Multi-modal Understanding and Generation

ZHAOQING WANG

Master of Philosophy

Supervisor: Prof. Tongliang Liu
Associate Supervisor: Dr. Baosheng Yu

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

23 March 2026

Abstract

The convergence of computer vision and natural language processing has been accelerated by large-scale pre-training, endowing machines with the ability to align visual concepts with semantic meaning. However, a significant gap persists between the coarse capabilities of existing foundation models and the requirements of advanced multi-modal intelligence, specifically, the ability to perform fine-grained dense prediction and unified generation in open-world scenarios. Current approaches are constrained by the high cost of pixel-level supervision, the inflexibility of closed-vocabulary training, and the architectural dichotomy between discriminative and generative tasks. This thesis argues that these challenges are fundamentally connected, and addresses them through a unified research agenda: improving vision-language alignment, extending it to scalable open-world supervision, and ultimately integrating diverse visual tasks into a single generative multi-modal framework.

The first part of the thesis focuses on fine-grained semantic grounding as the foundation of unified multi-modal learning. We investigate Referring Image Segmentation, where the goal is to segment a specific region based on a linguistic description. We propose CRIS, a CLIP-Driven Referring Image Segmentation framework. Unlike previous approaches that separately transfer language and vision knowledge, CRIS explicitly enforces text-to-pixel alignment through a novel vision-language decoder and contrastive learning. This allows the model to leverage the rich semantic knowledge of CLIP to achieve state-of-the-art performance in locating complex visual entities.

Building on this alignment foundation, the second part of the thesis addresses scalability in open-world settings, where fixed category vocabularies and expensive dense annotations limit real-world applicability. We introduce Unpair-Seg, a framework for Open-Vocabulary Segmentation with Unpaired Mask-Text Supervision. By utilizing readily available standalone images, image-text pairs, and image-mask pairs, without requiring strictly aligned triplets, this method significantly reduces annotation costs. We employ a large vision-language model

(LVLM) to generate precise entity descriptions and design a multi-scale matching strategy to align masks with textual entities in a shared embedding space, narrowing the performance gap between fully-supervised and weakly-supervised methods.

The third part of the thesis moves beyond task-specific dense understanding to architectural unification across vision tasks. We present LaVin-DiT, a Large Vision Diffusion Transformer capable of handling over 20 computer vision tasks within a single generative framework. By incorporating a spatial-temporal variational autoencoder and a joint diffusion transformer, LaVin-DiT models visual data as a continuous latent space rather than discrete tokens. Through in-context learning, the model adapts to diverse tasks (e.g., from depth estimation to video generation) without fine-tuning, demonstrating the potential of diffusion models as unified vision foundation models.

Taken together, these three frameworks form a coherent progression toward a unified multi-modal learning paradigm. CRIS establishes fine-grained vision-language alignment at the pixel level. Unpair-Seg makes such alignment scalable under open-vocabulary and weakly supervised conditions. And, LaVin-DiT generalizes these ideas into a universal generative architecture that bridges understanding and generation. Collectively, the thesis advances multi-modal learning from coarse image-level supervision toward scalable, fine-grained, and unified foundation models for open-world visual intelligence.

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Name: _____

Date: 21 March 2026

Declaration of Use of Generative AI

I acknowledge the use of Gemini to improve the grammatical structure and clarity of the writing.

The ideas, experimental designs, and interpretations of the results presented in this thesis are entirely my own. I have reviewed and edited all outputs generated by the AI tools and take full responsibility for the accuracy and integrity of the final content.

Name: _____

Date: 21 March 2026

Authorship Attribution Statement

(1): Chapter 2 of this thesis is published as [1].

I led the project with co-authors, analyzed data, designed algorithms, conducted experiments, and wrote paper drafts.

(2): Chapter 3 of this thesis is submitted as [2].

I led the project with co-authors, collected and analyzed data, designed algorithms, conducted experiments, and wrote paper drafts.

(3): Chapter 4 of this thesis is published as [3].

I led the project with co-authors, collected and analyzed data, designed algorithms, conducted experiments, and wrote paper drafts.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Name: _____

Date: 21 March 2026

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Name: _____

Date: 21 March 2026

Sidere mens eadem mutato.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my brilliant supervisors, Prof. Tongliang Liu and Prof. Mingming Gong, for their invaluable guidance, patience, and support throughout my PhD study. I also gratefully acknowledge the financial support provided by the Faculty of Engineering Research Stipend Scholarship, which made the research reported in this thesis possible. My supervisors' profound insights into Machine Learning and Computer Vision, along with their constant encouragement to explore new ideas, have been instrumental in shaping this work. Thank you for believing in me, even when experiments failed or results were elusive.

I am also incredibly grateful to my impressive list of collaborators for their constructive feedback and for pushing the quality of our work to a higher standard: Qiang Li, Yu Lu, Xunqiang Tao, Guoxin Zhang, Pengfei Wan, Wen Zheng, Nannan Wang, Xiaowei Chi, Jiaming Liu, Ming Lu, Rongyu Zhang, Yandong Guo, Kaixin Chen, Xiaoqi Li, Enhua Wu, Chuang Zhang, Ming Wu, Dongdong Yu, Jinlin Liu, Changhu Wang, Bo Han, Xiao He, Ziyu Chen, Haodong Chen, Yanwu Xu, Xiangyu Kong, Zhanbei Cui, Jiepeng Wang, Wenping Wang, Yijie Huang, Shaokun Zhang, and Xiaoli Wei.

To the members, both past and present, of the Trustworthy Machine Learning Lab and the Sydney AI Center at The University of Sydney, thank you for creating such a vibrant and enjoyable working environment. Special thanks to Runnan Chen, Qiang Qu, Yuhao Wu, Yongli Xiang, Zhenchen Wan, Quan Tran, Tianyu Huang, Jiaxin Huang, Li He, Jun Wang, Ziming Hong, Xiuchuan Li, Muyang Li, Yexiong Lin, Jiyang Zheng, Suqin Yuan, Zhuo Huang, Runqi Lin, Chaojian Yu, Xiaobo Xia, Yingbin Bai, Huaxi Huang, Yu Yao, Keshen Zhou, Xiangyu Sun, and Hui Kang.

On a personal note, to my father and mother: there are no words to express my gratitude for your unconditional love and support. Thank you for the sacrifices you made to provide

me with the education that brought me here. Your faith in me has been my foundation, and everything I have achieved is thanks to you.

Last but certainly not least, I want to thank my girlfriend, Emily Song. Thank you for your endless patience, understanding, and encouragement, especially during the stressful weeks leading up to deadlines. Thank you for listening to my ramblings about research problems, for reminding me to take breaks, and for always being by my side. I could not have finished this journey without you.

Contents

List of Figures

CHAPTER 1

Introduction

The ultimate goal of Artificial General Intelligence (AGI) [4], [5] is to create systems that can perceive the physical world with the acuity of human vision and reason about it with the flexibility of natural language. Historically, Computer Vision (CV) [1], [2], [3], [6], [7], [8], [9], [10] and Natural Language Processing (NLP) [11], [12], [13], [14] evolved as separate disciplines with distinct architectures and datasets. However, the landscape has been fundamentally reshaped by the advent of large-scale Multi-modal Learning. Foundation models, such as Contrastive Language-Image Pre-training (CLIP) [15], have moved the field away from training on limited, closed-set discrete labels toward learning from vast, web-scale image-text pairs. This paradigm shift has equipped machines with robust, open-world representations, enabling impressive capabilities in zero-shot classification and retrieval.

Despite these strides, the transition from coarse image-level alignment to a truly unified, fine-grained, and generative understanding of the visual world remains incomplete. Applying current foundation models to complex downstream tasks reveals three critical bottlenecks that this thesis seeks to address: the granularity gap, the supervision bottleneck, and the architectural fragmentation.

The Granularity Gap in Dense Prediction. First, most vision-language models [15], [16], [17] are pre-trained on global objectives, aligning a single vector representation of an entire image with a sentence. While effective for retrieval, this global pooling operation discards the spatial information necessary for dense prediction tasks, such as Referring Image Segmentation (RIS) [18], [19], [20], [21]. In RIS, the model must locate a specific object (e.g., the man in the yellow vest) among multiple similar instances. Direct application of image-level

features fails to capture these pixel-level nuances. Consequently, prior approaches [18], [22], [23], [24] often resorted to complex, multi-stage pipelines that trained vision and language encoders separately, failing to learn the intricate, fine-grained interactions required to ground linguistic concepts into specific pixels.

The Supervision Bottleneck and Open-World Scalability. Second, achieving high performance in segmentation [25], [26] has traditionally required expensive triplet supervision, datasets containing images, precise pixel-wise masks, and aligned text descriptions. While image-text pairs are abundant on the internet, high-quality segmentation masks are labor-intensive to annotate [2]. This reliance on strictly aligned data restricts models to closed vocabularies defined by the training set [27], [28]. To achieve true Open-vocabulary segmentation, models must generalize to concepts seen only in text during training [29], [30], [31]. The challenge lies in learning these capabilities without incurring the prohibitive cost of manual dense annotation, necessitating new frameworks that can leverage noisy, "unpaired" data (e.g., standalone images or loose image-text pairs) to approximate the performance of fully supervised systems.

Architectural Fragmentation and Unified Generation. Third, the computer vision ecosystem is currently fragmented between understanding and generation. Furthermore, existing attempts to build unified Large Vision Models (LVMs) often blindly adapt architectures from NLP [12], treating images as sequences of discrete tokens in an autoregressive manner [32]. This approach is computationally inefficient for high-dimensional visual data and often disrupts the essential 2D spatial structure of images and the temporal coherence of videos. There is a pressing need for a generalist foundation model, one that avoids the pitfalls of discrete tokenization, preserves spatial-temporal continuity, and unifies perception and generation within a single, scalable framework.

To address the challenges outlined above, this thesis presents three distinct but interconnected frameworks. We progress from refining specific fine-grained alignment, to scaling segmentation via weak supervision, and finally to constructing a unified generalist model.

Fine-Grained Transfer via CLIP-Driven Referring Image Segmentation. To address the granularity gap in dense prediction, our first contribution investigates the transfer of global multi-modal knowledge to fine-grained pixel-level tasks. Existing foundation models like CLIP [15] are optimized for image-level alignment, which often leads to a loss of detail when applied to tasks requiring precise localization, such as Referring Image Segmentation. We propose CLIP-Driven Referring Image Segmentation (CRIS), an end-to-end framework that explicitly enforces text-to-pixel alignment rather than simple feature concatenation. By designing a specialized vision-language decoder and introducing a novel text-to-pixel contrastive learning objective, we enable the model to propagate semantic information from textual representations directly to the relevant pixel-level activations. This approach effectively distinguishes the referred entity from complex backgrounds, achieving state-of-the-art performance on benchmark datasets and demonstrating that global pre-training can be successfully adapted for dense prediction without heavy post-processing.

Scalable Open-Vocabulary Learning with Unpaired Supervision. Addressing the supervision bottleneck, our second contribution challenges the conventional reliance on expensive, strictly aligned annotations for segmentation. We introduce Unpair-Seg, a weakly-supervised framework designed to achieve Open-Vocabulary Segmentation using heterogeneous, unpaired data sources, for instance, standalone images, image-text pairs, and image-mask pairs, rather than difficult-to-acquire triplets. To bridge the semantic disconnect between visual masks and textual concepts in the absence of direct labels, we leverage a Large Vision-Language Model (LVLM) [33] as a knowledge engine to generate precise captions and extract entities from unlabelled images. We then implement a multi-scale bipartite matching strategy in the CLIP embedding space to reliably associate these noisy textual entities with predicted masks. This methodology significantly reduces annotation costs while narrowing the performance gap between weakly-supervised and fully-supervised approaches, offering a scalable path toward open-world recognition.

Unified Understanding and Generation via Diffusion Transformers. Finally, to resolve the architectural fragmentation between understanding and generation, our third contribution moves beyond task-specific models to a unified generalist architecture. We present LaVin-DiT

(Large Vision Diffusion Transformer), a foundation model capable of handling over 20 diverse computer vision tasks, ranging from depth estimation to video generation, within a single generative framework. Unlike standard Large Vision Models that inefficiently discretize visual data into autoregressive tokens, LaVin-DiT employs a Spatial-Temporal Variational Autoencoder (ST-VAE) to compress images and videos into a continuous latent space, thereby preserving their structural and temporal integrity. By utilizing a joint diffusion transformer guided by in-context learning, the model adapts to arbitrary tasks defined by input-target pairs without fine-tuning. This work establishes a robust alternative to token-based approaches, proving that diffusion models can serve as versatile, scalable generalist agents for both perception and synthesis.

The remainder of this thesis is structured as follows:

- Chapter 2 details the CRIS framework and related works, elaborating on the vision-language decoder design and the mathematical formulation of text-to-pixel contrastive loss.
- Chapter 3 presents Unpair-Seg and related works, discussing the pipeline for generating pseudo-labels via LVLMs and the optimization strategies for unpaired supervision.
- Chapter 4 introduces LaVin-DiT and related works, providing an in-depth analysis of the ST-VAE architecture, the joint diffusion mechanism, and the implementation of in-context learning for vision tasks.
- Chapter 5 concludes the thesis, summarizing the key findings and discussing future directions for creating more robust, efficient, and unified multi-modal systems.

CHAPTER 2

Clip-driven referring image segmentation

2.1 Introduction

Referring image segmentation [18], [19], [20] is a fundamental and challenging task at the intersection of vision and language understanding, which can be potentially used in a wide range of applications, including interactive image editing and human-object interaction. Unlike semantic and instance segmentation [34], [35], [36], [37], which requires segmenting the visual entities belonging to a pre-determined set of categories, referring image segmentation is not limited to indicating specific categories but finding a particular region according to the input language expression.

Since the image and language modality maintain different properties, it is difficult to explicitly align textual features with pixel-level activations. Benefiting from the powerful capacity of the deep neural network, early approaches [18], [22], [23], [24] concatenate textual features with each visual activation directly, and use these combined features to generate the segmentation mask. Subsequently, to address the lack of adequate interaction between two modalities, a series of methods [20], [39], [40], [41], [42] adopt the language-vision attention mechanism to better learn cross-modal features.

Existing methods [20], [39], [40], [41], [42] leverage external knowledge to facilitate learning in common, while they mainly utilize a single-modal pretraining (e.g., the pretrained image or text encoder), which is short of multi-modal correspondence information. By resorting to language supervision from large-scale unlabeled data, vision-language pretraining [38], [43], [44] is able to learn ample multi-modal representations. Recently, the remarkable success

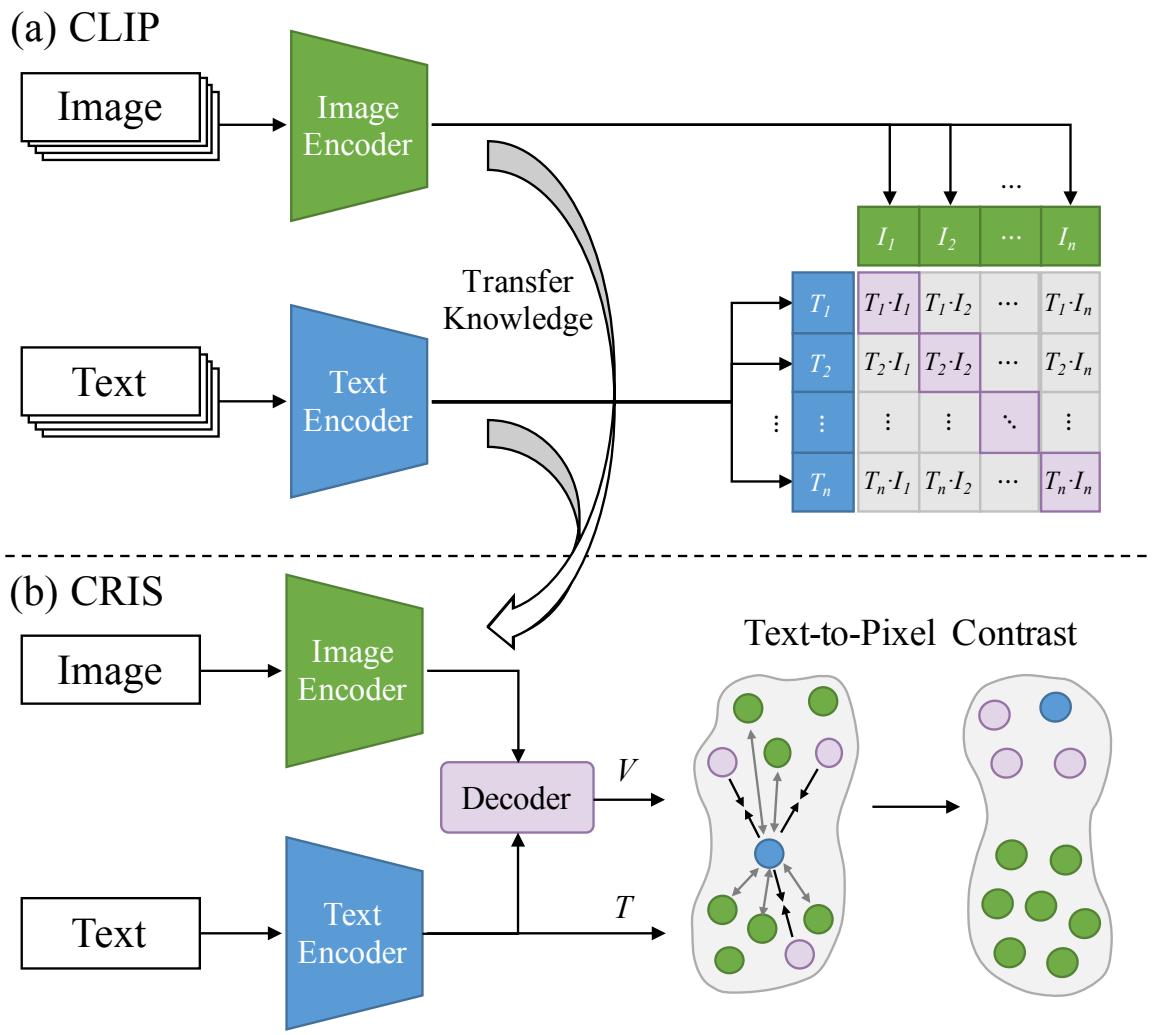


FIGURE 2.1. **An illustration of our main idea.** (a) CLIP [38] jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of image I and text T, which can capture the multi-modal corresponding information. (b) To transfer this knowledge of the CLIP model from image level to pixel level, we propose a CLIP-Driven Referring Image Segmentation (CRIS) framework. Firstly, we design a vision-language decoder to propagate fine-grained semantic information from textual features to pixel-level visual features. Secondly, we combine all pixel-level visual features V with the global textual feature T and adopt contrastive learning to pull text and related pixel-wise features closer and push other irrelevances away.

of the CLIP [38] has shown its capability of learning SOTA image-level visual concepts from 400 million image-text pairs, which assists many multi-modal tasks achieve significant improvements, including image-text retrieval [38], video-text retrieval [45], [46]. However, as

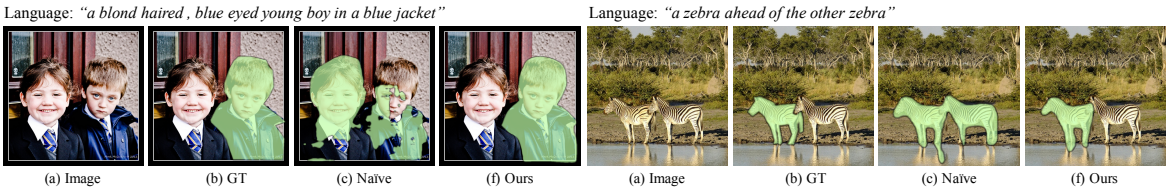


FIGURE 2.2. **Comparison between the direct fine-tuning and our proposed methods.** "Naive" denotes the direct fine-tuning mentioned in section 2.4. Compared with the direct fine-tuning, our method can not only leverage the powerful cross-modal matching capability of the CLIP, but also learn fine-grained visual representations.

shown in Figure 2.2, the direct usage of the CLIP can be sub-optimal for pixel-level prediction tasks, e.g., referring image segmentation, duo to the discrepancy between image-level and pixel-level prediction. The former focuses on the global information of an input image, while the latter needs to learn fine-grained visual representations for each spatial activation.

In this paper, we explore leveraging the powerful knowledge of the CLIP model for referring image segmentation, in order to enhance the ability of cross-modal matching. Considering the characteristics of referring image segmentation, we propose an effective and flexible framework named CLIP-Driven Referring Image Segmentation (CRIS), which can transfer ample multi-modal corresponding knowledge of the CLIP for achieving text-to-pixel alignment. Firstly, we propose a visual-language decoder that captures long-range dependencies of pixel-level features through the self-attention operation and adaptively propagate fine-structured textual features into pixel-level features through the cross-attention operation. Secondly, we introduce the text-to-pixel contrastive learning, which can align linguistic features and the corresponding pixel-level features, meanwhile distinguishing irrelevant pixel-level features in the multi-modal embedding space. Based on this scheme, the model can explicitly learn fine-grained visual concepts by interwinding the linguistic and pixel-level visual features.

Our main contributions are summarized as follow:

- We propose a CLIP-Driven Referring Image Segmentation framework (CRIS) to transfer the knowledge of the CLIP model for achieving text-to-pixel alignment.

- We take fully advantage of this multi-modal knowledge with two innovative designs, i.e., the vision-language decoder and text-to-pixel contrastive learning.
- The experimental results on three challenging benchmarks significantly outperform previous state-of-the-art methods by large margins (e.g., $+\,4.89\,IoU$ on RefCOCO, $+\,8.88\,IoU$ on RefCOCO+, $+\,5.47\,IoU$ on G-Ref).

2.2 Related Work

Vision-Language Pretraining. Vision-Language pretraining has made rapid progress in recent years and achieved impressive performance on various multi-modal downstream tasks. By resorting to semantic supervision from large-scale image data, several approaches [38], [43], [44] were proposed to learn visual representations from text representations. MIL-NCE [44] mainly explored leveraging noisy large-scale Howto100M [47] instructional videos to learn a better video encoder via an end-to-end manner. SimVLM [43] reduced the training complexity by leveraging large-scale weak supervision, and adopted a single prefix language modeling objective in an end-to-end manner. Benefit from the large-scale image and text pairs collected from the Internet, a recent approach, i.e., Contrastive Language-Image Pretraining (CLIP) [38], achieved the notable success of aligning two modalities representations in the embedding space. CLIP adopted contrastive learning with high-capacity language models and visual feature encoders to capture compelling visual concepts for zero-shot image classification. More recently, a series of works [45], [46], [48], [49] were proposed to transfer the knowledge of CLIP models to downstream tasks and achieved promising results, such as video caption, video-text retrieval, and image generation. Different from these works, we transfer these image-level visual concepts to referring image segmentation for leveraging multi-modal corresponding information.

Contrastive Learning Date back to [50], these approaches learned representations by contrasting positive pairs against negative pairs. Several approaches [51], [52], [53], [54], [55] were proposed to treat each image as a class and use contrastive loss-based instance discrimination for representation learning. Recently, VADeR and DenseCL [56], [57] proposed to

explore pixel-level contrastive learning to fill the gap between self-supervised representation learning and dense prediction tasks. Besides, CLIP [38] proposed a promising alternative that directly learns transferable visual concepts from large-scale collected image-text pairs by using cross-modal contrastive loss. In this paper, we propose a CLIP-Driven Referring Image Segmentation (CRIS) framework to transfer the knowledge of the CLIP model to referring image segmentation in an end-to-end manner.

Referring Image Segmentation Referring image segmentation is to segment a target region (e.g., object or stuff) in an image by understanding a given natural linguistic expression, which was first introduced by [18]. Early works [22], [23], [24] first extracted visual and linguistic features by CNN and LSTM, respectively, and directly concatenated two modalities to obtain final segmentation results by a FCN [58]. In [19], they proposed a two-stage method that first extracted instances using Mask R-CNN [37], and then adopted linguistic features to choose the target from those instances. Besides, MCN [59] designed a framework achieving impressive results. They learned to optimize two related tasks, i.e., referring expression comprehension and segmentation, simultaneously.

As the attention mechanism arouses more and more interests, a series of works are proposed to adopt the attention mechanism. It is powerful to extract the visual contents corresponding to the language expression. [40] used the vision-guided linguistic attention to aggregate the linguistic context of each visual region adaptively. [20] designed a Cross-Modal Self-Attention (CSMA) module to focus on informative words in the sentence and crucial regions in the image. [60] proposed a bi-directional relationship inferring network that adopted a language-guided visual and vision-guided linguistic attention module to capture the mutual guidance between two modalities. Besides, LTS [61] designs a strong pipeline that decouples the task into a "Locate-Then-Segment" scheme by introducing the position prior. EFNet [62] designs a co-attention mechanism to use language to refine the multi-modal features progressively, which can promote the consistent of the cross-modal information representation. More recently, VLT[63] employs transformer to build a network with an encoder-decoder attention mechanism for enhancing the global context information. Different from previous methods,

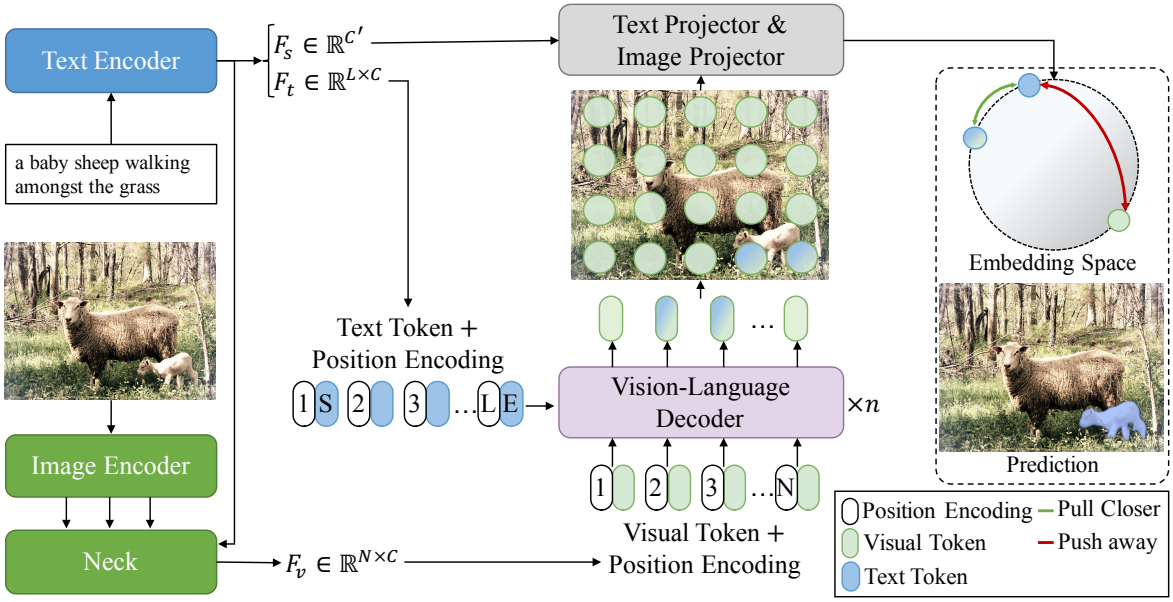


FIGURE 2.3. **The overview of the proposed CLIP-Driven Referring Image Segmentation (CRIS) framework.** CRIS mainly consists of a text encoder, an image encoder, a cross-modal neck, a vision-language decoder, and two projectors. The vision-language decoder is used to adaptively propagate semantic information from textual features to visual features. The text-to-pixel contrastive learning is used to explicitly learn fine-grained multi-modal corresponding information by interwinding the text features and pixel-level visual features.

we aim to leverage the knowledge of the CLIP, in order to improving the compatibility of multi-modal information and boost the ability of cross-modal matching.

2.3 Methodology

As illustrated in Figure 2.3, we introduce how the proposed CRIS framework transfers the knowledge of CLIP to referring image segmentation to achieve text-to-pixel alignment by leveraging multi-modal corresponding information. Firstly, we use a ResNet [64] and a Transformer [65] to extract image and text features respectively, which are further fused to obtain the simple multi-modal features. Secondly, these features and text features are fed into the vision-language decoder to propagate fine-grained semantic information from textual representations to pixel-level visual activations. Finally, we use two projectors to produce the

final prediction mask, and adopt the text-to-pixel contrastive loss to explicitly align the text features with the relevant pixel-level visual features.

2.3.1 Image & Text Feature Extraction

As illustrated in Figure 2.3, the input of our framework consists of an image I and a referring expression T.

Image Encoder. For an input image $I \in \mathbb{R}^{H\times W\times 3}$, we utilize multiple visual features from the 2th-4th stages of the ResNet, which are defined as $F_{v2} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C_2}$, $F_{v3} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C_3}$, and $F_{v4} \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C_4}$, respectively. Note that C is the feature dimension, H and W are the height and width of the original image.

Text Encoder. For an input expression $T \in \mathbb{R}^{L}$, we adopt a Transformer [65] modified by [15] to extract text features $F_t \in \mathbb{R}^{L\times C}$. The Transformer operates on a lower-cased byte pair encoding (BPE) representation of the text with a 49,152 vocab size [66], and the text sequence is bracketed with `[SOS]` and `[EOS]` tokens. The activations of the highest layer of the transformer at the `[EOS]` token are further transformed as the global textual representation $F_s \in \mathbb{R}^{C'}$. Note that C and C' are the feature dimension, L is the length of the referring expression.

Cross-modal Neck. Given multiple visual features and the global textual representation F_s, we obtain the simple multi-modal feature $F_{m4} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$ by fusing F_{v4} with F_s:

$$F_{m4} = Up(\sigma(F_{v4}W_{v4}) \cdot \sigma(F_sW_s)), \tag{2.1}$$

where $Up(\cdot)$ denotes $2\times$ upsampling, $\cdot$ denotes the element-wise multiplication, σ denotes ReLU, W_{v4} and W_s are two learnable matrices to transform the visual and textual representations into the same feature dimension. Then, the multi-modal features F_{m2} and F_{m3} are obtained by:

$$\begin{aligned} F_{m_3} &= [\sigma(F_{m_4}W_{m_4}),\ \sigma(F_{v_3}W_{v_3})], \\ F_{m_2} &= [\sigma(F_{m_3}W_{m_3}),\ \sigma(F'_{v_2}W_{v_2})],\ F'_{v_2} = Avg(F_{v_2}), \end{aligned} \tag{2.2}$$

where $Avg(\cdot)$ denotes a kernel size of 2×2 average pooling with 2 strides, respectively. $[,]$ is the concatenation operation. Subsequently, we concatenate three multi-modal features and use a 1×1 convolution layer to aggregate them:

$$F_m = Conv([F_{m_2},\ F_{m_3},\ F_{m_4}]), \tag{2.3}$$

where $F_m \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$. Finally, we concatenate a 2D spatial coordinate feature $F_{coord} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 2}$ with F_m and fuse that by a 3×3 convolution [67]. The visual feature $F_v \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ is calculated as follow,

$$F_v = Conv([F_m,\ F_{coord}]). \tag{2.4}$$

As shown in figure 2.3, we flatten the spatial domain of F_v into a sequence, forming the visual feature $F_v \in \mathbb{R}^{N \times C}$, $N = \frac{H}{16} \times \frac{W}{16}$, which is utilized in the following process.

2.3.2 Vision-Language Decoder

We design a vision-language decoder to adaptively propagate fine-grained semantic information from textual features to visual features. As shown in Figure 2.3, the decoder module takes textual features $F_t \in \mathbb{R}^{L \times C}$ and pixel-level visual features $F_v \in \mathbb{R}^{N \times C}$ as inputs, which can provide ample textual information corresponding to visual features. To capture positional information, the fixed sine spatial positional encodings are added to F_v [68] and F_t [65], respectively. The vision-language decoder composed of n layers is applied to generate a sequence of evolved multi-modal features $F_c \in \mathbb{R}^{N \times C}$. Following the standard architecture of the transformer [65], each layer consists of a multi-head self-attention layer, a multi-head cross-attention layer, and a feed-forward network. In one decoder layer, F_v is first sent into the multi-head self-attention layer to capture global contextual information:

$$F'_v = MHSA(LN(F_v)) + F_v, \tag{2.5}$$

where F'_v is the evolved visual feature, $MHSA(\cdot)$ and $LN(\cdot)$ denote the multi-head self-attention layer and Layer Normalization [69], respectively. The multi-head self-attention

mechanism is composed of three point-wise linear layers mapping F_v to intermediate representations, queries $Q \in \mathbb{R}^{N \times d_q}$, keys $K \in \mathbb{R}^{N \times d_k}$ and values $V \in \mathbb{R}^{N \times d_v}$. Multi-head self-attention is then calculated as follows,

$$MHSA(Q,\, K,\, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{2.6}$$

After that, the multi-head cross-attention layer is adopted to propagate fine-grained semantic information into the evolved visual features, where one point-wise linear layer maps F_v' to Q, and the other two linear layers map F_t to K and V. To obtain the multi-modal feature F_c, the output query Q is further computed by a MLP block of two layers with Layer Normalization and residual connections:

$$\begin{aligned} F_c' &= MHCA(LN(F_v'),\, F_t) + F_v', \\ F_c &= MLP(LN(F_c')) + F_c', \end{aligned} \tag{2.7}$$

where $MHCA(\cdot)$ denotes the multi-head cross-attention layer, and F_c' is the intermediate features. The evolved multi-modal feature F_c is utilized for the final segmentation mask. Note that the hyper-parameter n is discussed in the following experiment section.

2.3.3 Text-to-Pixel Contrastive Learning

Although the CLIP [38] learns powerful image-level visual concepts by aligning the textual representation with the image-level representation, this type of knowledge is sub-optimal for referring image segmentation, due to the lack of more fine-grained visual concepts.

To tackle this issue, we design a text-to-pixel contrastive loss, which explicitly aligns the textual features with the corresponding pixel-level visual features. As illustrated in Figure 2.3, image and text projector are adopted to transform F_c and F_s as follow:

$$\begin{aligned} z_v &= F_c' W_v + b_v,\; F_c' = Up(F_c), \\ z_t &= F_s W_t + b_t, \end{aligned} \tag{2.8}$$

where $z_t \in \mathbb{R}^D$, $z_v \in \mathbb{R}^{N \times D}$, $N = \frac{H}{4} \times \frac{W}{4}$, Up denotes $4\times$ upsampling, W_v and W_t are two learnable matrices to transform F_c and F_s into the same feature dimension D, b_v and b_t are two learnable biases.

Given a transformed textual feature z_t and a set of transformed pixel-level features z_v, a contrastive loss is adopted to optimize the relationship between two modalities, where z_t is encouraged to be similar with its corresponding z_v and dissimilar with other irrelevant z_v. With the similarity measured by dot product, the text-to-pixel contrastive loss can be formulated as:

$$L^i_{con}(z_t,\, z^i_v) = \begin{cases} -\log \sigma(z_t \cdot z^i_v), & i \in \mathcal{P}, \\ -\log(1 - \sigma(z_t \cdot z^i_v)), & i \in \mathcal{N}, \end{cases} \tag{2.9}$$

$$L_{con}(z_t,\, z_v) = \frac{1}{|\mathcal{P} \cup \mathcal{N}|} \sum_{i \in \mathcal{P} \cup \mathcal{N}} L^i_{con}(z_t,\, z^i_v), \tag{2.10}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the class of "1" and "0" in the ground truth, $|\mathcal{P} \cup \mathcal{N}|$ is the cardinality, σ is the sigmoid function. Finally, to obtain the final segmentation results, we reshape $\sigma(z_t \cdot z_v)$ into $\frac{H}{4} \times \frac{W}{4}$ and upsample it back to the original image size.

2.4 Experimental results

Our proposed framework is built on different image encoders (e.g., ResNet-50, ResNet-101 [64]) and compared with a series of state-of-the-art methods. To evaluate the effectiveness of each component in our method, we conduct extensive experiments on three benchmarks, including RefCOCO [70], RefCOCO+ [70], and G-Ref [71].

2.4.1 Datasets

RefCOCO [70] is one of the largest and most commonly used datasets for referring image segmentation. It contains 19,994 images with 142,210 referring expressions for 50,000 objects, which are collected from the MSCOCO [72] via a two-player game [70]. The dataset is split into 120,624 train, 10,834 validation, 5,657 test A, and 5,095 test B samples, respectively.

According to statistics, each image contains two or more objects and each expression has an average length of 3.6 words.

RefCOCO+ [70] dataset contains 141,564 language expressions with 49,856 objects in 19,992 images. The dataset is split into train, validation, test A, and test B with 120,624, 10,758, 5,726, and 4,889 samples, respectively. Compared with RefCOCO dataset, some kinds of absolute-location words are excluded from the RefCOCO+ dataset, which could be more challenging than the RefCOCO dataset.

G-Ref [73] includes 104,560 referring expressions for 54,822 objects in 26,711 images. Unlike the above two datasets, natural expressions in G-Ref are collected from Amazon Mechanical Turk instead of a two-player game. The average length of sentences is 8.4 words, which have more words about locations and appearances. It is worth mentioning that we adopt UNC partition in this paper.

2.4.2 Implementation Details

Experimental Settings. We initiate the text and image encoder with CLIP [38], and adopt ResNet-50 [64] as the image encoder for all ablation studies. Input images are resized to 416×416. Due to the extra `[SOS]` and `[EOS]` tokens, and the input sentences are set with a maximum sentence length of 17 for RefCOCO and RefCOCO+, and 22 for G-Ref. Each Transformer Decoder layer has 8 heads, and the feed-forward hidden dimension is set to 2048. We train the network for 50 epochs using the Adam optimizer with the learning rate $\lambda = 0.0001$. The learning rate is decreased by a factor of 0.1 at the 35th epoch. We train the model with a batch size of 64 on 8 Tesla V100 with 16 GPU VRAM.

During inference, we upsample the predicted results back to the original image size and binarize them at a threshold of 0.35 as the final result. No other post-processing operations are needed.

Metrics. Following previous works [22], [23], [24], [63], we adopt two metrics to verify the effectiveness: IoU and Precision@X. The IoU calculates intersection regions over

union regions of the predicted segmentation mask and the ground truth. The Precision@X measures the percentage of test images with an IoU score higher than the threshold $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, which focuses on the location ability of the method.

2.4.3 Ablation Study

The proposed CRIS framework consists of two main parts, i.e., text-to-pixel contrastive learning and vision-language decoder. To investigate each component in our method, we conduct extensive experiments on the validation set of three widely used datasets.

Effectiveness of Contrastive Learning & Vision-Language Decoder. Firstly, we remove the parts of the text-to-pixel contrastive learning and vision-language decoder from the framework to build our baseline, which is same as the naive setting in Figure 2.2.(c). As illustrated in Table 2.1, we introduce the contrastive learning scheme, which significantly increases the IoU accuracy of 1.98%, 2.98%, and 3.43% than the baseline network on three datasets, respectively. This superior performance gain proves that the contrastive loss can encourage the model to explicitly pull closer linguistic and relevant pixel-level visual representations and push away other irrelevances for learning fine-structured multi-modal corresponding information.

TABLE 2.1. **Ablation studies on validation set of three benchmarks.** $Con.$ denotes the proposed text-to-pixel contrastive learning. $Dec.$ denotes the proposed vision-language decoder. n denotes the number of layers in the vision-language decoder. We set $Num = 3$ as the default. "Params" and "FPS" denote the parameter complexity (M) and inference speed, respectively. Given an image $I \in \mathbb{R}^{416\times416\times3}$, they are calculated on a Tesla V100 GPU. Gray lines denote the baseline network.

Dataset	$Con.$	$Dec.$	n	IoU	Pr@50	Pr@60	Pr@70	Pr@80	Pr@90	Params	FPS
RefCOCO	-	-	-	62.66	72.55	67.29	59.53	43.52	12.72	131.86	27.30
	✓	-	-	64.64	74.89	69.58	61.70	45.50	13.31	134.22	25.79
	-	✓	1	66.31	77.66	72.99	65.67	48.43	14.81	136.07	23.02
	✓	✓	1	68.66	80.16	75.72	68.82	51.98	15.94	138.43	22.64
	✓	✓	2	69.13	80.96	76.60	69.67	52.23	16.09	142.64	20.68
	✓	✓	3	**69.52**	**81.35**	**77.54**	**70.79**	**52.65**	16.21	146.85	19.22
	✓	✓	4	69.18	80.99	76.74	69.32	52.57	**16.37**	151.06	18.26
RefCOCO+	-	-	-	50.17	54.55	47.69	40.19	28.75	8.21	131.86	27.30
	✓	-	-	53.15	58.28	53.74	46.67	34.01	9.30	134.22	25.79
	-	✓	1	54.73	63.31	58.89	52.46	38.53	11.70	136.07	23.02
	✓	✓	1	59.97	69.19	64.85	58.17	43.47	13.39	138.43	22.64
	✓	✓	2	60.75	70.69	66.83	60.74	45.69	13.42	142.64	20.68
	✓	✓	3	**61.39**	**71.46**	**67.82**	**61.80**	**47.00**	**15.02**	146.85	19.22
	✓	✓	4	61.15	71.05	66.94	61.25	46.98	14.97	151.06	18.26
G-Ref	-	-	-	49.24	53.33	45.49	36.58	23.90	6.92	131.86	25.72
	✓	-	-	52.67	59.27	52.45	44.12	29.53	8.80	134.22	25.33
	-	✓	1	51.46	58.68	53.33	45.61	31.78	10.23	136.07	22.57
	✓	✓	1	57.82	66.28	60.99	53.21	38.58	13.38	138.43	22.34
	✓	✓	2	58.40	67.30	61.72	54.70	39.67	13.40	142.64	20.61
	✓	✓	3	**59.35**	**68.93**	**63.66**	**55.45**	**40.67**	**14.40**	146.85	19.14
	✓	✓	4	58.79	67.91	63.11	55.43	39.81	13.48	151.06	17.84

Besides, we evaluate the performance of the proposed vision-language decoder. Compared with the baseline network, we use one layer in the decoder, bringing 3.65%, 4.56%, and 2.22% IoU improvements on RefCOCO, RefCOCO+, and G-Ref, respectively. In particular, the self-attention operation can help the model sufficiently capture long-range dependencies across each pixel, which is helpful for understanding complex scenarios. Furthermore, each word encoded by the text encoder is used in the cross-attention operation, which can propagate fine-grained semantic information from textual features to pixel-level features to generate more discriminative visual representations and obtain more accurate segmentation masks.

Finally, combining the proposed contrastive loss and vision-language decoder, the IoU and Precision are significantly better than the baseline solely with the contrastive loss or decoder module, which further achieves large margins at about 4% - 8% on three datasets. The reason of this obvious complementary phenomenon is that the contrastive loss can guide the decoder to find the more informative emphasis and transfer this knowledge to more accurate pixel-level visual representations, which boosts the ability of cross-modal matching and generates precise segmentation masks.

Numbers of Layers in Decoder. In Table 2.1, the results illustrate the effect of utilizing different number of layers in the vision-language decoder. When the visual representations are sequentially processed by more layers, our model can consistently get better IoU of 69.52%, 61.39%, and 59.35% on three benchmarks, respectively. The setting of $n = 1$ may not taking full advantage of the multi-modal corresponding information from both vision and language. Meanwhile, the setting of $n = 4$ introduces more parameters, which could increase the risk of over-fitting. Considering the performance and efficiency, we set $n = 3$ as the default in our framework.

TABLE 2.2. **Comparisons with the state-of-the-art approaches on three benchmarks.** We report the results of our method with various visual backbones. "⋆" denotes the post-processing of DenseCRF [74]. "-" represents that the result is not provided. IoU is utilized as the metric.

Method	Backbone	RefCOCO			RefCOCO+			G-Ref	
		val	test A	test B	val	test A	test B	val	test
RMI⋆ [22]	ResNet-101	45.18	45.69	45.57	29.86	30.48	29.50	-	-
DMN [24]	ResNet-101	49.78	54.83	45.13	38.88	44.22	32.29	-	-
RRN⋆ [23]	ResNet-101	55.33	57.26	53.95	39.75	42.15	36.11	-	-
MAttNet [19]	ResNet-101	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [75]	ResNet-101	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
CMSA⋆ [20]	ResNet-101	58.32	60.61	55.09	43.76	47.60	37.89	-	-
Lang2Seg [39]	ResNet-101	58.90	61.77	53.81	-	-	-	46.37	46.95
BCAN⋆ [60]	ResNet-101	61.35	63.37	59.57	48.57	52.87	42.13	-	-
CMPC⋆ [41]	ResNet-101	61.36	64.53	59.64	49.56	53.44	43.23	-	-
LSCM⋆ [42]	ResNet-101	61.47	64.99	59.55	49.34	53.12	43.50	-	-
MCN [59]	DarkNet-53	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
CGAN [76]	DarkNet-53	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69
EFNet [62]	ResNet-101	62.76	65.69	59.67	51.50	55.24	43.01	-	-
LTS [61]	DarkNet-53	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT [63]	DarkNet-53	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
CRIS (Ours)	ResNet-50	69.52	72.72	64.70	61.39	67.10	52.48	59.35	59.39
CRIS (Ours)	ResNet-101	**70.47**	**73.18**	**66.10**	**62.27**	**68.08**	**53.68**	**59.87**	**60.36**

2.4.4 Main Results

We compare our proposed approach, CLIP-Driven Referring Image Segmentation, with a series of state-of-the-art methods on three commonly used datasets. As illustrated in Table 2.2, our method surpasses other methods on each split of all datasets even though we utilize a shallow ResNet-50 [64].

On the RefCOCO dataset, our model significantly outperforms the state-of-the-art Vision Language Transformer [63] by 4.82%, 4.89% and 3.37% on three splits, respectively, which indicates that our model effectively transfer the knowledge of the CLIP model from image-level to pixel-level, enhancing the ability of cross-modal matching.

Besides, in Table 2.2, our method achieves remarkable performance gains of about 4~8% than a series of state-of-the-art works on the more challenging RefCOCO+ dataset. These obvious improvements over them suggest that our method can adequately leverage the powerful knowledge of the CLIP to accurately focus the region referred by the given language expression.

Furthermore, on another more complex G-Ref dataset where the average length of referring expressions is complicated, our proposed method consistently achieve notable improvement of around 5% IoU than the state-of-the-art Locate then Segmentation [61]. As shown in Table 2.2, the results demonstrate that our proposed approach manages to understand long and complex sentences that contain more information and more emphases, and simultaneously perceive the corresponding object. Apart from that, longer referring expressions could contain complex scenarios, which need a strong ability to model the global contextual information. Our proposed vision-language decoder is suitable to enhance the holistic understanding of vision and language features.

2.4.5 Qualitative Analysis

Visualization. As illustrated in Figure 2.4, we present some visualization results with different setting, which demonstrates the benefits of each component in our proposed method. Firstly,

FIGURE 2.4. **Qualitative examples with different settings.** (a) the input image. (b) the ground turth. (c) the baseline network. (d) CRIS without Vision-Language Decoder. (e) CRIS without Contrastive Learning. (f) our proposed CRIS. *Best viewed in color.*

FIGURE 2.5. **Qualitative examples of failure cases.** *Best viewed in color.*

compared with our full model, the baseline network without the contrastive learning and vision-language decoder generates worse segmentation masks, because the baseline network fails to interwind referring expressions with the corresponding regions. Secondly, the setting of (d) and (e) can obtain more accurate segmentation results than the baseline network, but the model is still confused in some hard regions. Finally, our proposed full model can generate

high-quality segmentation masks, which demonstrates the effectiveness of our proposed method, i.e., CRIS.

Failure Cases. We visualize some insightful failed cases in Figure 2.5. One type of failure is caused by the ambiguity of the input expression. For the top left example in Figure 2.5, the expression of "yellow" is not enough to describe the region of the man in the yellow snowsuit. Besides, for the top right example, some failures are also caused by the wrong label. It is obvious that the top region is unrelated to "fingers". As shown in the bottom left example, the boundaries of the referent cannot be accurately segmented, but this issue can be alleviated by introducing other technologies, such as the refine module. Finally, occlusion could cause failure cases, which is a challenging problem in many vision tasks.

2.5 Conclusion

In this paper, we have investigated to leverage the power of Contrastive Language-Image Pretraining (CLIP) models to achieve text-to-pixel alignment for referring image segmentation. And, we have proposed an end-to-end CLIP-Driven Referring Image Segmentation (CRIS) framework to well transfer the knowledge of the CLIP model. Compared with the direct fine-tuning, our proposed framework not only inherit the strong cross-modal matching ability of the CLIP, but also learn ample fine-structured visual representations. The designed vision-language decoder can adaptively propagate sufficient semantic information of the language expression into pixel-level visual features, promoting consistency between two modalities. Furthermore, the introduced text-to-pixel contrastive learning can explicitly interwind the text representation and relevant pixel-level visual features, learning fine-grained multi-modal corresponding information. Extensive ablation studies on three commonly used datasets have verified the effectiveness of each proposed component, and our approach significantly outperforms previous state-of-the-art methods without any post-processing.

CHAPTER 3

Open-Vocabulary Segmentation with Unpaired Mask-Text Supervision

---

3.1 Introduction

Open-vocabulary segmentation refers to the segmentation and categorization of objects from an expensive and unrestricted vocabulary, even though the object categories within the vocabulary are not encountered during training [29], [30], [31]. In contrast to traditional closed-vocabulary segmentation [27], [28], which relies on predefined training categories and cannot recognize unseen categories, open-vocabulary segmentation enables recognition based on arbitrary text descriptions. This innovative segmentation paradigm has recently received considerable attention [25], [26], unlocking potential applications in areas [77], [78] such as autonomous driving and human-computer interaction [79].

Cutting-edge methods in open-vocabulary segmentation typically depend on the supervision of triplet annotations composed of images, masks, and corresponding texts (i.e., categories) [80]. However, collecting such precisely aligned annotations between the mask and text is expensive and labor-intensive. To mitigate this burden, existing weakly-supervised methods are proposed to use image-text pairs, i.e., text supervision [81], [82], [83]. However, these approaches struggle with dense prediction tasks due to their limited capacity to model complex spatial relationships. Moreover, without instance-level positional cues, they often fail to distinguish multiple instances within the same semantic class. These limitations hinder their generalization and segmentation quality.

To address these challenges, we introduce Unpair-Seg, a novel weakly-supervised open-vocabulary segmentation framework designed to reduce annotation costs while significantly

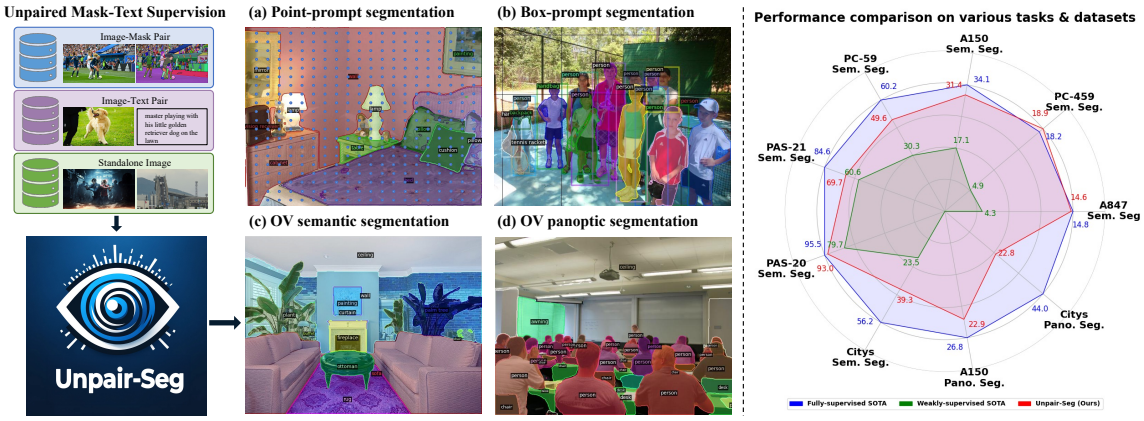


FIGURE 3.1. Unpair-Seg framework directly learns from **unpaired mask-text supervision**. Unlike labor-intensive image-mask-text annotations, image-mask pairs, image-text pairs, and standalone images are more accessible to collect from the Internet. With a single set of weights, Unpair-Seg excels at multiple image segmentation tasks. Extensive experimental results demonstrate that our method significantly narrows the gap between fully- and weakly-supervised methods.

improving segmentation performance. Unlike previous works requiring strict alignment, Unpair-Seg leverages unpaired supervision: independent image-mask pairs, image-text pairs, and standalone images. These data types are easier to collect at scale, as shown in Figure 3.1. By resorting to the data with weak supervision, Unpair-Seg learns strong segmentation abilities to group semantically similar pixels and align them with corresponding text entities within a shared embedding space, thereby enabling open-vocabulary segmentation.

Technically, when presented with an image-mask pair, we train a visual prompt encoder and a mask generator to predict a set of binary masks based on different types of visual prompts (e.g., points and bounding boxes). The collected image-text pair often contains certain texts that do not match the image [84], leading to an incorrect correspondence between masks and entities. Moreover, standalone images lack any text information. To tackle this, we employ a large vision language model (e.g., InternLM-XComposer [33]) to recaption each input image and extract more precise entity descriptions. Afterwards, we perform bipartite matching in the CLIP embedding space [38] to identify confident mask-entity pairs and generate pseudo labels. Due to the various scales of objects and stuff, we introduce a multi-scale matching that improves the quality of visual embedding, stabilizing mask-entity alignment. Given

these pseudo labels, we employ a semantic adapter to align regional semantic embeddings of predicted masks with entity embeddings of text descriptions into a shared CLIP embedding space. During inference, a zero-shot classifier constructed by embedding category names from the target dataset, assigns a category to each predicted mask, which enables the system to segment objects across an open vocabulary.

Before delving into details, we summarize our contributions.

- We introduce Unpair-Seg, a novel weakly-supervised open-vocabulary segmentation framework that significantly reduces the need for costly triplet annotations, making open-vocabulary segmentation more accessible and scalable.
- Considering the inherent noise in the vision-language correspondence, we employ the large vision language model to extract precise entities from input images and design a multi-scale matching strategy to exploit confident pairs of masks and entities.
- We conduct extensive experiments to justify our claims and demonstrate the superiority of Unpair-Seg. Unpair-Seg achieves impressive results of 22.8% PQ, 14.6%, and 19.6% mIoU in the challenging ADE20k panoptic segmentation, ADE-847, and PASCAL Context-459 semantic segmentation, respectively. Comprehensive ablation studies and discussions are also provided.

3.2 Related works

Generic segmentation. Given an image, segmentation of specific visual concepts has remained an ongoing research topic in computer vision, as indicated by the extensive literature on it [7], [37], [58]. Generic segmentation mainly includes semantic segmentation [58], [85], [86], [87], instance segmentation [37], [88], [89], and panoptic segmentation [7], [90], [91], related to different levels of granularity. In more detail, semantic segmentation [34], [87], [92], [93] aims to assign a label to each pixel of the input image according to its respective semantic classes. In addition, instance segmentation [94], [95], [96] attempts to distinguish different object instances of the same semantic class. Panoptic segmentation [97], [98], [99], [100] combines the characteristics of semantic segmentation and instance segmentation. Following

a close-vocabulary assumption, previous works can only predict predefined object categories. In this paper, we aim to build an advanced open-vocabulary segmentation framework that can categorize objects and stuff from an open set of vocabulary, learning from weak supervision.

Vision foundation models. Recent advancements in visual foundation models have diversified optimization techniques across various learning paradigms. These developments range from vision-only pretraining [32], [51], [101] to joint vision-language pre-training [38], [102], [103], [104], [105], [106], [107], [108], and extend to multi-modal frameworks that integrate visual prompting [109], [110], [111], [112]. A prime example of this evolution is SAM [9], [113], which shows the potential of extensive training for general segmentation, offering impressive generalizability and scalability. Despite its impressive capabilities, SAM cannot categorize predicted masks into different semantic classes, which is limited by the supervision of the image-mask pairs. More recently, Semantic-SAM [114] unifies different sources of human-annotated segmentation datasets and augments SAM by adding semantic labels and increasing levels of granularity. Our work aims to develop a more flexible vision foundation model, which can be trained with unpaired mask-text supervision (e.g., independent image-mask and image-text pairs) and can be easily adapted to different segmentation tasks.

Open-vocabulary segmentation. Open-vocabulary segmentation counters the constraints of closed-vocabulary segmentation by allowing the segmentation of a diverse range of classes, even those unseen during training [25], [115], [116], [117], [118], [119]. Existing works [80], [120], [121] leverage the pretrained vision-language models (e.g., CLIP [38] and ALIGN [102]) to perform open-vocabulary segmentation. Most open-vocabulary segmentation methods commonly use human-annotated supervision (i.e., the image-mask-text triplets) to generalize the capability of vision-language models from the image level to the pixel level. To reduce the dependency on this labor-intensive supervision, some weakly-supervised methods are proposed to use only text supervision [83], [122]. They learn to group image regions into shaped segments but struggle to distinguish different instances with the same semantic class, and the segmentation performance is unsatisfactory [81], [123]. This dilemma motivates our pursuit of a more advanced open-vocabulary segmentation framework that

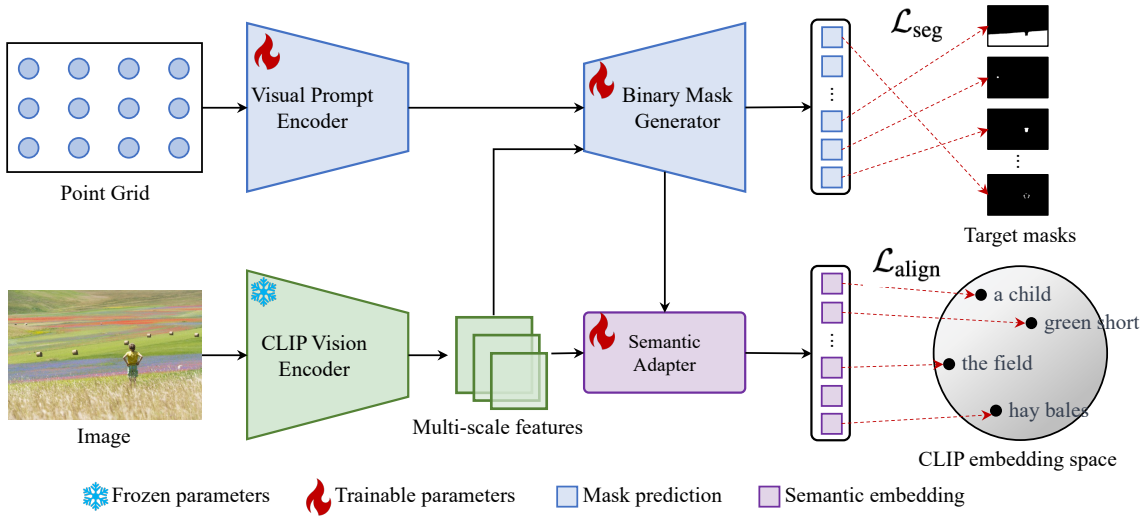


FIGURE 3.2. **Overview of the proposed Unpair-Seg framework.** Our framework consists of two parts, which are mask generation and mask-entity alignment. Given an input image, we uniformly sample a point grid as visual prompts and generate a set of corresponding binary mask proposals, which can be optimized with image-mask pairs. With these mask proposals, we apply a semantic adapter to extract semantic embeddings from multi-scale features. Using our designed matching strategy, we align these semantic embeddings with input entities into the CLIP embedding space, performing open-vocabulary segmentation. For simplicity, we omit the extraction of CLIP text embedding and our designed matching strategy.

minimizes annotation costs while achieving strong performance, narrowing the gap between fully-supervised and weakly-supervised methods.

3.3 Method

Problem definition. Given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the height and width of the image respectively, open-vocabulary segmentation aims to segment the image into a set of masks with associated semantic classes:

$$\{\mathbf{y}_i\}_{i=1}^{k} = \{(\mathbf{m}_i, \mathbf{c}_i)\}_{i=1}^{k}. \tag{3.1}$$

The k masks $\mathbf{m}_i \in \{0,1\}^{H \times W}$ include the associated ground truth class $\mathbf{c}_i$ [80]. Unlike traditional image segmentation tasks [7], [37], [58], open-vocabulary segmentation is more

challenging because inference classes are not observed during training. During the evaluation, the test categories $\mathbf{C}_{\text{test}}$ are different from $\mathbf{C}_{\text{train}}$, which contain novel categories that are not seen in training, i.e., $\mathbf{C}_{\text{train}} \neq \mathbf{C}_{\text{test}}$. Different from previous works [81], [121], in our setting, no paired human-annotated mask and semantic category is provided in advance for any training image. We assume that the category names of $\mathbf{C}_{\text{test}}$ are available, represented in the form of natural language.

A straightforward baseline. Before detailing our Unpair-Seg, we introduce a straightforward baseline using knowledge of image-text and image-mask pairs, for comparison. Specifically, we employ a CLIP model as a visual and text encoder, which is trained on a large amount of image-text pairs. Afterwards, we use the image-mask pairs to obtain a branch of mask generation, predicting a set of binary masks. To perform open-vocabulary segmentation, we crop and pool the CLIP image features based on these predicted masks, which are further classified by the CLIP text embeddings. Although this straightforward baseline enables open-vocabulary segmentation, it exhibits a noticeable knowledge gap between the image-level and pixel-level tasks. We also compare this baseline with our method in Section 3.4. Below, we introduce the technical details of the Unpair-Seg step by step.

3.3.1 Unpair-Seg Framework

Overview. Unpair-Seg is designed for weakly-supervised open-vocabulary segmentation, and is depicted in Figure 3.2. Specifically, on a macro level, our framework contains a CLIP model to extract features of both images and text descriptions. Optimizing with the image-mask pairs, our framework employs a branch of mask generation. This branch includes a visual prompt encoder and a binary mask generator, all of which work together to predict a set of binary mask proposals for an image, given input visual prompts (e.g., point grid, bounding box). With the image-text pairs, we employ a semantic adapter that gathers semantic embeddings from multi-scale features corresponding to the set of mask proposals. By resorting to our designed matching strategy, we align these semantic embeddings with entities in the text description into CLIP embedding space, performing open-vocabulary segmentation.

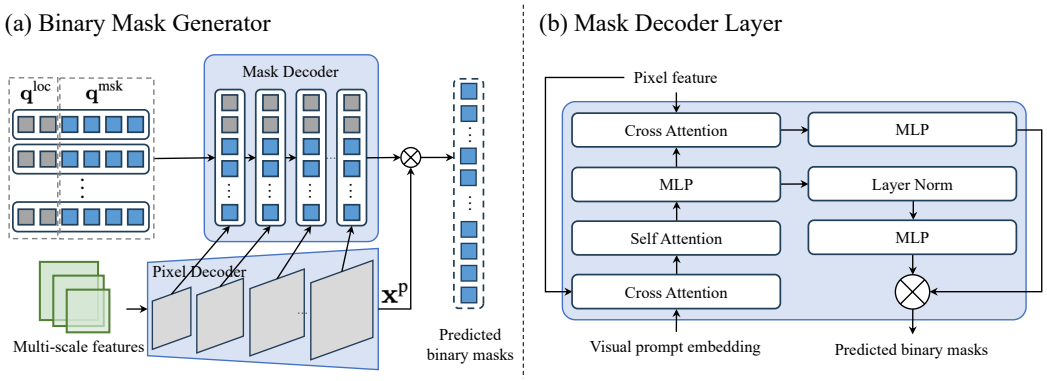


FIGURE 3.3. **Architecture of the binary mask generator and mask decoder layer.** (a) The mask generator consists of two parts, a pixel decoder and a mask decoder. (b) The mask decoder layer updates both visual prompt embeddings and pixel features by the cross-attention layers. The self-attention layer is used to update visual prompts. At each attention layer, positional encodings are added to the pixel features, and the entire original visual prompts (including position encoding) are added to the updated visual prompts.

Mask generation. Given an input image $\mathbf{I}$, mask generation aims to predict a set of binary masks $\mathbf{m}$. To meet interactivity and flexibility requirements, we employ a visual prompt encoder and a binary mask generator that associate predicted masks with input visual prompts.

Specifically, the points (x, y) and the boxes $(\mathrm{x}, \mathrm{y}, \mathrm{w}, \mathrm{h})$ are transformed into a unified format of anchor boxes. Each point is approximated by an anchor box with a pre-defined width and height. To address the challenge of predicting masks with variable granularity, we encode each point into two positional embeddings $\mathbf{q}_n^{\text{loc}} \in \mathbb{R}^{2\times d}$ using sinusoidal encoding [65], and combine them with M different mask embeddings $\mathbf{q}^{\text{msk}} \in \mathbb{R}^{M\times d}$. Each mask embedding is a learnable vector corresponding to a specific level of granularity. For box prompts, the process involves encoding into two position embeddings, followed by concatenation with a singular mask embedding, because each box is linked to a unique object. Afterward, we represent N input location prompts as a set of query embeddings $\mathbf{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_n\}$. The i-th query is calculated by the equation:

$$\mathbf{q}_i = \texttt{Concat}(\mathbf{q}_i^{\text{loc}}; \mathbf{q}^{\text{msk}}) + \mathbf{q}^{\text{type}} + \mathbf{q}^{\text{feat}}, \tag{3.2}$$

where $\texttt{Concat}(\cdot;\cdot)$ represents the concatenation operation, and $\mathbf{q}^{\text{type}} \in \mathbb{R}^d$ denotes the query type, chosen from either the point or the box. That $\mathbf{q}^{\text{feat}} \in \mathbb{R}^d$ is the content embedding

sampled from visual features. For box prompts, we sample visual features based on the center of each box.

For the binary mask generator shown in Figure 3.3, we adopt a pixel decoder equipped with multi-scale deformable attention [124] to model multi-scale features and output the enhanced pixel feature $\mathbf{x}^{\text{p}}$. Given each visual prompt, the corresponding binary mask $\mathbf{m}_n$ is derived through a matrix multiplication between the query embedding and the enhanced pixel feature,

$$\mathbf{m}_n = \texttt{Sigmoid}(\mathbf{q}_n^{\text{msk}} \cdot \mathbf{x}^{\text{p}}), \tag{3.3}$$

where $\mathbf{m}_n \in \mathbb{R}^{M \times H \times W}$ and $\texttt{Sigmoid}(\cdot)$ is the sigmoid function to normalize the mask values into $[0, 1]$. An input point is likely to exist in different granularity masks simultaneously. Given a point prompt, SAM [9] selects the prediction with the minimum cost to the ground truth for loss calculation, which cannot effectively model the inherent ambiguity of point prompts. In contrast, we automatically associate the ground-truth masks with the mask predictions of each point in a many-to-many manner, accurately predicting multi-granularity masks. Afterwards, we compute the segmentation loss for each mask prediction $\mathbf{m}_i$ with the corresponding ground-truth mask $\mathbf{y}_j^{\text{msk}}$. Besides, we estimate the quality $\mathbf{p}_i$ of each mask by regressing the intersection-over-union $\mathbf{y}_{i,j}^{\text{iou}}$ between the mask prediction $\mathbf{m}_i$ and the ground-truth mask $\mathbf{y}_j^{\text{msk}}$,

$$\begin{aligned} \mathcal{L}_{\text{seg}} = \lambda_{\text{bce}} \cdot \mathcal{L}_{\text{bce}}(\mathbf{m}_i, \mathbf{y}_j^{\text{msk}}) + \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{dice}}(\mathbf{m}_i, \mathbf{y}_j^{\text{msk}}) \\ + \lambda_{\text{iou}} \cdot \mathcal{L}_{\text{iou}}(\mathbf{p}_i, \mathbf{y}_{i,j}^{\text{iou}}), \end{aligned} \tag{3.4}$$

where λ_{bce}, λ_{dice}, and λ_{iou} are three weights used to balance the binary cross-entropy loss $\mathcal{L}_{\text{bce}}$, the dice loss $\mathcal{L}_{\text{dice}}$, and the quality loss $\mathcal{L}_{\text{iou}}$. Note that we do not apply any penalty to the remaining unmatched predicted masks.

Mask-entity alignment. In order to enable the model to categorize predicted masks from a wide range of vocabulary, we need to establish a connection between objects in the image and entities in the text description based on image-text pairs. However, datasets collected from the Internet often face several challenges, such as text-image misalignment, incomplete descriptions, and even missing text descriptions. To address these challenges, we utilize

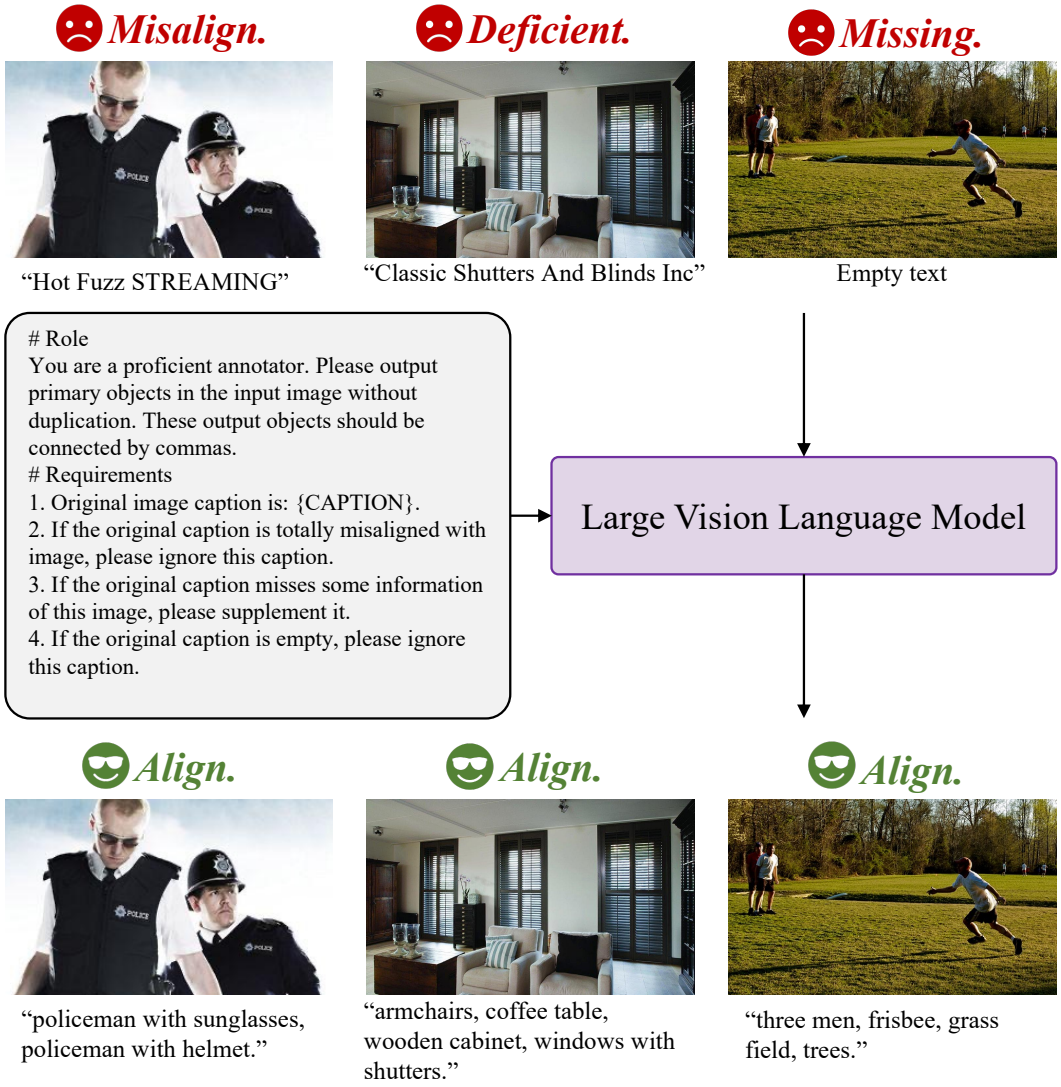


FIGURE 3.4. **Reception pipeline.** Large vision language model is used to extract entities from image-text pairs and sole images. "*Misalign.*", "*Deficient.*", and "*Missing.*" denote text–image misalignment, deficient description, and missing text.

the large vision language model (LVLM), e.g., InternLM-XComposer [33], to reception each input image with a pre-defined instruction. In doing so, we can accurately identify entities from the text descriptions, as illustrated in Figure 3.4. Afterwards, we extract the text embeddings $\mathbf{t}_k$ of each identified entity by the CLIP text encoder.

Matching mask proposals with these identified entities presents further challenges. Firstly, predicted masks are often redundant, and their granularity varies significantly, encompassing

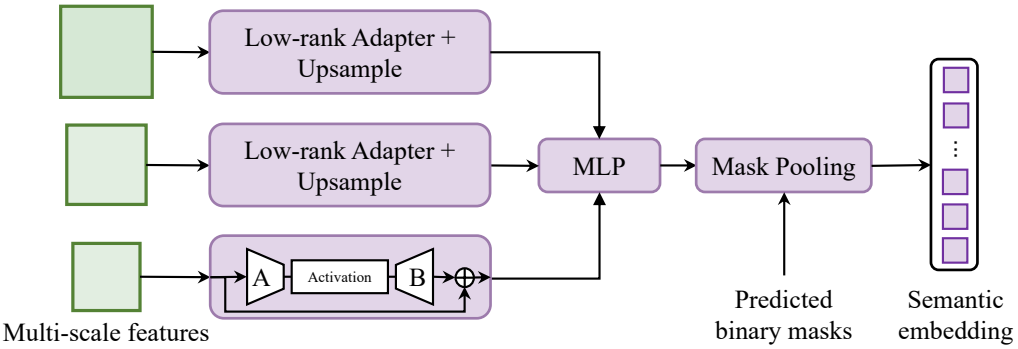


FIGURE 3.5. **Architecture of the semantic adapter.** We adopt low-rank adapters to transform multi-scale features, then fused by a MLP layer. Given predicted binary masks, we apply mask pooling operation to obtain a set of semantic embeddings.

part-level, object-level, and stuff-level segmentations. This diversity in granularity can lead to misalignment with entities, as most entities correspond to object-level or stuff-level concepts. Secondly, the scale of objects within images often varies considerably. Low-resolution image features are generally more suitable for recognizing stuff-level entities (e.g., sky, field, sea surface), whereas high-resolution features are better for recognizing smaller, object-level entities (e.g., cup, mobile phone, hay bales).

To tackle redundant masks and various granularities, as shown in Algorithm 1, we modify the Non-Maximum Suppression (NMS) algorithm that incorporates a coverage metric designed to filter out heavily overlapped and contained masks. This process preferentially retains object-level and stuff-level masks, which align more appropriately with the typical granularity of textual entities. Once we have non-redundant masks, we capture the information of objects at various scales from multi-resolution input images. A mask pooling layer and a projector F_{v} in CLIP visual encoder are employed to extract regional embeddings $\mathbf{r}_i$ from CLIP visual features $\mathbf{x}^{\text{v}}$ at each resolution,

$$\mathbf{r}_i = F_{\text{v}}(\texttt{AvgPool}(\mathbf{x}^{\text{v}}, \mathbf{m}_i)). \tag{3.5}$$

These multi-resolution regional embeddings are then ensembled to enhance the quality and robustness of the visual representation, $\hat{\mathbf{r}}_i = \sum_{s=1}^{S} \mathbf{r}_{i,s}/S$, where S denotes the number of

scales. Afterward, we calculate a cost matrix $\mathbf{ffi}$ between the region embeddings $\hat{\mathbf{r}}$ and text embeddings $\mathbf{t}$, i.e.,

$$\mathbf{ffi}_{i,j} = \frac{\exp(\mathbf{ffi}^{'}_{i,j})}{\sum_j \exp(\delta^{'}_{i,j})}, \quad \mathbf{ffi}^{'}_{i,k} = 1 - \frac{\mathbf{r}_i \cdot \mathbf{t}_k}{\|\mathbf{r}_i\|_2 \|\mathbf{t}_k\|_2}, \tag{3.6}$$

where $\mathbf{ffi}^{'}_{i,k}$ denotes the reverse cosine similarity between the i-th region embedding and the k-th text embedding. With this matrix, we apply the bipartite matching algorithm [125] to obtain the confident pairs.

After that, as illustrated in Figure 3.5, we design a semantic adapter to obtain semantic embeddings $\mathbf{r}^{'}$ for each binary mask from the multi-scale features. In the semantic adapter, we adopt multiple low-rank adapters to transform multi-scale features, where each layer incorporates two linear layers, denoted as A and B, with a non-linear activation function placed in between. Afterwards, we upsample each scale's feature into the same size and fuse them into the enhanced pixel feature $\mathbf{x}^{\mathrm{e}}$ with an MLP layer. Given the predicted masks, we apply an extract semantic embedding $\mathbf{r}^{'}$ by the mask pooling operation,

$$\mathbf{r}^{'}_i = \texttt{AvgPool}(\mathbf{x}^{\mathrm{e}}, \mathbf{m}_i). \tag{3.7}$$

Finally, we compute the cosine similarity loss for each paired semantic embedding $\mathbf{r}^{'}_i$ and text embedding $\mathbf{t}_k$:

$$\mathcal{L}_{\text{align}}(\mathbf{r}^{'}_i, \mathbf{t}_k) = 1 - \frac{\mathbf{r}^{'}_i \cdot \mathbf{t}_k}{\|\mathbf{r}^{'}_i\|_2 \|\mathbf{t}_k\|_2}. \tag{3.8}$$

Optimizing that, the predicted masks can be effectively associated with entities within the CLIP embedding space, thereby enabling open-vocabulary segmentation.

CLIP Visual & Text Encoder. In general, the CLIP encoder can be any architecture. Motivated by the scalability of different input resolutions, we employ a ConvNext-based CLIP model to serve as both the image and text encoders. The image encoder is configured as a ConvNext-Large model, comprising four stages. Each stage contains a different number of blocks: 3 in the first, 3 in the second, 27 in the third, and 3 in the fourth. In contrast, the text encoder is structured as a 16-layer transformer, each layer being 768 units wide and featuring 12 attention heads. We harness the power of multi-scale features extracted by the

image encoder. These features are represented as feature maps of varying widths and scales: a 192-wide feature map downscaled by a factor of 4, a 384-wide map downscaled by 8, a 768-wide map downscaled by 16, and a 1536-wide map downscaled by 32.

Open-vocabulary inference. During inference, we first equip the test categories $\mathbf{C}_{\text{test}}$ with prompt engineering [115], and their corresponding text embeddings are extracted through the CLIP text encoder for open-vocabulary segmentation. For each input image, a uniform grid of points serves as visual prompts to generate initial binary mask proposals. Subsequently, we apply Algorithm 1 to filter these proposals and obtain a set of non-redundant masks. Finally, for each retained mask, its category is predicted by calculating the cosine similarity between its visual embedding and all text embeddings, assigning the label corresponding to the highest similarity score.

Algorithm 1 Mask Non-Maximum Suppression

Require: Candidate masks $\mathbf{m} = \{m_1, m_2, \ldots, m_n\}$; quality scores $\mathbf{p} = \{p_1, p_2, \ldots, p_n\}$; overlap threshold τ_o; coverage threshold τ_c.
Ensure: Non-redundant masks $\mathbf{m}'$.

```
Initialize m' ← ∅.
Sort m in descending order according to p.
while m is not empty do
    Select m_i with the highest score in m.
    Add m_i to m' and remove it from m.
    for each m_j ∈ m do
        if IoU(m_i, m_j) > τ_o and Cover(m_i, m_j) > τ_c then
            Remove m_j from m.
        end if
    end for
end while
return m'.
```

3.4 Experiments

3.4.1 Implementation details

Datasets. During training, we randomly sample the 50% subset from the SA-1B dataset [9], which contains ∼5 million images and ∼0.5 billion masks. Although this supervision provides

diverse binary masks, it lacks the semantic class for each mask. Besides, we collect about 200k images [126] and adopt a vision-language large model [33] to refine captions, forming image-text pairs. Afterward, we extract entities from these text descriptions.

Training configuration. We adopt our visual and text encoders as our ConvNext-Large CLIP model [38], [127] from OpenCLIP [128]. For mask segmentation training, we mainly follow [28] and adopt a similar training recipe and losses without any particular design. The input size of the image is 1024×1024, and we set the batch size to 64. The model is optimized with AdamW [129], [130], where the learning rate is set to 1×10^{-4} and the weight decay is set to 5×10^{-2}. We train the model for 200k iterations and update the learning rate via the multi-step decay schedule. The parameters λ_{bce}, λ_{dice}, and λ_{iou} are set to 2, 1, and 5, respectively. Afterwards, during the training process of mask-entity alignment, we adopt a 20×20 point grid as the input visual prompt. The batch size and learning rate are set to 32 and 2×10^{-5}, respectively. We train this stage for 30k iterations. We conduct all training on 8 NVIDIA H100 80GB, where 109 hours for the first stage, and 7 hours for the second stage.

Evaluation & metrics. We evaluate our model mainly on four tasks: open-vocabulary semantic segmentation, open-vocabulary panoptic segmentation, point-prompt segmentation, and box-prompt segmentation. Following previous work [80], we adopt prompt engineering from [115], [121] and prompt templates from [131], [132]. For open-vocabulary semantic segmentation, we zero-shot evaluate the model on the COCO [72], ADE20K [133], and PASCAL [134] datasets. Following previous work [80], we adopt the prompt engineering from [115], [121], [131], [132]. The open-vocabulary semantic segmentation results are evaluated with the mean Intersection-over-Union (mIoU). We evaluated the model in COCO, ADE20K and Cityscapes [135] datasets for open-vocabulary panoptic segmentation. We report the panoptic quality (PQ), semantic quality (SQ), and recognition quality (RQ) for open-vocabulary panoptic segmentation. For point-prompt and box-prompt segmentation, from the oracle perspective, we report the 1-pt IoU and 1-box IoU on a wide range of datasets. "Oracle" indicates that we select the output mask with the max IoU by calculating the IoU between the prediction and the ground truth mask.

TABLE 3.1. **Open-vocabulary semantic segmentation performance.** We mainly compare with the fully-supervised and weakly-supervised methods. "COCO S.", "COCO P.", and "COCO C." denote the COCO stuff, panoptic, and caption datasets. "O365" denotes the Object 365 dataset. "M. 41M" denotes the merged 41M image dataset. We report mIoU for all datasets.

Method	Training Data	A-847	PC-459	A-150	PC-59	PAS-21	PAS-20	Citys
		mIoU (%)						
Fully-supervised method (image-text-mask pair)								
SimBaseline [25]	COCO S.	-	-	15.3	-	74.5	-	-
ZegFormer [30]	COCO S.	-	-	16.4	-	73.3	-	-
LSeg [29]	COCO S.	3.8	7.8	18.0	46.5	-	-	-
OVSeg [26]	COCO S.	9.0	12.4	29.6	55.7	-	94.5	-
SAN [117]	COCO S.	13.7	17.1	33.3	60.2	-	95.5	-
OpenSeg [115]	COCO P. + C.	6.3	9.0	21.1	42.1	-	-	-
ODISE [121]	COCO P.	11.1	14.5	29.9	57.3	84.6	-	-
X-Decoder [77]	COCO P. + C.	-	-	25.0	-	-	-	47.3
MaskCLIP [120]	COCO P.	8.2	10.0	23.7	45.9	-	-	-
FC-CLIP [80]	COCO P.	14.8	18.2	34.1	58.4	81.8	95.4	56.2
Weakly-supervised method (image-text pair or image-text & image-mask pair)								
GroupViT [81]	GCC + YFCC	4.3	4.9	10.4	23.4	52.3	79.7	18.5
ReCo [136]	ImageNet-1K	-	-	11.2	22.3	25.1	57.7	21.6
TCL [82]	GCC	-	-	17.1	33.9	55.0	83.2	24.3
OVSeg [83]	CC4M	-	-	5.6	-	53.8	-	-
SegCLIP [122]	CC3M + COCO C.	-	-	8.7	-	52.6	-	-
CLIPpy [137]	HQITP-134M	-	-	13.5	-	52.2	-	-
MixReorg [138]	CC12M	-	-	10.1	25.4	50.5	-	-
SAM-CLIP [139]	M. 41M	-	-	17.1	29.2	60.6	-	-
Baseline (Ours)	50% SA1B	12.1	13.3	27.0	40.6	60.3	90.6	27.2
Unpair-Seg (Ours)	50% SA1B + M. 200K	**14.6**	**19.6**	**32.4**	**52.4**	**75.8**	**92.8**	**44.7**

TABLE 3.2. **Open-vocabulary panoptic segmentation performance.** We mainly compare with the fully-supervised and unsupervised methods. "COCO P.", "COCO" and "IN 1K" denote the COCO panoptic, COCO image and ImageNet-1K datasets, respectively. We report PQ, SQ and RQ for all datasets. Fully-supervised method trained on the "COCO P." datasets, so we show the results in grey.

Method	Training Data	COCO			ADE20K			Cityscapes		
		PQ	SQ	RQ	PQ	SQ	RQ	PQ	SQ	RQ
Fully-supervised method (image-text-mask pair)										
MaskCLIP [120]	COCO P.	30.9	-	-	15.1	70.5	19.2	-	-	-
ODISE [121]	COCO P.	55.4	-	-	22.6	-	-	23.9	75.3	29.0
FC-CLIP [80]	COCO P.	54.4	83.0	64.8	26.8	71.6	32.3	44.0	75.4	53.6
OPSNet [140]	COCO P.	57.9	84.1	68.2	19.0	52.4	23.0	41.5	67.5	50.0
Unsupervised method (Unlabelled image)										
CutLER+STEGO [141]	IN 1K + COCO	12.4	64.9	15.5	-	-	-	12.4	36.1	15.2
U2Seg [142]	IN 1K + COCO	16.1	71.1	19.9	-	-	-	17.6	52.7	21.7
Weakly-supervised method (image-text & image-mask pair)										
Baseline (Ours)	50% SA1B	24.4	79.6	30.0	19.1	78.1	23.7	17.3	71.4	22.0
Unpair-Seg (Ours)	50% SA1B + M. 200K	**29.0**	**80.1**	**35.4**	**22.8**	**74.3**	**28.2**	**26.2**	**70.7**	**33.2**

TABLE 3.3. **Promptable segmentation performance.** We compare our method with SAM [9]. Given point or box prompts, we select the output masks with max IoU by calculating the IoU with the ground-truth masks. We report 1-pt IoU on the COCO and ADE20K panoptic segmentation, and 1-box IoU on the COCO and ADE20K instance segmentation datasets, respectively.

Method	Param	Training Data	COCO		ADE20K	
			1-pt IoU (%)	1-box IoU (%)	1-pt IoU (%)	1-box IoU (%)
SAM-Base	93M	100% SA-1B	42.7	72.8	41.6	74.7
SAM-Large	312M	100% SA-1B	64.3	77.9	61.1	77.7
SAM-Huge	641M	100% SA-1B	66.7	**78.0**	63.3	**78.3**
Unpair-Seg (Ours)	221M	50% SA-1B	**80.8**	76.9	**76.9**	77.3

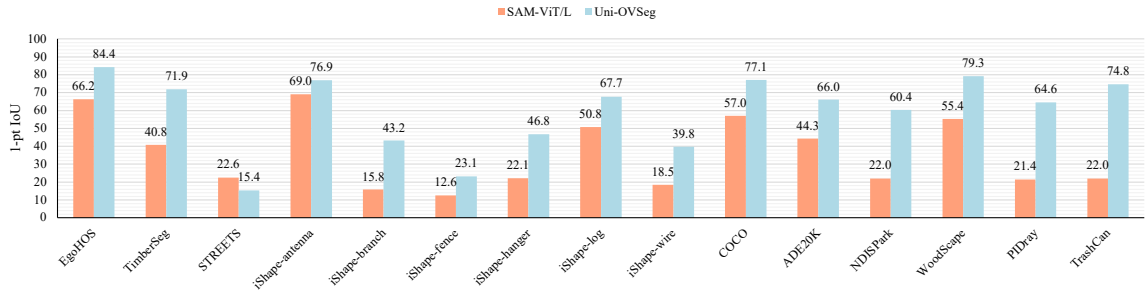


FIGURE 3.6. **Point-prompt segmentation performance.** We compare our method with SAM-Large [9]. Given a 20×20 point grid as visual prompt, we select the output masks with max IoU by calculating the IoU with the ground-truth masks. We report 1-pt IoU for all datasets.

3.4.2 Main results

Open-vocabulary semantic segmentation. To highlight the significant reduction in the performance gap between fully-supervised and weakly-supervised methods, we perform a comprehensive comparison across various datasets, including ADE20K (encompassing both 150 and 847 class variants) [133], PASCAL Context (459 and 59 class variants) [134], PASCAL VOC (with 20 and 21 class categories) [134], and Cityscapes [135]. As shown in Table 3.1, Unpair-Seg outperforms weakly-supervised methods across different datasets by utilizing independent image-mask and image-text pairs. In particular, Unpair-Seg performs impressively in the challenging PASCAL Context-459 and ADE20K-847 datasets, surpassing the state-of-the-art fully-supervised method FC-CLIP [80]. Upon further analysis, we observed that FC-CLIP achieves significantly higher accuracy for in-vocabulary classes (corresponding to the COCO dataset) than out-of-vocabulary ones on the PASCAL Context-459 and ADE20K-847 datasets. Leveraging the diversity of image-text pairs, our method demonstrates superior capability in categorizing general semantic classes.

Open-vocabulary panoptic segmentation. Existing weakly-supervised methods only utilize text supervision, which makes it challenging to distinguish different instances with the same semantic class. To the best of our knowledge, Unpair-Seg is the first attempt to tackle the difficult task of open-vocabulary panoptic segmentation with weak supervision. Due to the absence of benchmarks, we evaluate our method on the COCO, ADE20K, and Cityscapes

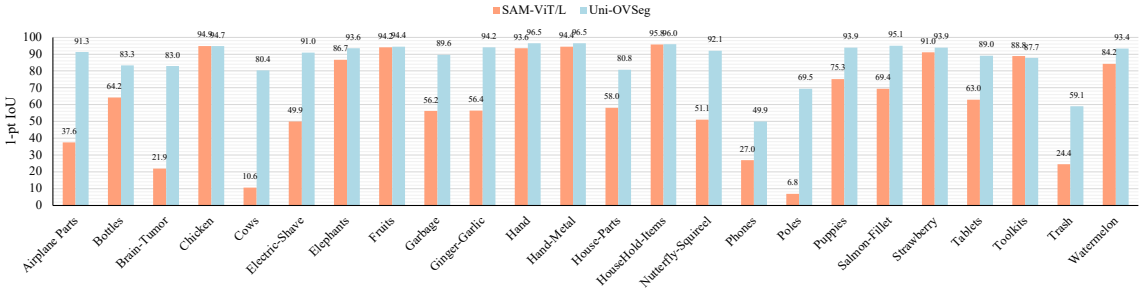


FIGURE 3.7. **Point-prompt segmentation performance on the SegInW dataset.** We compare our method with SAM-Large [9]. Given a 20×20 point grid as a visual prompt, we select the output masks with max IoU by calculating the IoU with the ground-truth masks. We report 1-pt IoU for all datasets.

datasets and compared it comprehensively with fully-supervised and unsupervised methods in Table 3.2. Compared to unsupervised methods, our approach outperforms U2Seg [142] by 12.9% PQ, 9.0% SQ, and 15.5% RQ on the COCO panoptic segmentation task. Even more impressively, Unpair-Seg exceeds specific fully-supervised methods. For instance, we achieve a noticeable gain of 3.8% PQ than OPSNet [140] on the ADE20K dataset and a remarkable improvement of 2.3% PQ than ODISE [121] on the Cityscapes panoptic segmentation. These results show that our method effectively captures fine-grained spatial relations through weak supervision, thereby enhancing the practical usefulness of weakly-supervised open-vocabulary segmentation.

Promptable segmentation. Following SAM [9], we evaluate the segmentation of an object from an input point, which is challenging as one point can refer to objects at multiple levels of detail. To provide visual cues, we have implemented a point grid as the interactive point prompt and use actual bounding boxes as box prompts. Specifically, For the point prompt, we adopt a uniform point grid $h \times w$ as input prompts (e.g., 20×20). For the box prompt, we use ground-truth bounding boxes as input prompts. 1-pt IoU denotes the oracle performance of one point by evaluating the intersection-over-union (IoU) of the predicted masks that best match ground truth. 1-box IoU denotes is similar to 1-pt IoU. We compare our method with three variants of SAM, as shown in Table 3.3. By using the proposed many-to-many matching, our method can predict multi-granularity masks when employing point prompts. In particular,

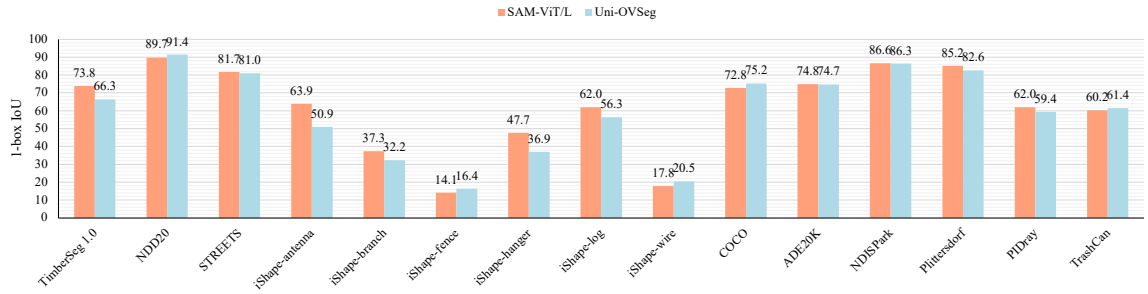


FIGURE 3.8. **Box-prompt segmentation performance.** We compare our method with SAM-Large [9]. Given a ground-truth box as the visual prompt, we select the output masks with max IoU by calculating the IoU with the ground-truth masks. We report 1-pt IoU for all datasets.

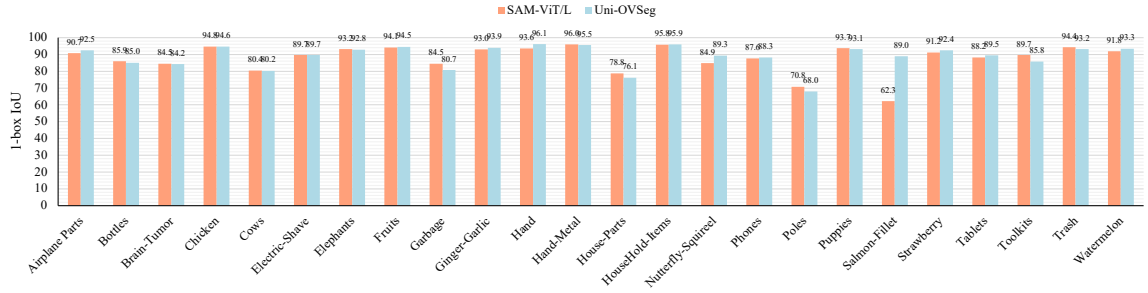


FIGURE 3.9. **Box-prompt segmentation performance on the SegInW dataset.** We compare our method with SAM-Large [9]. Given a ground-truth box as the visual prompt, we select the output masks with max IoU by calculating the IoU with the ground-truth masks. We report 1-box IoU for all datasets.

Unpair-Seg significantly outperforms SAM-Large in terms of 1-point IoU by 16.5% and 15.8% on the COCO and ADE20K panoptic segmentation datasets, respectively. Some visualizations are presented in Figure 3.10 and 3.13. Furthermore, we illustrate more experimental results on point-prompt segmentation in Figure 3.6 and 3.7, box-prompt segmentation in Figure 3.8 and 3.9. The details of the evaluated datasets are shown in Table 3.4. The iShape dataset has 6 subsets, including antenna, branch, fence, hanger, log, and wire.

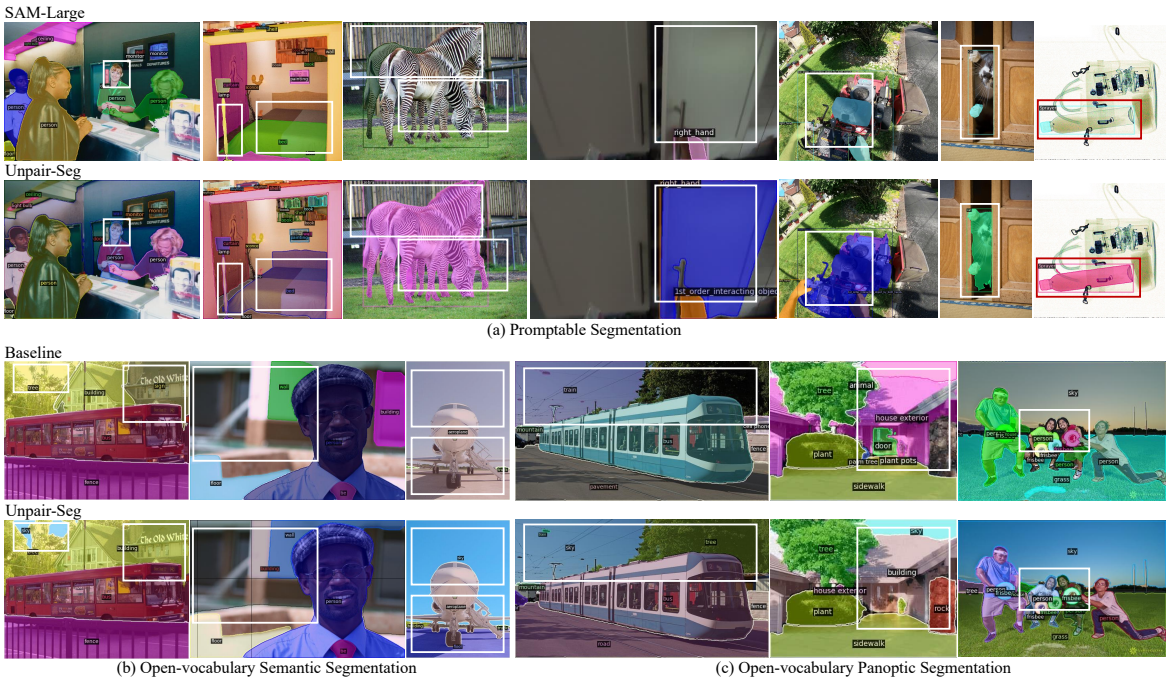


FIGURE 3.10. **Visualization.** We show prediction results on three tasks: promptable, open-vocabulary semantic, and open-vocabulary panoptic segmentation. The results are best viewed in color.

3.4.3 Ablation study

We conducted a thorough ablation study to evaluate the impact of key components in our Unpair-Seg framework, identifying each one's contribution toward open-vocabulary segmentation and mask-entity alignment. Compared to the simple baseline, Unpair-Seg achieved significant improvements of 2.5%, 6.3%, and 17.5% mIoU on the ADE20k-847, PASCAL Context-459, and Cityscapes semantic segmentation, and 4.6% and 3.7% PQ on the COCO and ADE20K panoptic segmentation, as shown in Table 3.1 and Table 3.2. These results demonstrate the effectiveness of our method in aligning objects in images and entities in text descriptions, bridging the image-level and pixel-level embedding alignment in the CLIP space. Qualitatively, Unpair-Seg achieves visibly superior segmentation quality, as shown in Figure 3.11 and 3.12, where it successfully delineates complex objects and instances compared to the baseline.

TABLE 3.4. Segmentation datasets used to evaluate promptable segmentation with point and box prompts. The 11 datasets cover a broad range of domains, which are illustrated in "image type".

Dataset	Image type	Mask type	Description
Egocentric Hand-Object Segmentation (EgoHOS) [143]	Egocentric	Instance	Fine-grained egocentric hand-object segmentation dataset. Dataset contains mask annotations for existing datasets.
TimberSeg 1.0 (TimberSeg) [144]	Logs	Instance	Segmentation masks of individual logs in piles of timber in various environments and conditions. Images are taken from an operator's point-of-view.
STREETS [145]	Traffic camera	Instance	Segmentation masks of cars in traffic camera footage.
iShape [146]	Irregular shapes	Instance	Segmentation masks of irregular shapes like antennas, logs, shapes fences, and hangers.
COCO [72]	Scenes	Instance	Segmentation masks of complex everyday scenes containing common objects in their natural context.
ADE20K [133]	Scenes	Instance	Object and part segmentation masks for images from SUN and Places datasets.
Night and Day Instance Segmented Park (NDISPart) [147]	Parking lots	Instance	Images of parking lots from video footage taken at day and night during different weather conditions and camera angles for vehicle segmentation.
WoodScape [148]	Fisheye driving	Instance	Fisheye driving dataset with segmentation masks. Images are driving taken from four surround-view cameras.
PIDray [149]	X-ray	Instance	Segmentation masks of prohibited items in X-ray images of baggage.
TrashCan [150]	Underwater	Instance	Segmentation masks of trash in images taken by underwater ROVs. Images are sourced from the J-EDI dataset.
Segmentation in the wild (SegInW) [77]	Multiple domain	Instance	This dataset consists of 25 free public Segmentation datasets, crowd-sourced on roboflow.com.

TABLE 3.5. **Vision-language large model.** "*LLaVA.*" and "*InternLM-VL.*" denote the LLaVA-v1.5-7B and InternLM-XComposer2-VL-7B models.

Method	**COCO**	**PC-59**
	PQ (%)	mIoU (%)
LLaVA.	25.4	42.6
InternLM-VL.	**29.0**	**52.4**

Furthermore, we explored the impact of different vision-language large models on the quality of image-text pairs. Our ablation study on two models, LLaVA [151] and InternLM-VL [33],

TABLE 3.6. **Design choice of the feature adapter.** "*Cross Attn.*" and "*Low Rank.*" denote the cross attention-based and low rank-base feature adapters.

Method	COCO	PC-59
	PQ (%)	mIoU (%)
Cross Attn.	27.1	47.4
Low Rank.	**29.0**	**52.4**

TABLE 3.7. **Mask-entity matching strategy.** "*Greedy.*" and "*Bipartite.*" denote the greedy and bipartite match strategies.

Method	COCO	PC-59
	PQ (%)	mIoU (%)
Greedy. (1 sigma)	26.1	43.7
Greedy. (2 sigma)	26.0	43.6
Bipartite.	**29.0**	**52.4**

TABLE 3.8. **Multi-scale ensemble.** We ensemble multi-scale visual embeddings extracted by CLIP to stabilize the matching process.

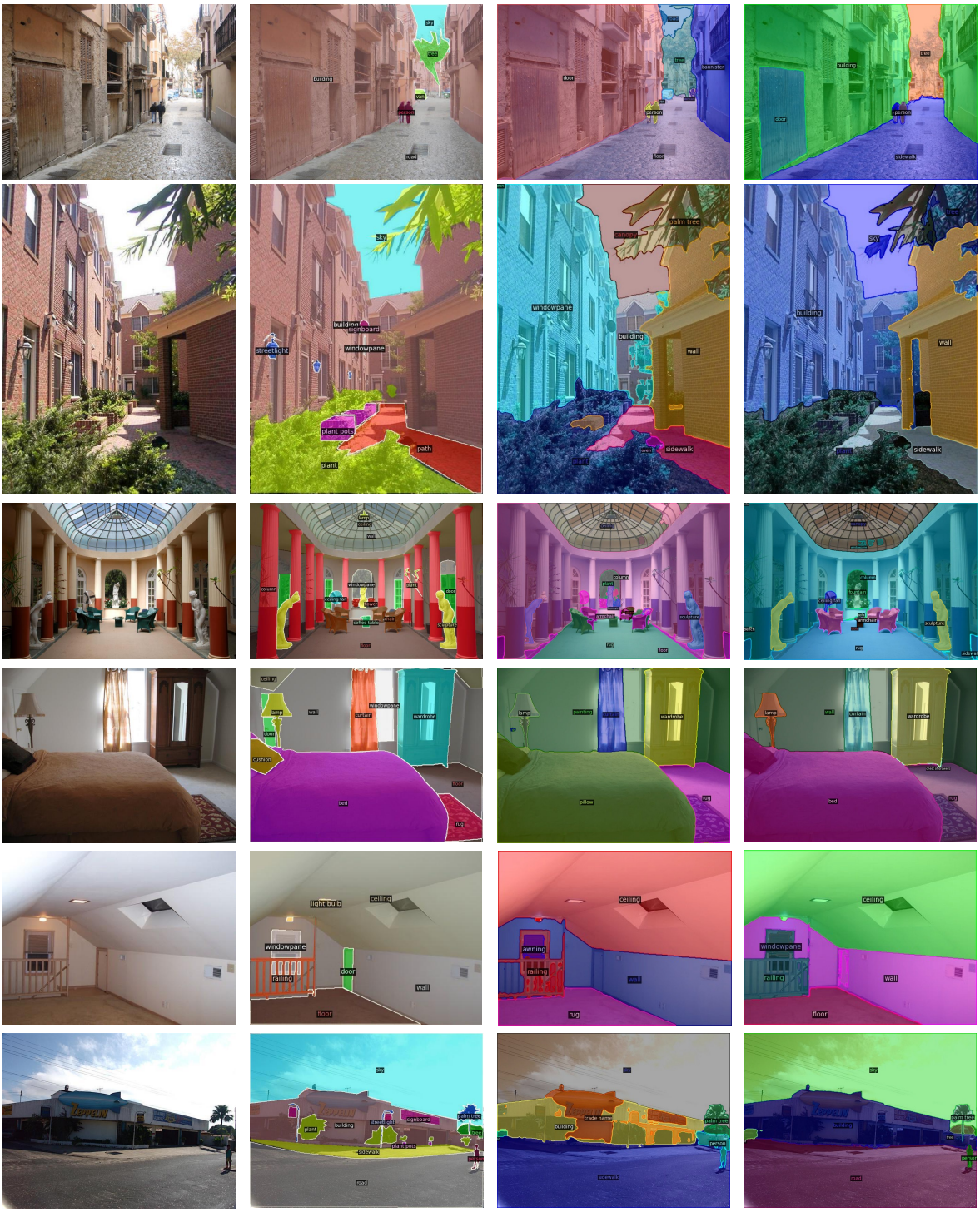
Size	COCO	PC-59
	PQ (%)	mIoU (%)
896	27.4	49.0
896, 1024	28.2	50.0
896, 1024, 1152	**29.0**	**52.4**

detailed in Table 3.5, revealed their influence on the performance of the model. High-quality captions were found to help the model learn accurate mask-entity alignment.

During the mask-entity alignment, we use a feature adapter to enhance regional embeddings, followed by alignment with text embedding. In Table 3.6, we design two alternative modules to aggregate multi-scale visual information and improve the quality of regional embedding. A low-rank feature adapter that is adopted in our framework achieves better results.

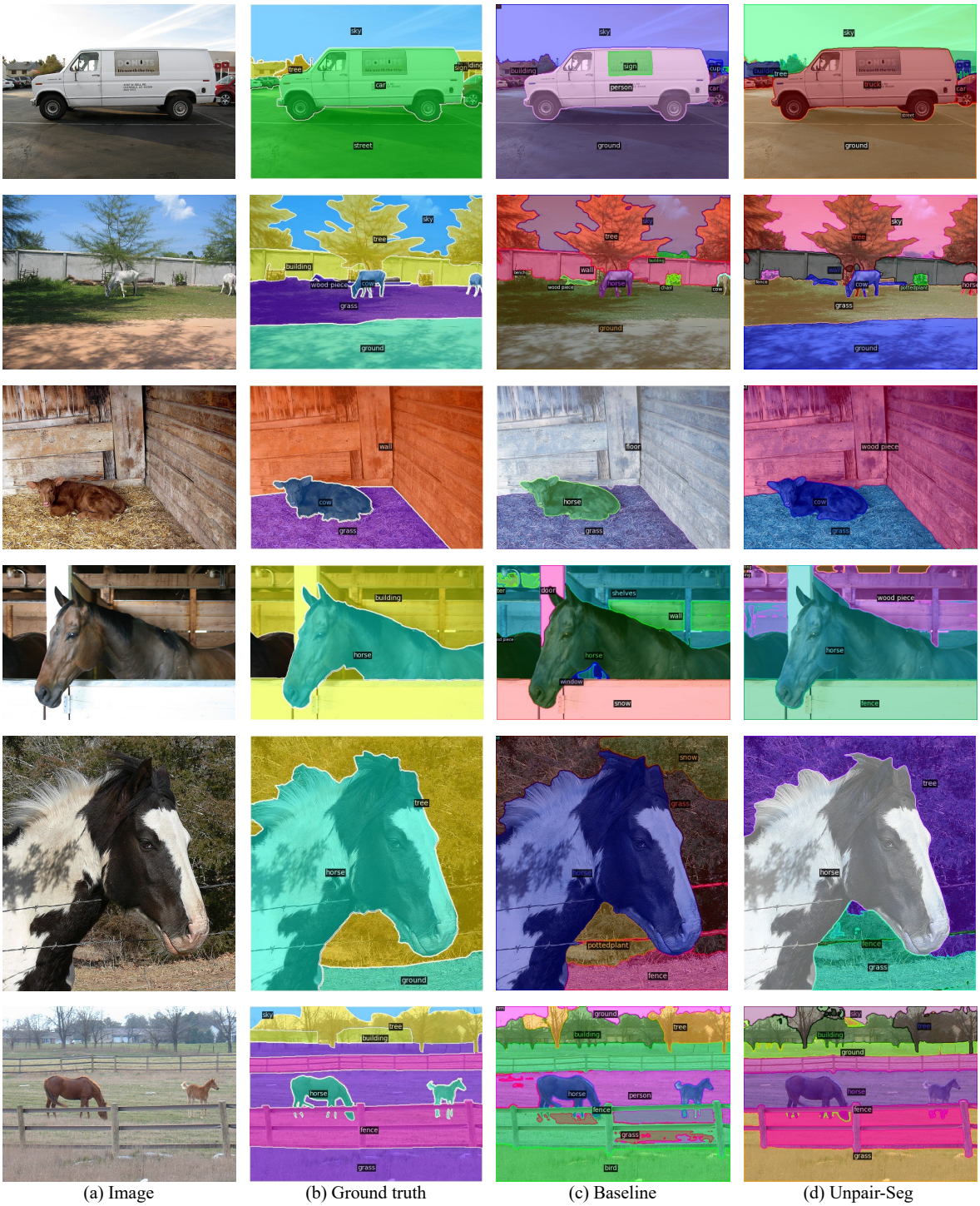
As illustrated in Table 3.7, we design two mask-entity matching strategies. The greedy method assigns an entity to each predicted mask based on the maximum confidence. The bipartite method optimizes an assignment plan to achieve the global minimum cost, which obtains better results than the greedy method.

Finally, Given the varied object scales, we employ a multi-scale ensemble strategy that extracts CLIP visual embeddings from images at multiple resolutions, combining them into a unified



(a) Image (b) Ground truth (c) Baseline (d) Unpair-Seg

FIGURE 3.11. Visualisation of open-vocabulary segmentation between the baseline and Unpair-Seg.



(a) Image (b) Ground truth (c) Baseline (d) Unpair-Seg

FIGURE 3.12. Visualisation of open-vocabulary segmentation between the baseline and Unpair-Seg.

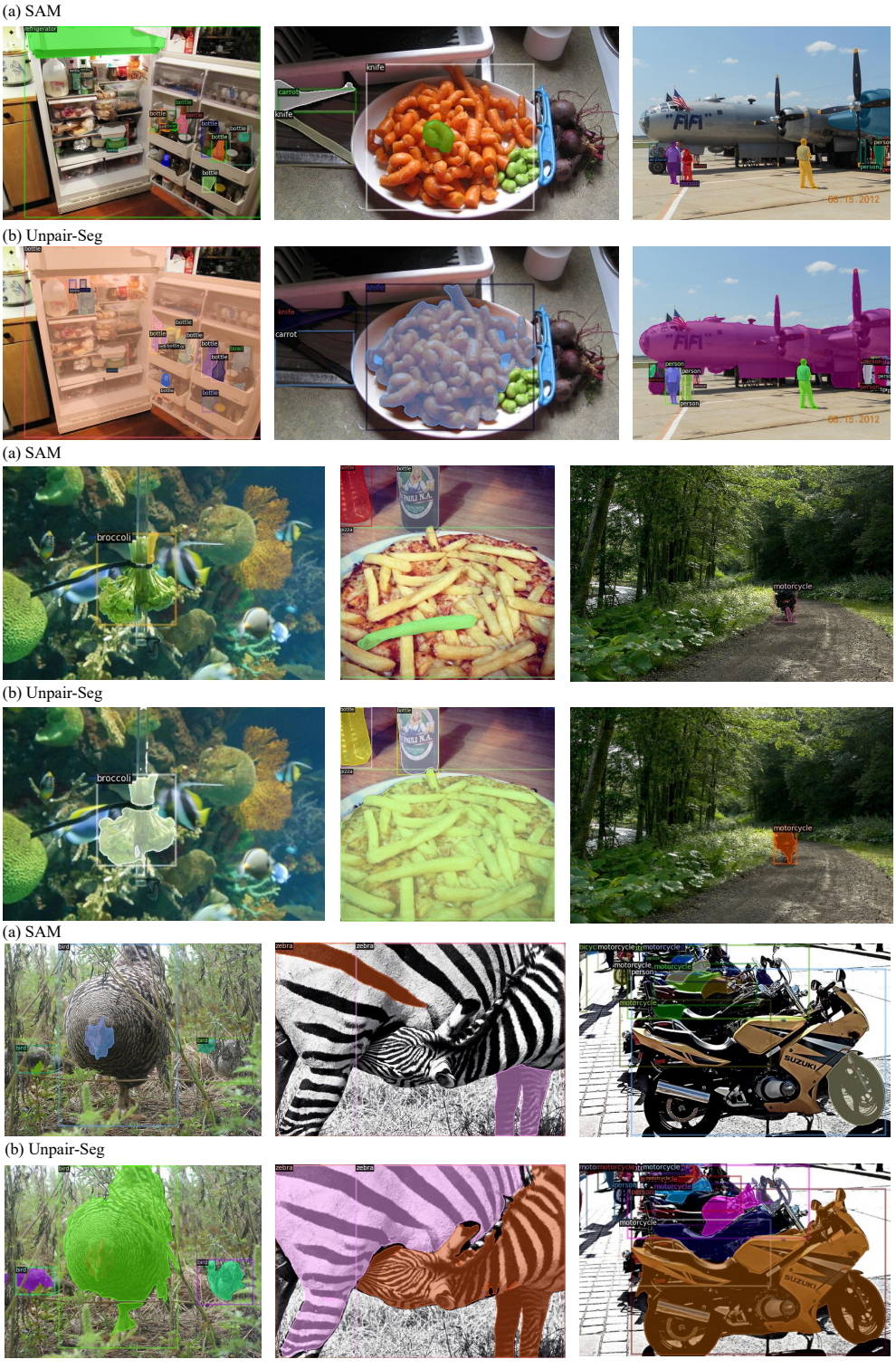


FIGURE 3.13. Visualisation of promptable segmentation between SAM-Large and Unpair-Seg.

embedding for improved mask-entity pairing. As shown in Table 3.8, we use default sizes of 869×896, 1024×1024, and 1152×1152 to capture diverse spatial details effectively.

3.5 Conclusion

In this paper, we introduce Unpair-Seg, a weakly-supervised open-vocabulary segmentation framework designed to use unpaired mask-text supervision, which reduces the need for labor-intensive annotations. To mitigate inherent noise between mask and text entities, we employ a vision-language large model for precise entity extraction and a multi-scale matching strategy for stable mask-entity alignment. By exploiting confident mask-text entity pairs, Unpair-Seg significantly outperforms previous weakly-supervised methods across multiple segmentation tasks. This significant improvement brings us closer to fully-supervised methods in real-world scenarios. However, it is important to acknowledge that the framework's performance may be influenced by the quality of image-text pairs and initial mask predictions in complex or ambiguous scenes. These aspects will be the focus of our future work.

CHAPTER 4

LaVin-DiT: Large Vision Diffusion Transformer

4.1 Introduction

Large language models (LLMs) like GPT [11] and LLaMA [13] have rapidly gained widespread attention and transformed the field, demonstrating the strong capability to handle a wide range of language tasks within a unified framework [12]. This breakthrough of integrating diverse language tasks into a single large model has sparked momentum to develop similar large models for computer vision. The potential to create large vision models (LVMs) capable of generalizing across multiple vision tasks represents a promising step toward a more versatile, scalable, and efficient approach to vision-based AI [32], [152], [153], [154], [155].

However, constructing LVMs presents greater complexity than LLMs due to the inherently diverse and high-dimensional nature of vision data, as well as the need to handle variations in scale, perspective, and lighting across tasks [1], [2], [6], [7], [8], [156]. To handle the problem, recent work [32] has developed a sequential modeling method that learns from purely vision data by representing images, videos, and annotations in a unified "visual sentence" format. This method enables the model to predict sequential vision tokens from a vast dataset, entirely independent of language-based inputs (see Figure 4.1(a)). Although this method has shown promising results in diverse vision tasks, it faces two primary challenges. Specifically, the first issue concerns the efficiency limitations inherent in autoregressive sequence modeling [157], as it demands token-by-token prediction, which is computationally intensive for high-dimensional vision data [158], [159]. The second issue is the disruption of

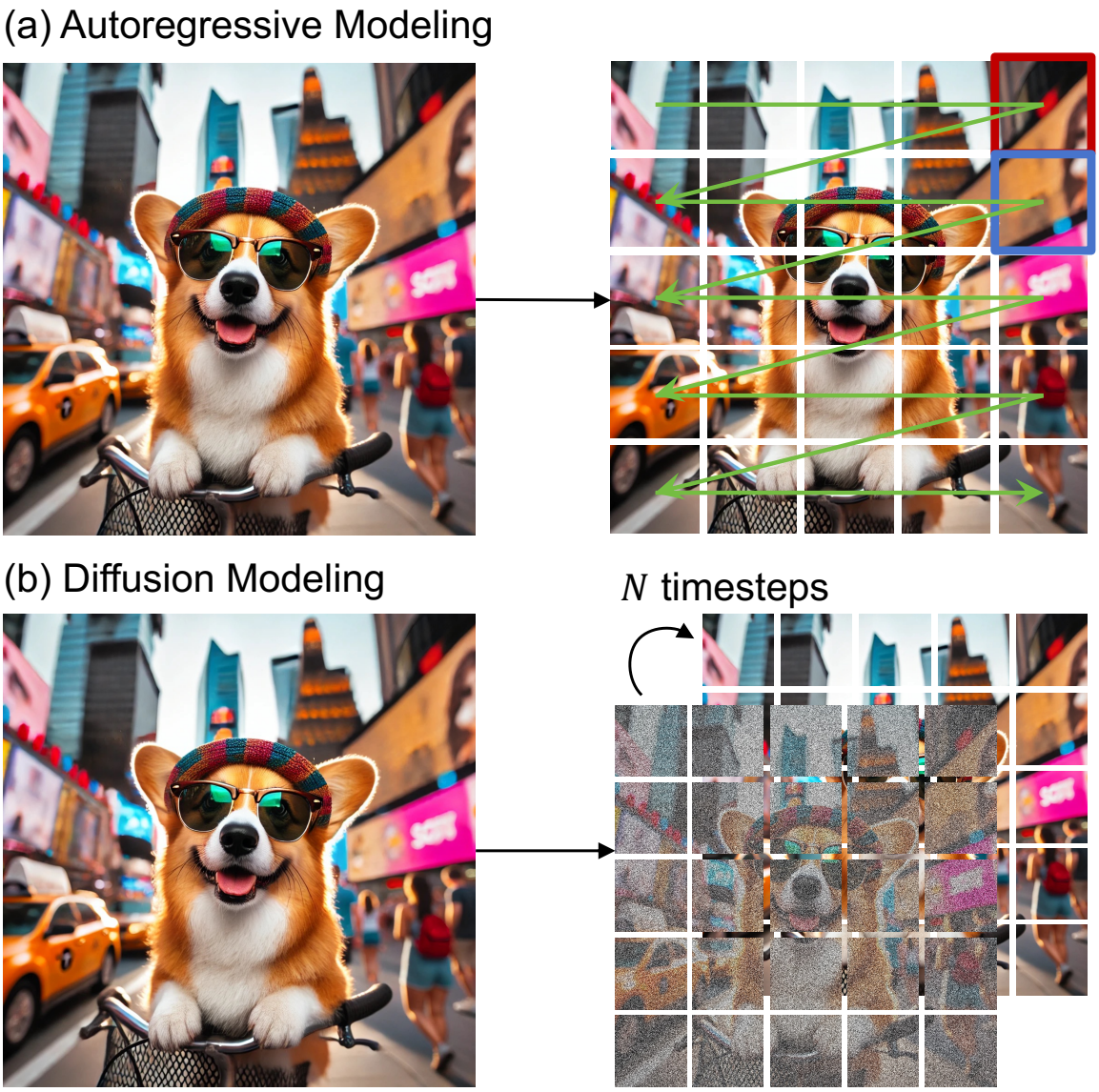


FIGURE 4.1. **Comparison of autoregressive and diffusion modeling.** (a) In **autoregressive modeling**, visual data is divided into a sequence of patches and transformed into a one-dimensional sequence. The model then predicts each token sequentially from left to right and top to bottom, which is computationally intensive for high-dimensional visual data. Besides, tokens marked in red and blue illustrate disrupted spatial dependencies, highlighting the limitations of preserving spatial coherence. (b) In contrast, **diffusion modeling** denoises all tokens in parallel across N timesteps, significantly improving computational efficiency and preserving essential spatial structures crucial for high-performance vision tasks.

spatial coherence when converting vision data into a sequential format, which compromises the preservation of spatial dependencies crucial for performance in vision tasks [160].

In this paper, we introduce a large vision diffusion transformer (LaVin-DiT) to advance the development of next-generation LVMs. LaVin-DiT enjoys better computational efficiency and effectively preserves spatial relationships within vision data, thereby achieving superior performance across diverse vision tasks (see Figure 4.1(b)). Technically, to tackle the high-dimensional nature of vision data, we introduce a spatial-temporal variational autoencoder [161] that encodes data (i.e., image and video) into a continuous latent space, allowing compact representation while preserving essential spatial and temporal features. This reduces computational demands and improves efficiency without sacrificing the model's ability to capture complex patterns. Besides, for generative modeling, we augment an existing diffusion transformer and propose a joint diffusion transformer with full-sequence joint attention. This module synthesizes visual outputs through parallel denoising steps, effectively reducing sequential dependencies to enhance processing efficiency while maintaining the spatial coherence essential for vision tasks. Moreover, to support unified multi-task training [162], we incorporate in-context learning [11], [163], [164], [165], [166], where input-target pairs guide the diffusion transformer in aligning outputs with specific tasks. During inference, LaVin-DiT leverages task-specific context sets and test data as queries to adapt to various tasks without fine-tuning. This capability enables LaVin-DiT to achieve robust generalization across diverse tasks, leading to a versatile solution for complex vision applications.

We conduct comprehensive experiments to demonstrate the superiority of LaVin-DiT. Results show that LaVin-DiT significantly outperforms the strongest baseline LVM [32] on various vision benchmarks. For instance, it achieves a 24 lower AbsRel in NYU-v2 depth estimation [156]. Besides, LaVin-DiT offers $1.7 \sim 2.3\times$ faster inference speeds than LVM [32] across resolutions ranging from 256×256 to 512×512. Evaluations across different model sizes showcase the scalability and fast convergence of LaVin-DiT across multiple complex vision tasks. Finally, we observe that increasing the task context length consistently enhances performance across a diverse array of tasks. These promising results establish LaVin-DiT

as a highly scalable, efficient, and versatile model, showing a new pathway for large vision foundation models.

4.2 Related Work

Large vision model. Developing a universal framework for diverse tasks across information sources is a longstanding goal in deep learning [153]. Natural language processing has achieved this with ChatGPT[1] that demonstrates versatility across numerous language tasks, e.g., summarization, reasoning, and translation. In contrast, computer vision is relatively lacking in universal frameworks, largely due to the complexity and diversity of visual data and tasks. Existing methods of universal vision frameworks generally follow two main pathways: image-resembling generation [152], [154], [167] and sequential modeling [32], [168].

The image-resembling generation methods reformulate visual tasks as image generation problems, which allows models to handle dense visual predictions through inpainting and reconstruction tasks [152]. For instance, Painter [154] formulates dense prediction tasks as masked image inpainting, demonstrating in-context capabilities across multiple vision tasks. By leveraging pre-trained diffusion models [158], [169], several methods [170], [171], [172], [173] utilize visual or textual instruction to guide generation and enhance adaptability across various tasks. The sequential modeling methods are largely inspired by breakthroughs in large language models and apply the sequence-to-sequence framework to visual data [13]. For these methods, visual data is typically quantized into sequences of discrete tokens [174]. The model is optimized through next-token prediction [11]. Recently, [32] introduce a framework that extends this concept to vision without relying on linguistic data, which treats visual data as a "visual sentence". By representing images and videos as one-dimensional sequences, this method [32] enables a unified transformer that can tackle image and video tasks within a single framework, expanding the scope of sequential modeling in computer vision.

In this paper, from the respective of image-resembling generation, we propose a universal diffusion framework with a transformer architecture tailored for visual data, which preserves

[1] https://openai.com/index/chatgpt/

spatial-temporal structure and minimizes information loss. Trained exclusively on visual data, our flexible framework unifies image and video tasks, advancing toward a generalist model in computer vision.

Diffusion transformer. By resorting to vision transformer (ViT) [175], [176], [177], recent advancements [178], [179], [180], [181], [182], [183], [184] in generative modeling achieves significant improvements in scalability and performance for both image [160], [185], [186], [187], [188], [189] and video generation [190], [191], [192], [193]. Among these advancements, U-ViT [179] treats all inputs as tokens by combining transformer blocks with a U-net architecture. DiT [178] employs a straightforward and non-hierarchical transformer structure, showcasing the scalability and versatility of diffusion transformers. MDT [181] and MaskDiT [180] enhance the training efficiency of DiT by using a masking strategy [101]. Subsequently, Stable Diffusion 3 [188] introduces a novel transformer-based architecture for text-to-image generation, which enables bidirectional interaction between image and text. Furthermore, diffusion transformers demonstrate robust capabilities for spatial-temporal modeling in video generation [194]. Previous methods [183], [193] utilize separate spatial and temporal attention mechanisms to reduce intensive computational costs. Besides, recent works [10], [190], [191], [192], [195] have proposed using 3D full attention to capture spatial-temporal information, ensuring consistency for large-moving objects. While diffusion transformers have shown impressive potential in visual content generation, their capability to serve as a large vision model unifying multiple vision tasks remains underexplored. In this paper, we introduce a new joint diffusion transformer with full-sequence joint attention that effectively integrates diverse vision tasks into a cohesive framework, elevating diffusion transformers to a new level of unified understanding and generation.

In-context learning. In-context learning is initially conceptualized with GPT-3 [11]. It has revolutionized the approach to task-specific model training by allowing models to infer and execute tasks based directly on contextual examples provided in prompts [196]. This paradigm shift enables models to perform complex reasoning and novel pattern recognition without direct training on those specific tasks. Extending beyond text, Flamingo [109] incorporates visual inputs and broadens in-context learning to multi-modal tasks such as image captioning,

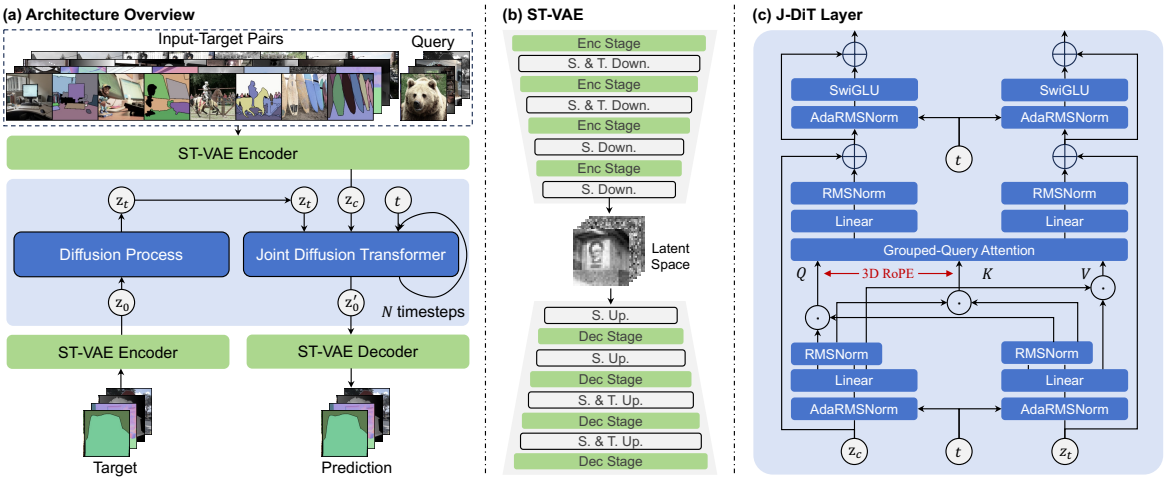


FIGURE 4.2. **Overview of Large Vision Diffusion Model (LaVin-DiT).** As shown in panel (a), the model initially compresses input visual data from the pixel space into a latent space, where multiple input-target pairs serve as the task context. A target is perturbed with Gaussian noise through a diffusion process. Guided by the task context and query, the Joint Diffusion Transformer (J-DiT) iteratively denoises this noisy target over N timesteps to recover a clean latent representation. The prediction is then generated via the ST-VAE decoder. Panels (b) and (c) provide architectural details of the ST-VAE and J-DiT, respectively. "Down." and "Up." indicate the downsampling and upsampling, respectively. Concatenation is represented by $\odot$.

visual question answering, and optical character recognition. This demonstrates the model's ability to integrate and interpret both textual and visual data, enhancing its application across different domains. In the realm of computer vision, the concept of in-context learning is explored through methods such as visual prompting [152], which infers tasks directly from concatenated image examples and queries. In this paper, we build on this idea. A set of examples are sampled as task definitions and concatenated with the input query for the model, to obtain predictions accordingly.

4.3 Method

Problem setup. Computer vision includes a series of tasks like object detection [6], [68], [197] and panoptic segmentation [7], [90], [198], which are typically handled by specialized models designed for specific input-target mappings [199]. While effective for single tasks,

this specialization restricts model adaptability and scalability across multiple tasks or diverse visual data. To overcome this limitation, we aim to design a *conditional generative framework* that unifies multiple vision tasks within a single cohesive model. Specifically, given a query $\boldsymbol{x}$ (e.g., an image or a video), the framework produces the corresponding prediction $\hat{\boldsymbol{y}}$ to approximate the target $\boldsymbol{y}$ conditioned on a set of input-target pairs $\boldsymbol{s}$. These conditioning pairs provide task definitions and guidance, enabling the model to flexibly adapt to different tasks according to the supplied examples. Formally, the objective is to model the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{s})$.

Framework overview. As shown in Figure 4.2(a), the proposed Large Vision Diffusion Transformer (**LaVin-DiT**) framework integrates a spatial-temporal variational autoencoder (ST-VAE) with a joint diffusion transformer to unify multiple vision tasks. Given a vision task, e.g., panoptic segmentation, we first sample a set of input-target pairs as the task definition. Afterward, the set and other visual examples are fed into ST-VAE, which are encoded into latent representations. Subsequently, the encoded representations are patchified and unfolded into a sequential format. The set and input visual data form the conditional latent presentation $\boldsymbol{z}_c$, while the target is perturbed with random Gaussian noise, yielding a noisy latent representation $\boldsymbol{z}_t$. Both $\boldsymbol{z}_c$ and $\boldsymbol{z}_t$ are then put into the joint diffusion transformer (J-DiT), which denoises $\boldsymbol{z}_t$ to recover a clean latent representation within the shared latent space. Lastly, the recovered latent representation is passed through the ST-VAE decoder to reconstruct the target in raw pixel space. Below we provide a detailed technical exposition of ST-VAE and J-DiT.

4.3.1 LaVin-DiT Modules

4.3.1.1 ST-VAE

It is computationally demanding to process visual data in raw pixel space [158]. To address this, we propose to use a spatial-temporal variational autoencoder (ST-VAE) [194], [200], [201]. ST-VAE can efficiently compress spatial and temporal information, and encode them from pixel space into compact latent space. As illustrated in Figure 4.2(b), ST-VAE uses

causal 3D convolutions and deconvolutions to compress and reconstruct visual data. It overall includes an encoder, a decoder, and a latent regularization layer. These components are structured into four symmetric stages with alternating $2\times$ downsampling and upsampling. The first two stages operate on both spatial and temporal dimensions, while the last stage affects only the spatial dimension, achieving an effective $4 \times 8 \times 8$ compression and reducing computational load. Besides, we apply a Kullback-Leibler (KL) constraint to regularize the Gaussian latent space.

To prevent future information leakage and its adverse effect on temporal predictions, we pad all locations at the start of the temporal convolution space. Additionally, to support both image and video processing, we treat the first frame of an input video independently, compressing it only spatially to maintain temporal independence. Subsequent frames are compressed along both spatial and temporal dimensions. The encoder of ST-VAE compresses the input to a lower-dimensional latent space, and the reconstruction is achieved through a decoding process. Training the ST-VAE occurs in two stages: we first train on images alone, then jointly on both images and videos. During each stage, we optimize the model using a combination of the mean squared error, perceptual loss [158], [202], and adversarial loss [158].

4.3.1.2 J-DiT

Diffusion transformers (DiT) [178] have emerged as a powerful method for generative modeling. Our joint diffusion transformer (J-DiT) builds upon DiT but introduces modifications to support the task-conditioned generation. A key distinction from the original DiT is our consideration of two conceptually different latent representations. The condition latent representation is clean, while the target latent representation is perturbed by Gaussian noise, resulting in potentially distinct value ranges for the two. To handle the difference and improve alignment between task-specific and visual information, we construct separate patch embeddings for the condition and target latents. Each embedding layer uses a patch size of 2×2, which allows for tailoring the representations for each latent type. As shown in Figure 4.2, the sampled timestep t, along with the condition and target sequences, is fed into a series of diffusion transformer layers. Building on the MM-DiT [188] architecture, we introduce condition-

and target-specific adaptive RMS normalization (AdaRN) to modulate each representation space independently. This is achieved through distinct timestep embeddings for the condition and target within AdaRN layers.

Full-sequence joint attention. Full-sequence joint attention is key in our transformer layers, which processes condition and noisy target sequences together to enhance task-specific alignment. As shown in Figure 4.2(c), the condition and target sequences are linearly projected, concatenated, and processed by a bidirectional attention module, allowing each to operate in its own space while considering the other. To improve speed and memory efficiency, we replace multi-head attention with grouped-query attention [203], which groups query heads to share a single set of key-value heads. This approach reduces parameters while retaining expressiveness, closely matching standard multi-head attention performance. Besides, to stabilize training with larger models and longer sequences, we add QK-Norm before query-key dot products to control attention entropy growth. Following [204], we also apply sandwich normalization after each attention and FFN layer to maintain activation magnitudes amid residual connections.

3D rotary position encoding. Unlike [32], we argue that it is sub-optimal to model visual data as a one-dimensional sequence, because 1D positional embedding is limited in capturing precise spatial-temporal positions. Instead, by treating multiple image-annotation pairs or video clips as a single continuous sequence, we can use 3D Rotary Position Encoding (3D RoPE) [205] to represent spatial-temporal relationships concisely. Then, each location in a video can be expressed by a 3D coordinate. Specifically, each token in an input sequence is associated with a 3D coordinate (t, x, y), representing its position in temporal and spatial dimensions. The 3D RoPE encodes positional information by decomposing it into three separate 1D RoPEs along the temporal and spatial axes, allowing the model to capture relative positional relationships across all dimensions inherently.

Technically, for each axis $a \in \{t, x, y\}$, we define a rotation matrix $R_p^{(a)}$ that operates on a dedicated subspace of an embedding vector z. The embedding vector is partitioned accordingly: $z = [z^{(t)},\ z^{(x)},\ z^{(y)}]$, where each subvector $z^{(a)} \in \mathbb{R}^{d_a}$ corresponds to axis a

and $d = d_t + d_x + d_y$. The rotation matrix $R_p^{(a)}$ is constructed in a block-wise manner, rotating each pair of dimensions $(2i, 2i+1)$ by an angle $\theta_{p,i}^{(a)} = p^{(a)} \cdot \omega_i^{(a)}$, where $\omega_i^{(a)} = \omega_{\text{base}}^{-2i/d_a}$ and ω_{base} is a predefined constant:

$$R_p^{(a)} = \begin{bmatrix} R_p^{(a,0)} & & \\ & \ddots & \\ & & R_p^{(a,d_a/2-1)} \end{bmatrix}, \quad \text{where} \tag{4.1}$$

$$R_p^{(a,i)} = \begin{bmatrix} \cos\left(\theta_{p,i}^{(a)}\right) & -\sin\left(\theta_{p,i}^{(a)}\right) \\ \sin\left(\theta_{p,i}^{(a)}\right) & \cos\left(\theta_{p,i}^{(a)}\right) \end{bmatrix}. \tag{4.2}$$

When computing self-attention, the rotated query q and key k are obtained by applying the rotation matrices: $q'^{(a)} = R_p^{(a)} q^{(a)}$ and $k'^{(a)} = R_p^{(a)} k^{(a)}$. The full rotated query and key are then $q' = [q'^{(t)},\ q'^{(x)},\ q'^{(y)}]$ and $k' = [k'^{(t)},\ k'^{(x)},\ k'^{(y)}]$. When computing the attention between tokens at positions j and k, the dot product incorporates the rotations from all axes:

$$(q_j'^{\top}) k_k' = \sum_{a \in \{t,x,y\}} \left(q^{(a)}\right)^{\top} R_j^{(a)\top} R_k^{(a)} k^{(a)}. \tag{4.3}$$

The key property of rotation matrices is that the product of two rotation matrices corresponds to a rotation by the difference of their angles:

$$R_j^{(a)\top} R_k^{(a)} = R_{j-k}^{(a)}, \tag{4.4}$$

where $R_{p-q}^{(a)}$ is the rotation matrix for the relative position $j^{(a)} - k^{(a)}$, constructed as:

$$R_{j-k}^{(a)} = \begin{bmatrix} R_{j-k}^{(a,0)} & & \\ & \ddots & \\ & & R_{j-k}^{(a,N_a-1)} \end{bmatrix}, \quad \text{where} \tag{4.5}$$

$$R_{j-k}^{(a,i)} = \begin{bmatrix} \cos\left(\Delta_{jk}^{(a)} \omega_i^{(a)}\right) & -\sin\left(\Delta_{jk}^{(a)} \omega_i^{(a)}\right) \\ \sin\left(\Delta_{jk}^{(a)} \omega_i^{(a)}\right) & \cos\left(\Delta_{jk}^{(a)} \omega_i^{(a)}\right) \end{bmatrix}, \tag{4.6}$$

$$\Delta_{jk}^{(a)} = j^{(a)} - k^{(a)}. \tag{4.7}$$

This block-wise matrix format explicitly shows that the attention score depends on the relative positions $j^{(a)} - k^{(a)}$ along each axis a. With the introduction of 3D RoPE, we provide a

unified and accurate spatial-temporal representation of positional encoding for various vision tasks.

Training procedure of J-DiT. We train J-DiT using flow matching [206] in the latent space. Specifically, given a representation $\boldsymbol{z}_0$ and noise $\boldsymbol{z}_1 \sim \mathcal{N}(0,1)$, flow matching defines a linear interpolation based forward process: $\boldsymbol{z}_t = t\boldsymbol{z}_0 + (1-t)\boldsymbol{z}_1$, where the timestep $t \in [0,1]$. This forward process induces a time-dependent velocity field $v(\boldsymbol{z}_t, t)$ that drives the flow along the linear path in the direction of $(\boldsymbol{z}_0 - \boldsymbol{z}_1)$. The velocity field defines an ordinary differential equation (ODE): $\mathrm{d}\boldsymbol{z}_t = v(\boldsymbol{z}_t, t)\mathrm{d}t$. We employ J-DiT that is parameterized by $\boldsymbol{\theta}$, to predict the velocity field that transforms noise into a clean latent representation. The training objective of flow matching is to directly regress the target velocity field, leading to the Conditional Flow Matching (CFM) loss [206]:

$$\ell_{\text{CFM}} = \int_0^1 \mathbb{E}[|v_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) - (\boldsymbol{z}_0 - \boldsymbol{z}_1)|_2^2]\mathrm{d}t. \tag{4.8}$$

We also provide the algorithm flow of the training procedure in Algorithm 2.

Algorithm 2 LaVin-DiT Training Procedure

Require: ST-VAE encoder $\mathrm{Enc}(\cdot)$, dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^K$, initialized parameters $\boldsymbol{\theta}$ of vector field $v_{\boldsymbol{\theta}}(\boldsymbol{z}, t)$, total iterations T, learning rate η.

for $n = 1$ to T **do**
  Sample $\boldsymbol{x} \sim \mathcal{D},\ \boldsymbol{c} \sim \mathcal{D}$
  Compute latents: $\boldsymbol{z}_0 \leftarrow \mathrm{Enc}(\boldsymbol{x}),\ \boldsymbol{z}_c \leftarrow \mathrm{Enc}(\boldsymbol{c})$
  Initialize random latent: $\boldsymbol{z}_1 \sim \mathcal{N}(0,1)$
  Sample time step: $t \sim \mathrm{LogitNormal}(0,1)$
  Interpolate: $\boldsymbol{z}_t \leftarrow (1-t)\boldsymbol{z}_1 + t\boldsymbol{z}_0$
  Target vector: $\boldsymbol{u} \leftarrow \boldsymbol{z}_0 - \boldsymbol{z}_1$
  Predicted vector: $\boldsymbol{v} \leftarrow v_{\boldsymbol{\theta}}(\boldsymbol{z}_t,\ \boldsymbol{z}_c,\ t)$
  Compute loss: $\mathcal{L} \leftarrow \mathbb{E}[|\boldsymbol{v} - \boldsymbol{u}|_2^2]$
  Update parameters: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{L}$
end for

Generation procedure of J-DiT. Upon completion of J-DiT training, we use it to generate new representations by integrating from the noise distribution toward representation distribution. Specifically, starting from noise $\boldsymbol{z}_1' \sim \mathcal{N}(0,1)$ at $t = 1$, we integrate the learned J-DiT backward to $t = 0$ to obtain a representation $\boldsymbol{z}_0'$. For instance, using the Euler method, we discretize the time interval [0,1] to N steps with a negative step size $\Delta t = -1/N$ to indicate

backward integration in time. At each step $k = 0, 1/N, \ldots, (N-1)/N$, we update the time and generated representation as follows:

$$t^{(k+1/N)} = t^{(k)} + \Delta t, \tag{4.9}$$

$$\boldsymbol{z}^{(k+1/N)} = \boldsymbol{z}^{(k)} + v_{\boldsymbol{\theta}}(\boldsymbol{z}^{(k)}, t^{(k)})\Delta t, \tag{4.10}$$

where $t^{(0)} = 1$, $t^{(1)} = 0$, $\boldsymbol{z}^{(0)} = \boldsymbol{z}'_1$, and $\boldsymbol{z}^{(1)} = \boldsymbol{z}'_0$. By iteratively applying these updates, we obtain a new presentation for the following decoding process of ST-VAE. The whole algorithm of generation procedure is illustrated in Algorithm 3.

Algorithm 3 LaVin-DiT Inference Procedure

Require: Trained vector field $v_{\boldsymbol{\theta}}(z, t)$, ST-VAE encoder $\mathrm{Enc}(\cdot)$, ST-VAE decoder $\mathrm{Dec}(\cdot)$, timesteps N, dataset $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{K}$.

Set step size $\Delta t \leftarrow \frac{1}{N}$, initialize $t^{(N)} \leftarrow 1$
Sample initial latent: $\boldsymbol{z}_1 \sim \mathcal{N}(0, 1)$
Encode condition: $\boldsymbol{z}_c \leftarrow \mathrm{Enc}(\boldsymbol{c}),\ \boldsymbol{c} \sim \mathcal{D}$
for $k = N$ down to 1 **do**
    Update time: $t^{(k-1)} \leftarrow t^{(k)} - \Delta t$
    Compute vector field: $\boldsymbol{v}^{(k)} \leftarrow v_{\boldsymbol{\theta}}(\boldsymbol{z}^{(k)},\ \boldsymbol{z}_c,\ t^{(k)})$
    Update latent: $\boldsymbol{z}^{(k-1)} \leftarrow \boldsymbol{z}^{(k)} - \Delta t \cdot \boldsymbol{v}^{(k)}$
end for
Decode sample: $\hat{\boldsymbol{y}} \leftarrow \mathrm{Dec}(\boldsymbol{z}_0)$

4.3.2 LaVin-DiT Inference

After completing the training of LaVin-DiT, the model becomes versatile and is ready to be applied across a range of downstream tasks. Specifically, when given a query (e.g., an image or a video) for any chosen task, we randomly sample a set of input-target pairs that define the task. These pairs, alongside the visual input and a Gaussian noise component, are then fed into the Joint Diffusion Transformer (J-DiT). Within J-DiT, these elements are processed to generate a latent representation. Finally, this latent representation is passed through the ST-VAE decoder, which transforms it into the raw pixel space to produce the desired prediction. To better understand this inference procedure, please refer to Figure 4.2(a).

4.4 Experiments

4.4.1 Setup

Training data. To unify multiple computer vision tasks, we construct a large-scale multi-task dataset that encompasses indoor and outdoor environments, spanning real-world and synthetic domains. This dataset comprises approximately 3.2 million unique images [72], [133], [135], [207], [208] and 0.6 million unique videos [209], [210], [211], covering over 20 tasks:

- *Image-based tasks*: object detection, instance segmentation, panoptic segmentation, pose estimation, edge extraction, depth estimation, surface normal estimation, inpainting, colorization, image restoration tasks (e.g., de-raining, de-glass blur, and de-motion blur), depth-to-image, and normal-to-image generation.
- *Video-based tasks*: frame prediction, video depth estimation, video surface normal estimation, video optical flow estimation, video instance segmentation, depth-to-video, and normal-to-video generation.

To overcome the limitations of large-scale annotations for depth and surface normal estimation, we generate pseudo depth and normal maps on ImageNet-1K [207] by utilizing Depth-anything V2 [212] and Stable-Normal (turbo) [213], respectively.

Implementation details. We conduct training in two stages, progressively increasing the image resolution. In the first stage, we train at a 256×256 resolution for 100,000 steps, leveraging DeepSpeed ZeRO-2 [214] optimization and gradient checkpointing to manage memory and computational efficiency. We employ a global batch size of 640 and use an AdamW optimizer [130] with a learning rate of 0.0001, betas set to 0.9 and 0.95, and weight decay of 0.01. This setup provides stable training across configurations without the need for a warmup or additional regularization techniques. In the second stage, we upscale the resolution to 512×512 and continue training for an additional 20,000 steps, while the learning rate is adjusted to 0.00005. Other hyperparameters are retained from the first stage. This two-stage strategy enables efficient scaling, ensuring optimal performance across resolutions.

TABLE 4.1. Configurations of LaVin-DiT with different numbers of parameters.

	LaVin-DiT		
	0.1B	**1.0B**	**3.4B**
Latent channels	16	16	16
Patch size	2×2	2×2	2×2
Hidden channels	512	1024	2304
Num. layers	12	28	22
Num. heads	8	16	32
K.V. groups	-	-	4
Drop path	0.0	0.1	0.1
Uncond. ratio	0.1	0.1	0.1
Grad. clip	1.0	1.0	1.0
EMA moment.	0.9999	0.9999	0.9999
Extra norm.	-	S-Norm.	S-Norm.
Position embed.	3D-RoPE	3D-RoPE	3D-RoPE

By default, we utilize 20 timesteps ($N = 20$) during inference. All experiments are conducted on 64×NVIDIA A100-80G GPUs.

Here, we detail the architecture of the LaVin-DiT models. Table 4.1 outlines the configurations for three parameter scales: 0.1B, 1.0B, and 3.4B. Each configuration is characterized by key architectural hyperparameters, including the number of latent channels, patch size, hidden channels, and the number of layers. Additionally, the configurations specify the number of attention heads, key-value groups, drop path rates, and unconditional ratios. To further enhance model training, we incorporate advanced techniques such as gradient clipping and the Exponential Moving Average (EMA). All models utilize 3D-RoPE to ensure consistent spatial and temporal encoding across scales. For large models, we employ sandwich normalization to improve training stability.

4.4.2 Large-Scale Multi-Task Dataset Composition

We build a large-scale multi-task dataset to unify diverse computer vision tasks. We integrate multiple public image-level and video-level task benchmarks into a large-scale dataset for training. Details are listed in Table 4.2 and 4.3.

TABLE 4.2. The first part of summary of the large-scale multi-task dataset used in LaVin-DiT, including the number of examples and annotation types for each component dataset. Tasks range from visual understanding and generation.

Task	**Dataset**	**Number of Samples**	**Annotation Type**
Single Object Detection	COCO 2017 train [72]	117,266	Ground Truth
	Object365 train [208]	1,728,778	Ground Truth
Instance Segmentation	COCO 2017 train [72]	117,266	Ground Truth
	ADE20K train+val [133]	19,020	Ground Truth
	Cityscapes train+val [135]	3,457	Ground Truth
Panoptic Segmentation	COCO 2017 train [72]	117,266	Ground Truth
	ADE20K train+val [133]	19,020	Ground Truth
	Cityscapes train+val [135]	3,457	Ground Truth
Pose Estimation	COCO 2017 train [72]	64,115	Ground Truth
Pose-to-Image Generation	COCO 2017 train [72]	64,115	Ground Truth
Depth Estimation	ImageNet1K train [207]	1,281,167	Depth-anything V2
Depth-to-Image Generation	ImageNet1K train [207]	1,281,167	Depth-anything V2
Surface Normal Estimation	COCO 2017 train [72]	117,266	Stable-Normal (turbo)
	ADE20K train+val [133]	19,020	Stable-Normal (turbo)
	Cityscapes train+val [135]	3,457	Stable-Normal (turbo)
Normal-to-Image Generation	COCO 2017 train [72]	117,266	Stable-Normal (turbo)
	ADE20K train+val [133]	19,020	Stable-Normal (turbo)
	Cityscapes train+val [135]	3,457	Stable-Normal (turbo)
Edge Detection	ImageNet1K [207] train	1,281,167	Canny (OpenCV)
	COCO 2017 train [72]	117,266	Canny (OpenCV)

Evaluation protocols. In this work, we assess our model on a comprehensive range of computer vision tasks spanning both image and video domains. We provide quantitative results for 10 tasks (The others are presented with visualization results). Following established protocols, Here we introduce the evaluation metrics for these 10 tasks.

TABLE 4.3. The second part of summary of the large-scale multi-task dataset used in LaVin-DiT, including the number of examples and annotation types for each component dataset. Tasks range from visual understanding and generation.

Task	**Dataset**	**Number of Samples**	**Annotation Type**
Inpainting	ImageNet1K train [207]	1,281,167	Crop (OpenCV)
	COCO 2017 train [72]	117,266	Crop (OpenCV)
Colorization	ImageNet1K train [207]	1,281,167	Grayscale (OpenCV)
	COCO 2017 train [72]	117,266	Grayscale (OpenCV)
De-glass Blur	ImageNet1K train [207]	1,281,167	Albumentations
	COCO 2017 train [72]	117,266	Albumentations
De-motion Blur	ImageNet1K train [207]	1,281,167	Albumentations
	COCO 2017 train [72]	117,266	Albumentations
De-raining	ImageNet1K train [207]	1,281,167	Albumentations
	COCO 2017 train [72]	117,266	Albumentations
Frame Prediction	UCF101 train [209]	7,629	N/A
	Kinetic 700 train+val [210]	570,465	N/A
	Kubric train [211]	48,689	N/A
Video Depth Estimation	Kubric train [211]	48,689	Ground Truth
Depth-to-Video Generation	Kubric train [211]	48,689	Ground Truth
Video Surface Normal Estimation	Kubric train [211]	48,689	Ground Truth
Normal-to-Video Generation	Kubric train [211]	48,689	Ground Truth
Video Optical Flow Estimation	Kubric train [211]	48,689	Ground Truth
Video Instance Segmentation	Kubric train [211]	48,689	Ground Truth

Colorization. We randomly sample 1,000 images from ImageNet-1K validation set [207] and convert them into grayscale. We adopt LPIPS [202] and mean squared error (MSE) as metrics.

Inpainting. We randomly sample 1,000 images from ImageNet-1K validation set [207] and mask out a 128×128 region for each image. We adopt the LPIPS [202] and Frechet Inception Distance (FID) as metrics.

Depth Estimation. We evaluate our model on NYUv2 test set [156], including 654 images. Following the protocol of affine-invariant depth evaluation [215], we first align the prediction to the ground truth with the least squares fitting. Afterwards, we adopt Absolute Mean Relative Error (AbsRel) and Mean Squared Error (MSE) as metrics.

Surface Normal Estimation. We evaluate our model on NYUv2 test set [156]. Following the protocol used in [216], we calculate the angular error between the prediction and the ground-truth normal maps and use the mean angular error as the metric.

Depth-to-Image Generation. We adopt all samples in the NYUv2 dataset [156], including 1,449 images. Given the pseudo label generated via Depth-anything V2 or Stable-Normal (turbo), we generate the corresponding RGB image and use the LPIPS [202] and Frechet Inception Distance (FID) as metrics.

Normal-to-Image Generation. The metrics are the same those in Depth-to-Image Generation.

Single Object Detection. We evaluate the model on the Pascal-5i dataset [217] and adopt the mean intersection-over-union (mIoU) as the metric.

Foreground Segmentation. We evaluate our model on the Pascal-5i dataset [217], including 4 different test splits. Following the protocol in [152], we extract binary masks from our predictions and report the mIoU.

Deraining. We randomly sample 1,000 images from ImageNet-1K validation set [207] and apply the raining filter on them. We adopt the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) as metrics.

De-motion Blur. We randomly sample 1,000 images from the ImageNet-1K validation set [207] and apply motion blur on these images. We adopt the PSNR and SSIM as metrics.

TABLE 4.4. **Comparison on foreground segmentation, single object detection, and colorization.** For foreground segmentation and single object detection, we report "mIoU" (higher is better). For colorization, we report "LPIPS" [202] and "MSE" (lower is better). Note that foreground segmentation and single object detection are *unseen* tasks during our training.

Method	**Foreground Segmentation** (mIoU ↑)				**Single Object Detection** (mIoU ↑)				**Colorization** ↓	
	Split 1	Split 2	Split 3	Split 4	Split 1	Split 2	Split 3	Split 4	MSE	LPIPS
MAE [152]	17.42	25.70	18.64	16.53	5.49	4.98	5.24	5.84	0.43	0.55
MAE-VQGAN [152]	27.83	30.44	26.15	24.25	24.19	25.20	25.36	25.23	0.67	0.40
LVM [32]	48.94	51.29	47.66	50.82	48.25	49.60	50.08	48.92	0.51	0.46
LaVin-DiT	**67.87**	**75.80**	**66.98**	**66.90**	**67.85**	**69.32**	**68.76**	**68.88**	**0.24**	**0.26**

TABLE 4.5. **Comparison on NYU-v2 depth estimation, surface normal estimation and ImageNet inpainting** [156], [207]. For depth estimation, we report absolute relative difference (AbsRel) and threshold accuracy (δ_1). For surface normal estimation, we report mean angular error (MAE) and angle accuracy within a threshold ($< 11.25^\circ$). We report FID for inpainting. † denotes evaluations on the official 7B model released by [32].

Method	**Depth Estimation**		**Normal Estimation**		**Inpainting**
	AbsRel (↓)	δ_1 (↑)	MAE (↓)	$< 11.25^\circ$ (↑)	FID (↓)
DPT [218]	9.8	90.3	-	-	-
StableNormal [213]	-	-	19.707	53.042	-
Marigold [219]	6.0	95.9	20.864	50.457	-
LVM† [32]	30.2	52.3	23.433	44.836	4.05
LaVin-DiT	**6.2**	**96.1**	**15.901**	**58.382**	**1.65**

4.4.3 Main Results

Quantitative analysis. To assess the effectiveness of our proposed method, we conduct extensive experiments across a broad range of computer vision tasks and report results of the 3.4B model by default, as summarized in Tables 4.4 and 4.5. Our method consistently outperforms existing baselines across multiple tasks, including challenging cases such as unseen foreground segmentation and single-object detection, demonstrating superior generalization and adaptability across diverse scenarios. Note that unless otherwise specified, we report LaVin-Dit (3.4B) performance.

As shown in Table 4.4, we report the performance on foreground segmentation and single object detection across different splits. Our LaVin-DiT achieves significant improvements over baseline methods in all splits. Specifically, in the foreground segmentation task, we attain mIoUs of 67.87%, 75.80%, 66.98%, and 66.90% across four splits, consistently outperforming previous methods such as LVM [32] and MAE-VQGAN [152] by a substantial margin. Additionally, for single object detection, our model demonstrates strong performance, achieving top results in all splits. Notably, we achieve a mIoU of 68.88% in Split 4, which is a considerable margin of 19.96% over the best-performing baseline LVM. These significant gains highlight our model's ability to effectively segment and detect objects across a range of scenarios, even when faced with tasks unseen during training. Following prior work [32], [152], we further evaluate our model in the colorization task, where lower LPIPS and MSE values indicate superior performance. As shown in Table 4.4, our method achieves an LPIPS of 0.26 and an MSE of 0.24, significantly outperforming all baselines. These results underscore our model's capability to generate realistic and natural colors from grayscale images, which is essential in restoration and artistic fields.

To validate the ability of our model to understand the geometric structure of 3D scenes, we evaluate it on NYU-v2 depth estimation and surface normal estimation tasks [156], as shown in Table 4.5. As [32] do not report related results in their paper, we conduct evaluations using their official 7B model[2]. For depth estimation, our model achieves an AbsRel of 6.2 and a

[2] https://huggingface.co/Emma02/LVM_ckpts

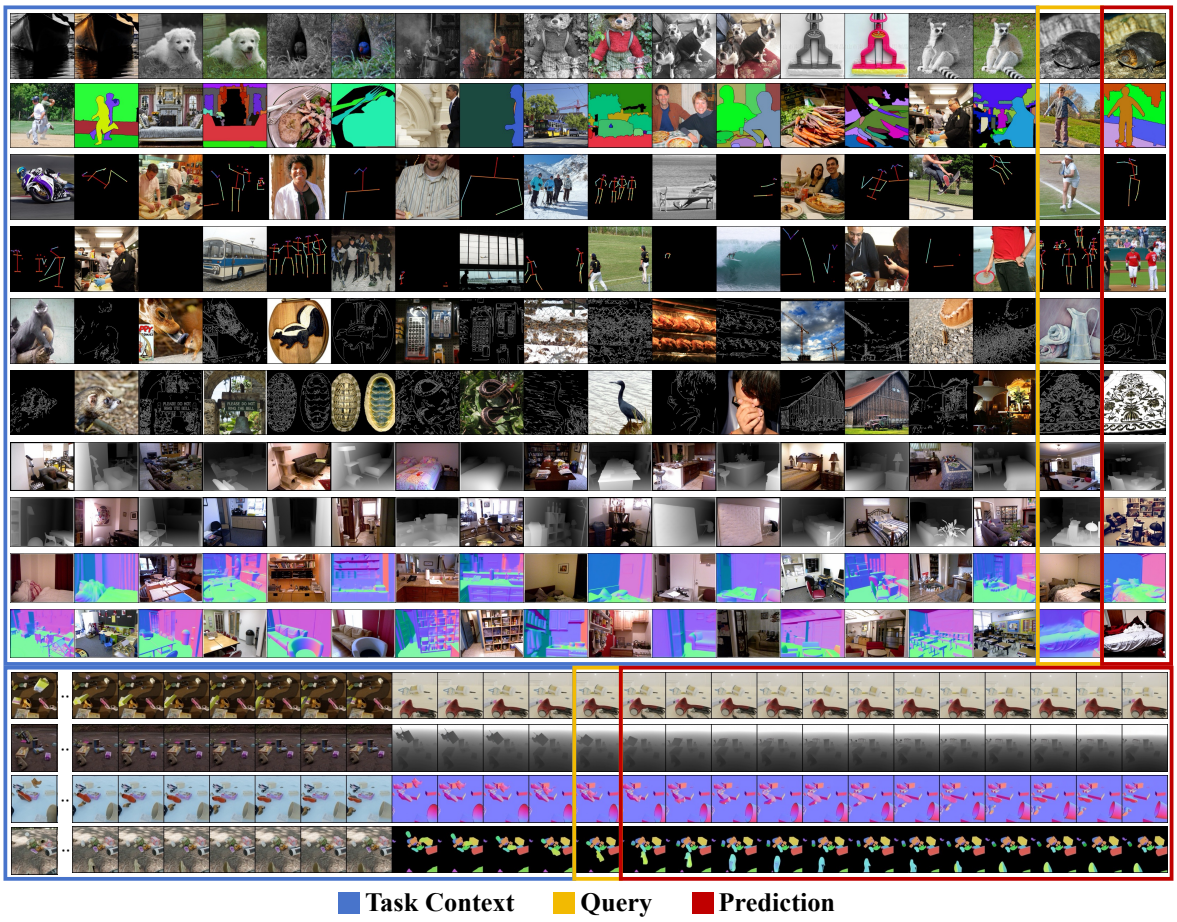


FIGURE 4.3. **Qualitative results on diverse image and video-based tasks.** The first ten rows show image-based tasks, where each row contains a sequence of images interleaved with annotations, followed by a query. The last image is predicted by the model (marked in red). The last four rows show video-based tasks, where each row includes a video sequence with a series of target frames as task context, followed by a query frame. A set of frames in the red box indicates the model's predictions. *Best viewed in color.*

threshold accuracy δ_1 of 96.1%, demonstrating competitive performance compared to expert models such as Marigold [219] and DPT [218]. In the surface normal estimation task, our method achieves an MAE of 15.901 and accuracy within a $< 11.25^\circ$ threshold of 58.382, surpassing the powerful expert model StableNormal [213]. This performance underscores our model's proficiency in estimating surface orientations accurately, enhancing its applicability in tasks requiring precise geometrical understanding, such as augmented reality and 3D reconstruction. These results reflect our model's capability to comprehend the geometric structure of 3D scenes with precision, even in complex environments, which is crucial for

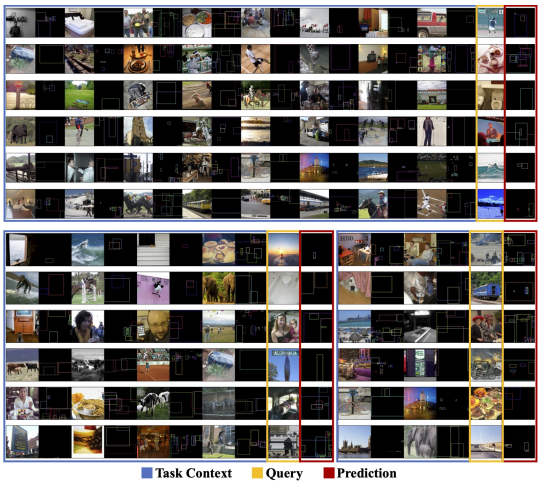


FIGURE 4.4. **Visualization on object detection.**

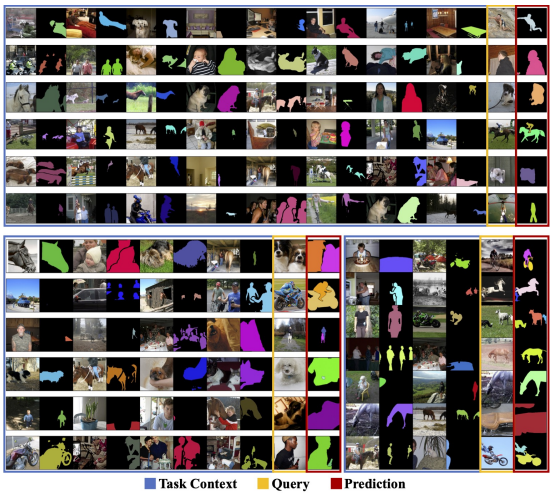


FIGURE 4.5. **Visualization on foreground segmentation.**

FIGURE 4.6. **Visualization on panoptic segmentation.**

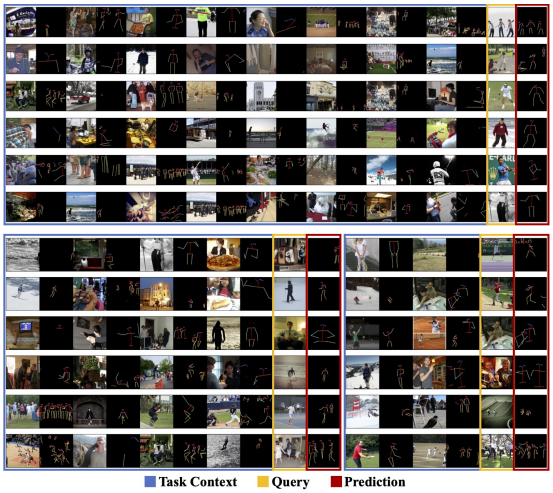


FIGURE 4.7. **Visualization on pose estimation.**

real-world applications like 3D scene reconstruction and spatial perception. Furthermore, we compare our LaVin-DiT to LVM on the inpainting task. Using 2,500 randomly selected images from the ImageNet-1K validation set, our model achieves an FID of 1.65, which greatly improves over the FID of 4.05 obtained by LVM.

Qualitative analysis. As shown in Figures 4.3, we present qualitative results in a wide variety of image-based and video-based tasks. Our model consistently follows task contexts and precisely generates the corresponding predictions. Furthermore, given sequential frames with

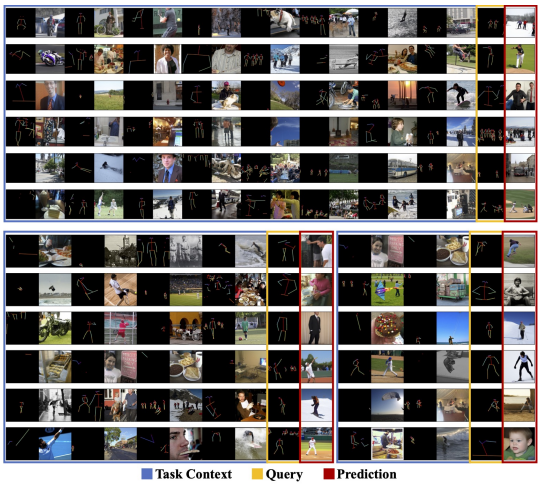


FIGURE 4.8. **Visualization on pose-to-image generation.**

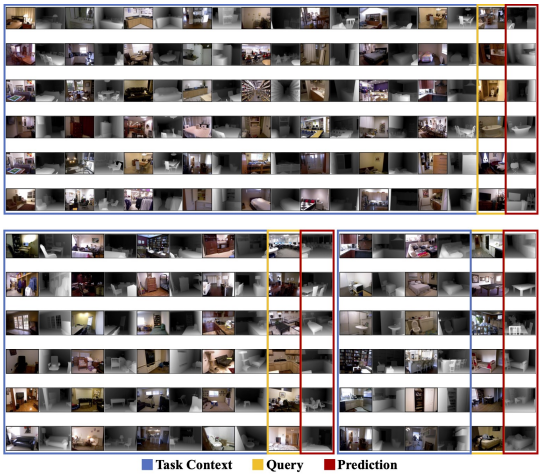


FIGURE 4.9. **Visualization on depth estimation.**

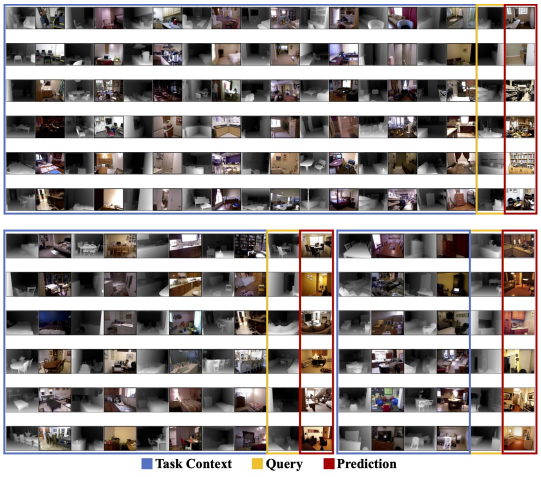


FIGURE 4.10. **Visualization on depth-to-image generation.**

FIGURE 4.11. **Visualization on surface normal estimation.**

task contexts, our model generates predictions for the subsequent 12 frames, which exhibits its ability to handle temporal consistency and scene dynamics effectively. Furthermore, we show more visualization results for each task, including object detection (Figure 4.4), foreground segmentation (Figure 4.5), panoptic segmentation (Figure 4.6), pose estimation (Figure 4.7), pose-to-image generation (Figure 4.8), depth estimation (Figure 4.9), depth-to-image generation (Figure 4.10), surface normal estimation (Figure 4.11)), normal-to-image

FIGURE 4.12. **Visualization on surface normal-to-image generation.**

FIGURE 4.13. **Visualization on edge detection.**

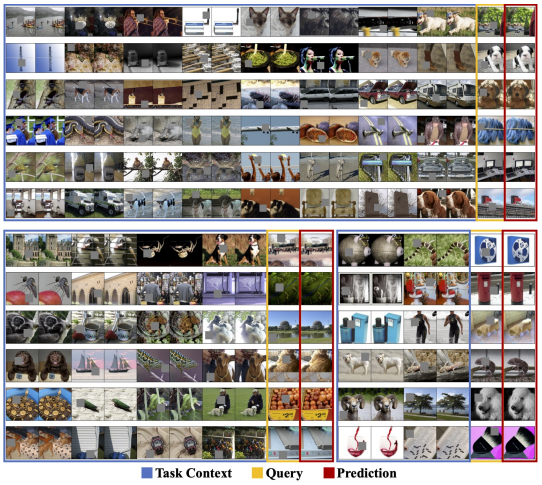


FIGURE 4.14. **Visualization on image inpainting.**

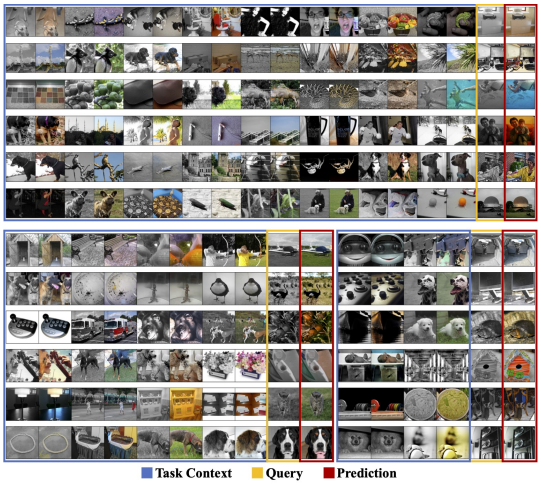


FIGURE 4.15. **Visualization on image colorization.**

generation (Figure 4.12), edge detection (Figure 4.13), inpainting (Figure 4.14), colorization (Figure 4.15).

4.4.4 Scalability

To investigate the scalability of the proposed LaVin-DiT, we conduct experiments with three model sizes, i.e., 0.1B, 1.0B, and 3.4B parameters. We train the three models for 100,000 steps.

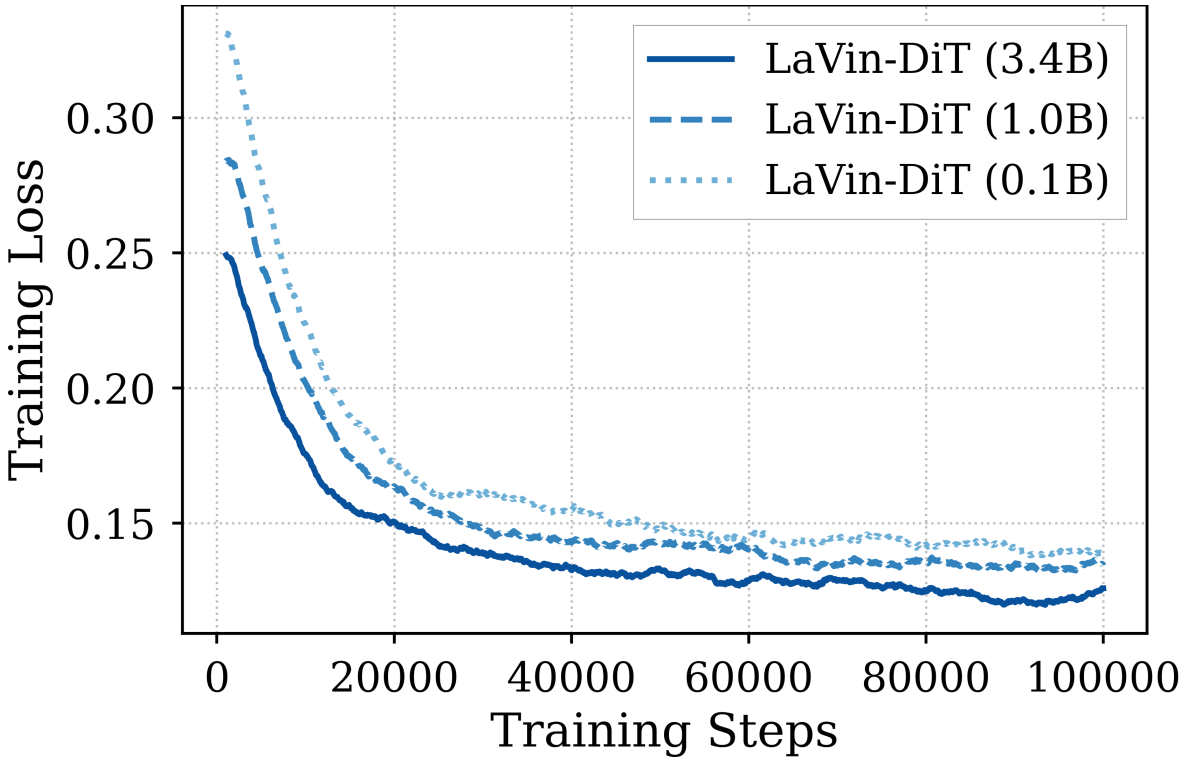


FIGURE 4.16. **Training loss curves for LaVin-DiT of varying model sizes.** The 3.4B model demonstrates faster convergence, achieving lower training losses than smaller models as training progresses.

Figure 4.16 illustrates the training loss curves, which shows that larger models consistently achieve lower loss values. Additionally, the 3.4B model converges more rapidly, reaching smaller loss values in fewer training steps. This accelerated convergence suggests that larger models are better equipped to capture complex data patterns, leading to improved learning efficiency. The observed training dynamics underscore the advantages of scaling up model capacity for complex vision tasks, where larger models can more effectively capture diverse data characteristics.

Beyond training dynamics, the model size also has a substantial impact on downstream task performance. This is evident in colorization and depth estimation tasks, which were selected for their distinct requirements in capturing color fidelity and spatial structure. As seen in Figure 4.17, model performance improves consistently as its scale increases. Specifically, for colorization, the 3.4B model achieves an MSE of 0.273, significantly outperforming the 1.0B and 0.1B models that achieve MSEs of 0.311 and 0.609, respectively. Similarly,

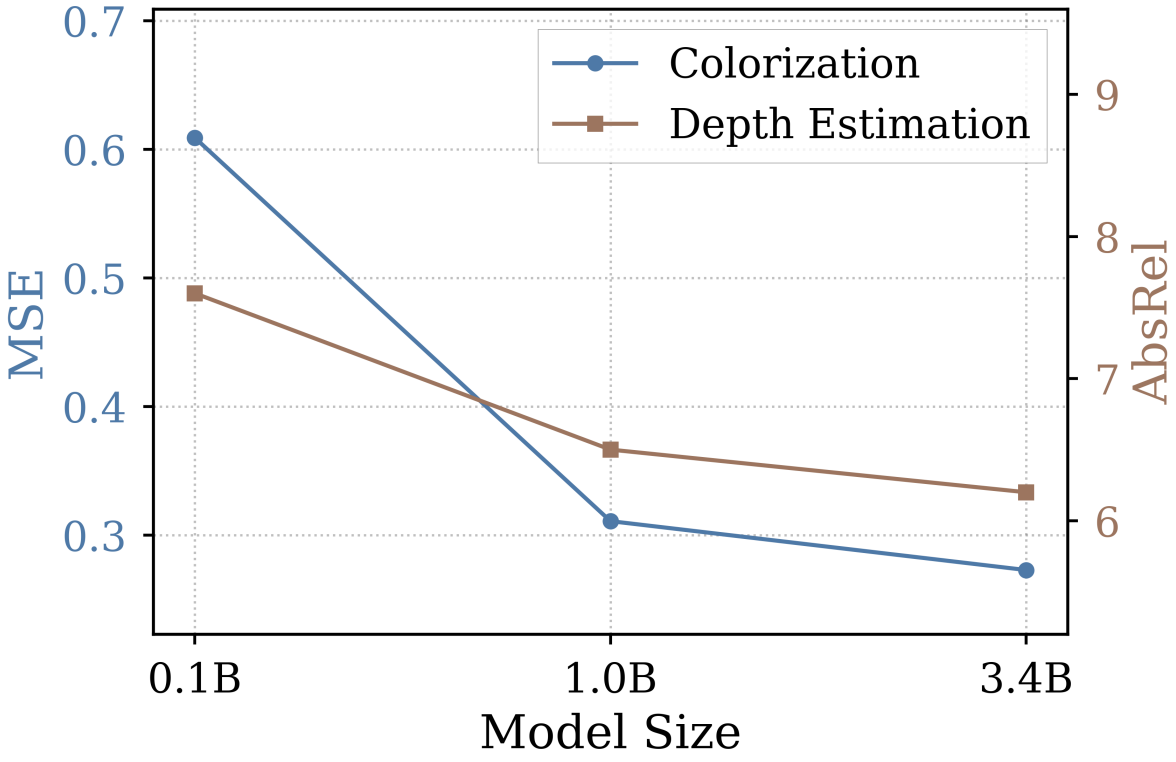


FIGURE 4.17. **Performance for LaVin-DiT of varying sizes.** Comparison of LaVin-DiT with different parameters on colorization (MSE) and depth estimation (AbsRel). Lower values indicate better performance.

in depth estimation, the 3.4B model attains an AbsRel of 6.2, compared to 6.5 and 7.6 for the 1.0B and 0.1B models. These results demonstrate that larger models indeed deliver enhanced performance across multiple tasks, affirming LaVin-DiT as a scalable and adaptable framework for high-performance vision applications.

4.4.5 Inference Latency Analysis

As demonstrated in Figure 4.18, we compare the inference latency of LaVin-DiT and LVM (both 7B models) across increasing resolutions, demonstrating that our method is consistently more efficient. At a resolution of 256, LaVin-DiT requires only 4.67 seconds per example, while LVM takes 8.1 seconds, with this efficiency gap widening at higher resolutions (e.g., 20.1 seconds *v.s.* 47.2 seconds at 512). This difference underscores a key advantage of diffusion models for vision tasks: unlike autoregressive models that process tokens sequentially and become increasingly time-intensive with larger inputs, diffusion models process tokens in

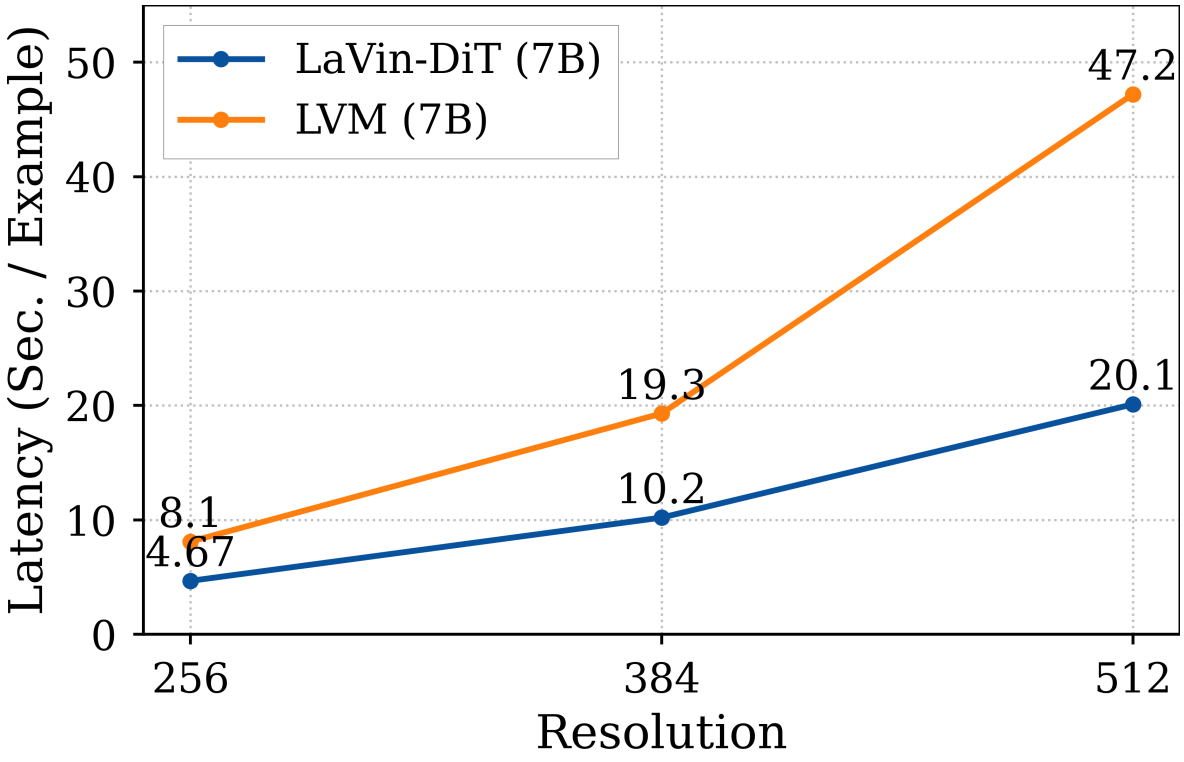


FIGURE 4.18. **Inference latency comparison.** LaVin-DiT consistently achieves lower latency than LVM [32] across different resolutions, as tested on an A100-80G GPU with 8 input-target pairs.

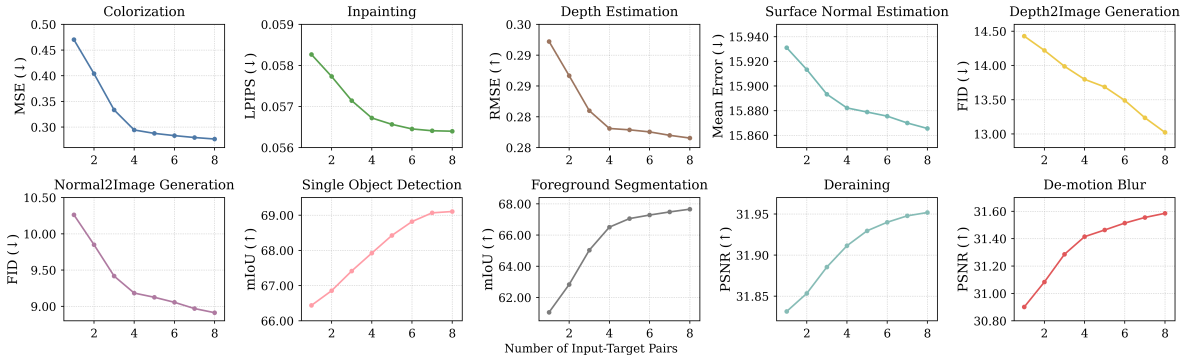


FIGURE 4.19. **Effect of task context length.** Longer task context can consistently improve the performance of downstream tasks.

parallel, allowing them to scale more effectively. This parallelism makes our LaVin-DiT a more suitable choice for large-scale vision applications.

FIGURE 4.20. **Potential application of single-view scene reconstruction.** Given an RGB image and predicted depth map, we lift this image into a 3D space. We illustrate three views of this scene. *Best viewed in color.*

4.4.6 Effect of Task Context Length

In-context learning enables the model to adapt to new tasks using a few examples, with performance generally improving as more examples are provided. We investigate this by assessing the effect of task context length across ten downstream tasks. As shown in Figure 4.19, the model consistently benefits from longer task contexts, achieving notable performance

gains. For instance, with more input-target pairs, LaVin-DiT achieves lower FID in depth-to-image generation and higher PSNR in de-motion blur tasks. These results demonstrate that LaVin-DiT effectively leverages extended task context, highlighting its capacity to utilize additional information for enhanced task adaptation and accuracy.

4.5 Potential Applications

LaVin-DiT opens transformative possibilities for tackling open-world computer vision challenges by unifying diverse vision tasks within a single generative framework. For instance, it can seamlessly generalize across tasks such as text-to-image generation, text-to-video generation, video understanding, 3D reconstruction (Figure 4.20), and 2D/3D visual editing without supervised fine-tuning. By leveraging its spatial-temporal variational autoencoder and joint diffusion transformer, LaVin-DiT excels at capturing the complexity of high-dimensional visual data while maintaining task-specific alignment through in-context learning. This capability positions LaVin-DiT as a foundation model capable of addressing dynamic realistic vision problems, including autonomous driving perception, robotic scene understanding, and interactive AI systems in mixed-reality environments, significantly advancing the frontier of adaptable and scalable AI systems.

4.6 Conclusion

We present LaVin-DiT, a scalable and unified foundation model for computer vision that integrates a spatial-temporal variational autoencoder and a diffusion transformer to efficiently process high-dimensional vision data while preserving spatial and visual coherence. Through in-context learning, LaVin-DiT adapts effectively to a wide range of tasks without fine-tuning, which shows remarkable versatility and adaptability. Extensive experiments validate LaVin-DiT's scalability and performance, positioning it as a promising framework for developing generalist vision models.

Limitations. Despite its advantages, LaVin-DiT is limited by current constraints in large-scale training data, diverse task annotations, and computational resources, especially in comparison to large language models. While our model achieves strong results on seen tasks and related unseen tasks, it struggles with generalization when task definitions deviate significantly from the training distribution. This limitation highlights a key challenge in developing vision models that can generalize effectively to entirely new tasks defined solely by task context.

Future work. Future research should explore scaling LaVin-DiT further in terms of model capacity, dataset diversity, and task complexity to push the boundaries of vision generalization. We anticipate that as these elements expand, LaVin-DiT and similar models may gain the ability to handle arbitrary (out-of-training) vision tasks, guided only by a few input-target pairs. Additionally, investigating methods to select optimal task context automatically could provide a rapid and effective pathway to enhance model performance, ensuring that it leverages the most relevant examples for each task. These directions will drive further advances in developing robust, adaptable, and highly generalized foundation models for computer vision.

CHAPTER 5

Conclusion

This thesis has explored the critical challenges of achieving fine-grained multi-modal alignment, scalable open-vocabulary learning, and unified understanding and generation within computer vision. Over the course of three interconnected research contributions, we have demonstrated that significant progress can be made by rethinking how foundation models leverage and combine visual and linguistic information.

Firstly, we addressed the granularity gap in referring image segmentation with CRIS. Our work showed that by explicitly enforcing text-to-pixel alignment through a vision-language decoder and contrastive learning, the rich, image-level semantics learned by foundation models like CLIP can be effectively transferred to fine-grained pixel-level tasks. This established a crucial bridge between global understanding and local perception, achieving state-of-the-art performance without complex post-processing.

Secondly, we tackled the supervision bottleneck by developing Unpair-Seg, a framework for open-vocabulary segmentation using unpaired data. By leveraging Large Vision-Language Models (LVLMs) to generate pseudo-labels and employing a novel bipartite matching strategy, we proved that robust segmentation can be achieved even when strict, pixel-level annotations are unavailable. This significantly reduces the data annotation burden, making open-vocabulary segmentation more scalable and accessible, and highlighting the power of weakly-supervised learning combined with advanced language models.

Thirdly, we proposed LaVin-DiT, a Large Vision Diffusion Transformer, to unify understanding and generation and overcome the limitations of autoregressive visual modeling. By encoding visual data into a continuous latent space via a Spatial-Temporal Variational

Autoencoder (ST-VAE) and employing in-context learning with a joint diffusion transformer, we demonstrated a scalable and versatile approach. LaVin-DiT can perform over 20 diverse vision tasks, from depth estimation to video generation, without task-specific fine-tuning, showcasing diffusion models as a promising foundation for generalist vision agents that preserve spatial-temporal continuity.

In conclusion, this thesis has demonstrated that by carefully designing multi-modal architectures, leveraging novel supervision paradigms, and embracing generative approaches, we can move closer to creating AI systems that understand and generate visual content with unprecedented fluency and versatility. The developed frameworks provide a foundation for future research towards more general, scalable, and efficient artificial visual intelligence.

Future works. The research presented in this thesis opens several promising avenues for future work. Building upon our findings, we see opportunities to further enhance the capabilities and applicability of multi-modal AI systems. Extending our frameworks, particularly LaVin-DiT, to incorporate 3D understanding and interaction is a natural next step. Integrating representations of 3D geometry and physics into diffusion models could enable embodied agents to perceive and act within real-world environments more effectively. Inspired by Unpair-Seg's use of LVLMs for supervision, future research could focus on creating closed-loop systems where generative models also generate their own high-quality training data for discriminative tasks. This could lead to exponential improvements in model performance and adaptability. While diffusion models offer advantages, their iterative nature can limit real-time performance. Research into techniques like consistency distillation, latent flow matching, or more efficient sampling strategies could significantly speed up inference, making unified models more practical for real-world deployment.

Bibliography

[1] Z. Wang et al., 'Cris: Clip-driven referring image segmentation,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 686–11 695.

[2] Z. Wang et al., 'Open-vocabulary segmentation with unpaired mask-text supervision,' *arXiv preprint arXiv:2402.08960*, 2024.

[3] Z. Wang et al., 'Lavin-dit: Large vision diffusion transformer,' in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 060–20 070.

[4] B. Goertzel, 'Artificial general intelligence: Concept, state of the art, and future prospects,' *Journal of Artificial General Intelligence*, vol. 5, no. 1, pp. 1–48, 2014.

[5] B. M. Lake, T. D. Ullman, J. B. Tenenbaum and S. J. Gershman, 'Building machines that learn and think like people,' *Behavioral and brain sciences*, vol. 40, e253, 2017.

[6] S. Ren, K. He, R. Girshick and J. Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[7] A. Kirillov, K. He, R. Girshick, C. Rother and P. Dollár, 'Panoptic segmentation,' in *CVPR*, 2019, pp. 9404–9413.

[8] R. Zhang, P. Isola and A. A. Efros, 'Colorful image colorization,' in *ECCV*, 2016, pp. 649–666.

[9] A. Kirillov et al., 'Segment anything,' *ICCV*, 2023.

[10] J. Wang et al., 'Mmgen: Unified multi-modal image generation and understanding in one go,' *arXiv preprint arXiv:2503.20644*, 2025.

[11] T. B. Brown, 'Language models are few-shot learners,' *arXiv preprint arXiv:2005.14165*, 2020.

[12] J. Achiam et al., 'Gpt-4 technical report,' *arXiv preprint arXiv:2303.08774*, 2023.

[13] H. Touvron et al., 'Llama: Open and efficient foundation language models,' *arXiv preprint arXiv:2302.13971*, 2023.

[14] G. Comanici et al., 'Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities,' *arXiv preprint arXiv:2507.06261*, 2025.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., 'Language models are unsupervised multitask learners,' *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[16] J. Li, D. Li, C. Xiong and S. Hoi, 'Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,' in *International conference on machine learning*, PMLR, 2022, pp. 12 888–12 900.

[17] X. Zhai, B. Mustafa, A. Kolesnikov and L. Beyer, 'Sigmoid loss for language image pre-training,' in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[18] R. Hu, M. Rohrbach and T. Darrell, 'Segmentation from natural language expressions,' in *European Conference on Computer Vision*, Springer, 2016, pp. 108–124.

[19] L. Yu et al., 'Mattnet: Modular attention network for referring expression comprehension,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.

[20] L. Ye, M. Rochan, Z. Liu and Y. Wang, 'Cross-modal self-attention network for referring image segmentation,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 502–10 511.

[21] L. Ji, Y. Du, Y. Dang, W. Gao and H. Zhang, 'A survey of methods for addressing the challenges of referring image segmentation,' *Neurocomputing*, vol. 583, p. 127 599, 2024.

[22] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu and A. Yuille, 'Recurrent multimodal interaction for referring image segmentation,' in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1271–1280.

[23] R. Li et al., 'Referring image segmentation via recurrent refinement networks,' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5745–5753.

[24] E. Margffoy-Tuay, J. C. Pérez, E. Botero and P. Arbeláez, 'Dynamic multimodal instance segmentation guided by natural language queries,' in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 630–645.

[25] M. Xu et al., 'A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model,' in *ECCV*, 2022, pp. 736–753.

[26] F. Liang et al., 'Open-vocabulary semantic segmentation with mask-adapted clip,' in *CVPR*, 2023, pp. 7061–7070.

[27] W. Zhang, J. Pang, K. Chen and C. C. Loy, 'K-net: Towards unified image segmentation,' in *NeurIPS*, 2021, pp. 10 326–10 338.

[28] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov and R. Girdhar, 'Masked-attention mask transformer for universal image segmentation,' in *CVPR*, 2022, pp. 1290–1299.

[29] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun and R. Ranftl, 'Language-driven semantic segmentation,' in *ICLR*, 2022.

[30] J. Ding, N. Xue, G.-S. Xia and D. Dai, 'Decoupling zero-shot semantic segmentation,' in *CVPR*, 2022, pp. 11 583–11 592.

[31] C. Zhu and L. Chen, 'A survey on open-vocabulary detection and segmentation: Past, present, and future,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[32] Y. Bai et al., 'Sequential modeling enables scalable learning for large vision models,' in *CVPR*, 2024, pp. 22 861–22 872.

[33] X. Dong et al., 'Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model,' *arXiv preprint arXiv:2401.16420*, 2024.

[34] J. Fu et al., 'Dual attention network for scene segmentation,' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[35] J. He, Z. Deng, L. Zhou, Y. Wang and Y. Qiao, 'Adaptive pyramid context network for semantic segmentation,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7519–7528.

[36] T. Wu et al., 'Ginet: Graph interaction network for scene parsing,' in *European Conference on Computer Vision*, Springer, 2020, pp. 34–51.

[37] K. He, G. Gkioxari, P. Dollár and R. Girshick, 'Mask r-cnn,' in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[38] A. Radford et al., 'Learning transferable visual models from natural language supervision,' in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[39] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin and M.-H. Yang, 'Referring expression object segmentation with caption-aware consistency,' *arXiv preprint arXiv:1910.04748*, 2019.

[40] H. Shi, H. Li, F. Meng and Q. Wu, 'Key-word-aware network for referring expression image segmentation,' in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 38–54.

[41] S. Huang et al., 'Referring image segmentation via cross-modal progressive comprehension,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 488–10 497.

[42] T. Hui et al., 'Linguistic structure guided context modeling for referring image segmentation,' in *European Conference on Computer Vision*, Springer, 2020, pp. 59–75.

[43] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov and Y. Cao, 'Simvlm: Simple visual language model pretraining with weak supervision,' *arXiv preprint arXiv:2108.10904*, 2021.

[44] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic and A. Zisserman, 'End-to-end learning of visual representations from uncurated instructional videos,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9879–9889.

[45] H. Luo et al., 'Clip4clip: An empirical study of clip for end to end video clip retrieval,' *arXiv preprint arXiv:2104.08860*, 2021.

[46] H. Fang, P. Xiong, L. Xu and Y. Chen, 'Clip2video: Mastering video-text retrieval via image clip,' *arXiv preprint arXiv:2106.11097*, 2021.

[47] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev and J. Sivic, 'Howto100m: Learning a text-video embedding by watching hundred million narrated video clips,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2630–2640.

[48] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li and X. Li, 'Clip4caption: Clip for video caption,' in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4858–4862.

[49] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or and D. Lischinski, 'Styleclip: Text-driven manipulation of stylegan imagery,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2085–2094.

[50] R. Hadsell, S. Chopra and Y. LeCun, 'Dimensionality reduction by learning an invariant mapping,' in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE, vol. 2, 2006, pp. 1735–1742.

[51] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, 'Momentum contrast for unsupervised visual representation learning,' in *CVPR*, 2020, pp. 9729–9738.

[52] X. Chen, H. Fan, R. Girshick and K. He, 'Improved baselines with momentum contrastive learning,' *arXiv preprint arXiv:2003.04297*, 2020.

[53] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, 'A simple framework for contrastive learning of visual representations,' in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[54] Z. Wu, Y. Xiong, S. X. Yu and D. Lin, 'Unsupervised feature learning via non-parametric instance discrimination,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[55] S. Li, X. Xia, S. Ge and T. Liu, 'Selective-supervised contrastive learning with noisy labels,' *arXiv preprint arXiv:2203.04181*, 2022.

[56] X. Wang, R. Zhang, C. Shen, T. Kong and L. Li, 'Dense contrastive learning for self-supervised visual pre-training,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3024–3033.

[57] P. O. O Pinheiro, A. Almahairi, R. Benmalek, F. Golemo and A. C. Courville, 'Unsupervised learning of dense visual representations,' *Advances in Neural Information Processing Systems*, vol. 33, pp. 4489–4500, 2020.

[58] J. Long, E. Shelhamer and T. Darrell, 'Fully convolutional networks for semantic segmentation,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[59] G. Luo et al., 'Multi-task collaborative network for joint referring expression comprehension and segmentation,' in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 10 034–10 043.

[60] Z. Hu, G. Feng, J. Sun, L. Zhang and H. Lu, 'Bi-directional relationship inferring network for referring image segmentation,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4424–4433.

[61] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li and T. Tan, 'Locate then segment: A strong pipeline for referring image segmentation,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9858–9867.

[62] G. Feng, Z. Hu, L. Zhang and H. Lu, 'Encoder fusion network with co-attention embedding for referring image segmentation,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 506–15 515.

[63] H. Ding, C. Liu, S. Wang and X. Jiang, 'Vision-language transformer and query generation for referring segmentation,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 321–16 330.

[64] K. He, X. Zhang, S. Ren and J. Sun, 'Deep residual learning for image recognition,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[65] A. Vaswani et al., 'Attention is all you need,' *Advances in neural information processing systems*, vol. 30, 2017.

[66] R. Sennrich, B. Haddow and A. Birch, 'Neural machine translation of rare words with subword units,' *arXiv preprint arXiv:1508.07909*, 2015.

[67] R. Liu et al., 'An intriguing failing of convolutional neural networks and the coordconv solution,' *Advances in neural information processing systems*, vol. 31, 2018.

[68] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, 'End-to-end object detection with transformers,' in *European conference on computer vision*, Springer, 2020, pp. 213–229.

[69] J. Lei Ba, J. R. Kiros and G. E. Hinton, 'Layer normalization,' *ArXiv e-prints*, arXiv–1607, 2016.

[70] S. Kazemzadeh, V. Ordonez, M. Matten and T. Berg, 'Referitgame: Referring to objects in photographs of natural scenes,' in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.

[71] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille and K. Murphy, 'Generation and comprehension of unambiguous object descriptions,' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.

[72] T.-Y. Lin et al., 'Microsoft coco: Common objects in context,' in *European conference on computer vision*, Springer, 2014, pp. 740–755.

[73] V. K. Nagaraja, V. I. Morariu and L. S. Davis, 'Modeling context between objects for referring expression understanding,' in *European Conference on Computer Vision*, Springer, 2016, pp. 792–807.

[74] P. Krähenbühl and V. Koltun, 'Efficient inference in fully connected crfs with gaussian edge potentials,' *Advances in neural information processing systems*, vol. 24, 2011.

[75] D. Liu, H. Zhang, F. Wu and Z.-J. Zha, 'Learning to assemble neural module tree networks for visual grounding,' in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4673–4682.

[76] G. Luo et al., 'Cascade grouped attention network for referring expression segmentation,' in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1274–1282.

[77] X. Zou et al., 'Generalized decoding for pixel, image, and language,' in *CVPR*, 2023, pp. 15 116–15 127.

[78] X. Zou et al., 'Segment everything everywhere all at once,' *arXiv preprint arXiv:2304.06718*, 2023.

[79] H. You et al., 'Ferret: Refer and ground anything anywhere at any granularity,' *arXiv preprint arXiv:2310.07704*, 2023.

[80] Q. Yu, J. He, X. Deng, X. Shen and L.-C. Chen, 'Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip,' *NeurIPS*, 2023.

[81] J. Xu et al., 'Groupvit: Semantic segmentation emerges from text supervision,' in *CVPR*, 2022, pp. 18 134–18 144.

[82] J. Cha, J. Mun and B. Roh, 'Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs,' in *CVPR*, 2023, pp. 11 165–11 174.

[83] J. Xu et al., 'Learning open-vocabulary semantic segmentation models from natural language supervision,' in *CVPR*, 2023, pp. 2935–2944.

[84] B. Thomee et al., 'Yfcc100m: The new data in multimedia research,' *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[85] H. Zhang et al., 'Context encoding for semantic segmentation,' in *CVPR*, 2018, pp. 7151–7160.

[86] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen and J. Wang, 'Ocnet: Object context for semantic segmentation,' *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2375–2398, 2021.

[87] B. Cheng, A. Schwing and A. Kirillov, 'Per-pixel classification is not all you need for semantic segmentation,' in *NeurIPS*, 2021, pp. 17 864–17 875.

[88] D. Bolya, C. Zhou, F. Xiao and Y. J. Lee, 'Yolact: Real-time instance segmentation,' in *ICCV*, 2019, pp. 9157–9166.

[89] K. Chen et al., 'Hybrid task cascade for instance segmentation,' in *CVPR*, 2019, pp. 4974–4983.

[90] B. Cheng et al., 'Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation,' in *CVPR*, 2020, pp. 12 475–12 485.

[91] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille and L.-C. Chen, 'Axial-deeplab: Stand-alone axial-attention for panoptic segmentation,' in *ECCV*, 2020, pp. 108–126.

[92] B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen et al., 'Segvit: Semantic segmentation with plain vision transformers,' in *NeurIPS*, 2022, pp. 4971–4982.

[93] C. Liang, W. Wang, J. Miao and Y. Yang, 'Gmmseg: Gaussian mixture based generative semantic segmentation models,' in *NeurIPS*, 2022, pp. 31 360–31 375.

[94] Z. Tian, C. Shen and H. Chen, 'Conditional convolutions for instance segmentation,' in *ECCV*, 2020, pp. 282–298.

[95] X. Wang, R. Zhang, C. Shen, T. Kong and L. Li, 'Solo: A simple framework for instance segmentation,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8587–8601, 2021.

[96] Z. Tian, B. Zhang, H. Chen and C. Shen, 'Instance and panoptic segmentation using conditional convolutions,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 669–680, 2022.

[97] H. Wang, Y. Zhu, H. Adam, A. Yuille and L.-C. Chen, 'Max-deeplab: End-to-end panoptic segmentation with mask transformers,' in *CVPR*, 2021, pp. 5463–5474.

[98] Q. Yu et al., 'K-means mask transformer,' in *ECCV*, 2022, pp. 288–307.

[99] Q. Yu et al., 'Cmt-deeplab: Clustering mask transformers for panoptic segmentation,' in *CVPR*, 2022, pp. 2560–2570.

[100] T. Chen, L. Li, S. Saxena, G. Hinton and D. J. Fleet, 'A generalist framework for panoptic segmentation of images and videos,' in *ICCV*, 2023, pp. 909–919.

[101] K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, 'Masked autoencoders are scalable vision learners,' in *CVPR*, 2022, pp. 16 000–16 009.

[102] C. Jia et al., 'Scaling up visual and vision-language representation learning with noisy text supervision,' in *ICML*, 2021, pp. 4904–4916.

[103] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini and Y. Wu, 'Coca: Contrastive captioners are image-text foundation models,' *arXiv preprint arXiv:2205.01917*, 2022.

[104] P. Wang et al., 'Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution,' *arXiv preprint arXiv:2409.12191*, 2024.

[105] D. Guo et al., 'Seed1.5-vl technical report,' *arXiv preprint arXiv:2505.07062*, 2025.

[106] Z. Wang et al., 'Mosaic representation learning for self-supervised visual pre-training,' in *The eleventh international conference on learning representations*, 2024.

[107] Z. Wang et al., 'Exploring set similarity for dense self-supervised representation learning,' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 590–16 599.

[108] Y. Bai et al., 'Rsa: Reducing semantic shift from aggressive augmentations for self-supervised learning,' *Advances in Neural Information Processing Systems*, vol. 35, pp. 21 128–21 141, 2022.

[109] J.-B. Alayrac et al., 'Flamingo: A visual language model for few-shot learning,' *NeurIPS*, vol. 35, pp. 23 716–23 736, 2022.

[110] M. Cai, J. Yang, J. Gao and Y. J. Lee, 'Matryoshka multimodal models,' in *ICLR 2025*, 2025.

[111] J. Wu et al., 'Visual prompting in multimodal large language models: A survey,' *arXiv preprint arXiv:2409.15310*, 2024.

[112] Y. Li et al., 'Omnibench: Towards the future of universal omni-language models,' *arXiv preprint arXiv:2409.15272*, 2024.

[113] N. Ravi et al., 'Sam 2: Segment anything in images and videos,' *arXiv preprint arXiv:2408.00714*, 2024.

[114] F. Li et al., 'Semantic-sam: Segment and recognize anything at any granularity,' *arXiv preprint arXiv:2307.04767*, 2023.

[115] G. Ghiasi, X. Gu, Y. Cui and T.-Y. Lin, 'Scaling open-vocabulary image segmentation with image-level labels,' in *ECCV*, 2022, pp. 540–557.

[116] H. Zhang et al., 'A simple framework for open-vocabulary segmentation and detection,' in *ICCV*, 2023, pp. 1020–1031.

[117] M. Xu, Z. Zhang, F. Wei, H. Hu and X. Bai, 'Side adapter network for open-vocabulary semantic segmentation,' in *CVPR*, 2023, pp. 2945–2954.

[118] H. Zhou et al., 'Rethinking evaluation metrics of open-vocabulary segmentation,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[119] H. Niu, J. Hu, J. Lin, G. Jiang and S. Zhang, 'Eov-seg: Efficient open-vocabulary panoptic segmentation,' in *AAAI*, vol. 39, 2025, pp. 6254–6262.

[120] Z. Ding, J. Wang and Z. Tu, 'Open-vocabulary universal image segmentation with maskclip,' in *ICML*, 2023.

[121] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang and S. De Mello, 'Open-vocabulary panoptic segmentation with text-to-image diffusion models,' in *CVPR*, 2023, pp. 2955–2966.

[122] H. Luo, J. Bao, Y. Wu, X. He and T. Li, 'Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation,' in *ICML*, 2023, pp. 23 033–23 044.

[123] C. Zhou, C. C. Loy and B. Dai, 'Extract free dense labels from clip,' in *ECCV*, 2022, pp. 696–712.

[124] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai, 'Deformable detr: Deformable transformers for end-to-end object detection,' in *ICLR*, 2020.

[125] R. M. Karp, U. V. Vazirani and V. V. Vazirani, 'An optimal algorithm for on-line bipartite matching,' in *STOC*, 1990, pp. 352–358.

[126] S. Y. Gadre et al., 'Datacomp: In search of the next generation of multimodal datasets,' vol. 36, 2024.

[127] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, 'A convnet for the 2020s,' in *CVPR*, 2022, pp. 11 976–11 986.

[128] G. Ilharco et al., *Openclip*, version 0.1, Jul. 2021. DOI: 10.5281/zenodo.5143773. [Online]. Available: https://doi.org/10.5281/zenodo.5143773.

[129] D. P. Kingma, 'Adam: A method for stochastic optimization,' *arXiv preprint arXiv:1412.6980*, 2014.

[130] I. Loshchilov and F. Hutter, 'Decoupled weight decay regularization,' *arXiv preprint arXiv:1711.05101*, 2017.

[131] X. Gu, T.-Y. Lin, W. Kuo and Y. Cui, 'Open-vocabulary object detection via vision and language knowledge distillation,' *arXiv preprint arXiv:2104.13921*, 2021.

[132] Z. Li et al., 'Panoptic segformer: Delving deeper into panoptic segmentation with transformers,' in *CVPR*, 2022, pp. 1280–1289.

[133] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, 'Scene parsing through ade20k dataset,' in *CVPR*, 2017, pp. 633–641.

[134] M. Everingham, L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, 'The pascal visual object classes (voc) challenge,' *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[135] M. Cordts et al., 'The cityscapes dataset for semantic urban scene understanding,' in *CVPR*, 2016, pp. 3213–3223.

[136] G. Shin, W. Xie and S. Albanie, 'Reco: Retrieve and co-segment for zero-shot transfer,' vol. 35, pp. 33 754–33 767, 2022.

[137] K. Ranasinghe, B. McKinzie, S. Ravi, Y. Yang, A. Toshev and J. Shlens, 'Perceptual grouping in contrastive vision-language models,' in *ICCV*, 2023, pp. 5571–5584.

[138] K. Cai et al., 'Mixreorg: Cross-modal mixed patch reorganization is a good mask learner for open-world semantic segmentation,' in *ICCV*, 2023, pp. 1196–1205.

[139] H. Wang et al., 'Sam-clip: Merging vision foundation models towards semantic and spatial understanding,' *arXiv preprint arXiv:2310.15308*, 2023.

[140] X. Chen, S. Li, S.-N. Lim, A. Torralba and H. Zhao, 'Open-vocabulary panoptic segmentation with embedding modulation,' *ICCV*, 2023.

[141] X. Wang, R. Girdhar, S. X. Yu and I. Misra, 'Cut and learn for unsupervised object detection and instance segmentation,' in *CVPR*, 2023, pp. 3124–3134.

[142] D. Niu, X. Wang, X. Han, L. Lian, R. Herzig and T. Darrell, 'Unsupervised universal image segmentation,' *arXiv preprint arXiv:2312.17243*, 2023.

[143] L. Zhang, S. Zhou, S. Stent and J. Shi, 'Fine-grained egocentric hand-object segmentation: Dataset, model, and applications,' in *ECCV*, 2022, pp. 127–145.

[144] J.-M. Fortin, O. Gamache, V. Grondin, F. Pomerleau and P. Giguère, 'Instance segmentation for autonomous log grasping in forestry operations,' in *IROS*, 2022, pp. 6064–6071.

[145] C. Snyder and M. Do, 'Streets: A novel camera network dataset for traffic flow,' in *NeurIPS*, 2019.

[146] L. Yang et al., 'Ishape: A first step towards irregular shape instance segmentation,' *arXiv preprint arXiv:2109.15068*, 2021.

[147] L. Ciampi, C. Santiago, J. P. Costeira, C. Gennaro and G. Amato, 'Domain adaptation for traffic density estimation.,' in *VISIGRAPP*, 2021, pp. 185–195.

[148] S. Yogamani et al., 'Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving,' in *ICCV*, 2019, pp. 9308–9318.

[149] L. Zhang, L. Jiang, R. Ji and H. Fan, 'Pidray: A large-scale x-ray benchmark for real-world prohibited item detection,' *IJCV*, vol. 131, no. 12, pp. 3170–3192, 2023.

[150] J. Hong, M. Fulton and J. Sattar, 'Trashcan: A semantically-segmented dataset towards visual detection of marine debris,' *arXiv preprint arXiv:2007.08097*, 2020.

[151] H. Liu, C. Li, Q. Wu and Y. J. Lee, 'Visual instruction tuning,' *arXiv preprint arXiv:2304.08485*, 2023.

[152] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson and A. Efros, 'Visual prompting via image inpainting,' in *NeurIPS*, 2022, pp. 25 005–25 017.

[153] J. Lu, C. Clark, R. Zellers, R. Mottaghi and A. Kembhavi, 'Unified-io: A unified model for vision, language, and multi-modal tasks,' in *ICLR*, 2022.

[154] X. Wang, W. Wang, Y. Cao, C. Shen and T. Huang, 'Images speak in images: A generalist painter for in-context visual learning,' in *CVPR*, 2023, pp. 6830–6839.

[155] R. Luo et al., 'Deem: Diffusion models serve as the eyes of large language models for image perception,' *arXiv preprint arXiv:2405.15232*, 2024.

[156] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, 'Indoor segmentation and support inference from rgbd images,' in *ECCV*, 2012, pp. 746–760.

[157] I. Sutskever, 'Sequence to sequence learning with neural networks,' *arXiv preprint arXiv:1409.3215*, 2014.

[158] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, 'High-resolution image synthesis with latent diffusion models,' in *CVPR*, 2022, pp. 10 684–10 695.

[159] X. Wei, Z. Wang, Y. Guo, C. Zhang, T. Liu and M. Gong, 'Training-free robust interactive video object segmentation,' *arXiv preprint arXiv:2406.05485*, 2024.

[160] L. Zhuo et al., 'Lumina-next: Making lumina-t2x stronger and faster with next-dit,' *arXiv preprint arXiv:2406.18583*, 2024.

[161] D. P. Kingma, 'Auto-encoding variational bayes,' *arXiv preprint arXiv:1312.6114*, 2013.

[162] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai and L. Van Gool, 'Multi-task learning for dense prediction tasks: A survey,' *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3614–3633, 2021.

[163] Q. Dong et al., 'A survey on in-context learning,' *arXiv preprint arXiv:2301.00234*, 2022.

[164] J. Wei et al., 'Finetuned language models are zero-shot learners,' in *ICLR*, 2022.

[165] S. Zhang et al., 'Ideal: Influence-driven selective annotations empower in-context learners in large language models,' in *ICLR*, 2024.

[166] R. Lin, B. Han, F. Li and T. Liu, 'Understanding and enhancing the transferability of jailbreaking attacks,' in *The Thirteenth International Conference on Learning Representations*, 2025.

[167] T. Chen, S. Saxena, L. Li, D. J. Fleet and G. Hinton, 'Pix2seq: A language modeling framework for object detection,' *arXiv preprint arXiv:2109.10852*, 2021.

[168] A. Kolesnikov, A. Susano Pinto, L. Beyer, X. Zhai, J. Harmsen and N. Houlsby, 'Uvim: A unified modeling approach for vision with learned guiding codes,' in *NeurIPS*, 2022, pp. 26 295–26 308.

[169] C. Zhao et al., 'Diception: A generalist diffusion model for visual perceptual tasks,' *arXiv preprint arXiv:2502.17157*, 2025.

[170] Z. Wang et al., 'In-context learning unlocked for diffusion models,' in *NeurIPS*, 2023, pp. 8542–8562.

[171] Z. Geng et al., 'Instructdiffusion: A generalist modeling interface for vision tasks,' in *CVPR*, 2024, pp. 12 709–12 720.

[172] C. Qin et al., 'Unicontrol: A unified diffusion model for controllable visual generation in the wild,' in *NeurIPS*, 2024.

[173] W. Van Gansbeke and B. De Brabandere, 'A simple latent diffusion approach for panoptic segmentation and mask inpainting,' *arXiv preprint arXiv:2401.10227*, 2024.

[174] A. Van Den Oord, O. Vinyals et al., 'Neural discrete representation learning,' in *NeurIPS*, 2017.

[175] A. Kolesnikov et al., 'An image is worth 16x16 words: Transformers for image recognition at scale,' in *ICLR*, 2021.

[176] Z. Liu et al., 'Swin transformer: Hierarchical vision transformer using shifted windows,' in *ICCV*, 2021, pp. 10 012–10 022.

[177] H. Fan et al., 'Multiscale vision transformers,' in *ICCV*, 2021, pp. 6824–6835.

[178] W. Peebles and S. Xie, 'Scalable diffusion models with transformers,' in *ICCV*, 2023, pp. 4195–4205.

[179] F. Bao et al., 'All are worth words: A vit backbone for diffusion models,' in *CVPR*, 2023, pp. 22 669–22 679.

[180] H. Zheng, W. Nie, A. Vahdat and A. Anandkumar, 'Fast training of diffusion models with masked transformers,' *arXiv preprint arXiv:2306.09305*, 2023.

[181] S. Gao, P. Zhou, M.-M. Cheng and S. Yan, 'Masked diffusion transformer is a strong image synthesizer,' in *ICCV*, 2023, pp. 23 164–23 173.

[182] F. Bao et al., 'One transformer fits all distributions in multi-modal diffusion at scale,' in *ICML*, 2023, pp. 1692–1717.

[183] S. Chen et al., 'Gentron: Diffusion transformers for image and video generation,' in *CVPR*, 2024, pp. 6441–6451.

[184] K. Crowson, S. A. Baumann, A. Birch, T. M. Abraham, D. Z. Kaplan and E. Shippole, 'Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers,' in *ICML*, 2024.

[185] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, 'Hierarchical text-conditional image generation with clip latents,' *arXiv preprint arXiv:2204.06125*, 2022.

[186] J. Chen et al., 'Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis,' *arXiv preprint arXiv:2310.00426*, 2023.

[187] Y. Tian, Z. Tu, H. Chen, J. Hu, C. Xu and Y. Wang, 'U-dits: Downsample tokens in u-shaped diffusion transformers,' *arXiv preprint arXiv:2405.02730*, 2024.

[188] P. Esser et al., 'Scaling rectified flow transformers for high-resolution image synthesis,' in *ICML*, 2024.

[189] Z. Wan, Y. Xu, Z. Wang, F. Liu, T. Liu and M. Gong, 'Ted-viton: Transformer-empowered diffusion models for virtual try-on,' *arXiv preprint arXiv:2411.17017*, 2024.

[190] A. Gupta et al., 'Photorealistic video generation with diffusion models,' *arXiv preprint arXiv:2312.06662*, 2023.

[191] A. Polyak et al., 'Movie gen: A cast of media foundation models,' *arXiv preprint arXiv:2410.13720*, 2024.

[192] Z. Yang et al., 'Cogvideox: Text-to-video diffusion models with an expert transformer,' *arXiv preprint arXiv:2408.06072*, 2024.

[193] X. Ma et al., 'Latte: Latent diffusion transformer for video generation,' *arXiv preprint arXiv:2401.03048*, 2024.

[194] T. Brooks et al., 'Video generation models as world simulators,' 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators.

[195] J. Zheng, S. Pan, Y. Yao, Z. Wang, D. Wang and T. Liu, 'Aligning what matters: Masked latent adaptation for text-to-audio-video generation,' in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[196] R. Zhang, S. Frei and P. L. Bartlett, 'Trained transformers learn linear models in-context,' *Journal of Machine Learning Research*, vol. 25, no. 49, pp. 1–55, 2024.

[197] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, 'Feature pyramid networks for object detection,' in *CVPR*, 2017, pp. 2117–2125.

[198] Y. Xiong et al., 'Upsnet: A unified panoptic segmentation network,' in *CVPR*, 2019, pp. 8818–8826.

[199] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang and D. Tao, 'Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping,' in *CVPR*, 2019, pp. 2427–2436.

[200] J. Wang, Y. Jiang, Z. Yuan, B. Peng, Z. Wu and Y.-G. Jiang, 'Omnitokenizer: A joint image-video tokenizer for visual generation,' *arXiv preprint arXiv:2406.09399*, 2024.

[201] S. Zhao et al., 'Cv-vae: A compatible video vae for latent generative video models,' *arXiv preprint arXiv:2405.20279*, 2024.

[202] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, 'The unreasonable effectiveness of deep features as a perceptual metric,' in *CVPR*, 2018.

[203] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron and S. Sanghai, 'Gqa: Training generalized multi-query transformer models from multi-head checkpoints,' in *EMNLP*, 2023.

[204] G. Team et al., 'Gemma: Open models based on gemini research and technology,' *arXiv preprint arXiv:2403.08295*, 2024.

[205] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo and Y. Liu, 'Roformer: Enhanced transformer with rotary position embedding,' *Neurocomputing*, vol. 568, p. 127 063, 2024.

[206] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel and M. Le, 'Flow matching for generative modeling,' in *ICLR*, 2023.

[207] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, 'Imagenet: A large-scale hierarchical image database,' in *CVPR*, 2009, pp. 248–255.

[208] S. Shao et al., 'Objects365: A large-scale, high-quality dataset for object detection,' in *ICCV*, 2019, pp. 8430–8439.

[209] K. Soomro, 'Ucf101: A dataset of 101 human actions classes from videos in the wild,' *arXiv preprint arXiv:1212.0402*, 2012.

[210] W. Kay et al., 'The kinetics human action video dataset,' *arXiv preprint arXiv:1705.06950*, 2017.

[211] K. Greff et al., 'Kubric: A scalable dataset generator,' in *CVPR*, 2022, pp. 3749–3761.

[212] L. Yang et al., 'Depth anything v2,' *arXiv preprint arXiv:2406.09414*, 2024.

[213] C. Ye et al., 'Stablenormal: Reducing diffusion variance for stable and sharp normal,' *arXiv preprint arXiv:2406.16864*, 2024.

[214] S. Rajbhandari, J. Rasley, O. Ruwase and Y. He, 'Zero: Memory optimizations toward training trillion parameter models,' in *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, pp. 1–16.

[215] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler and V. Koltun, 'Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,' *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[216] G. Bae and A. J. Davison, 'Rethinking inductive biases for surface normal estimation,' in *CVPR*, 2024, pp. 9535–9545.

[217] A. Shaban, S. Bansal, Z. Liu, I. Essa and B. Boots, 'One-shot learning for semantic segmentation,' *arXiv preprint arXiv:1709.03410*, 2017.

[218] R. Ranftl, A. Bochkovskiy and V. Koltun, 'Vision transformers for dense prediction,' in *ICCV*, 2021, pp. 12 179–12 188.

[219] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt and K. Schindler, 'Repurposing diffusion-based image generators for monocular depth estimation,' in *CVPR*, 2024, pp. 9492–9502.