International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02171-y

LLMFormer: Large Language Model for Open-Vocabulary Semantic Segmentation

Hengcan Shi[1,2] · Son Duy Dao[2] · Jianfei Cai[2]

Received: 16 December 2023 / Accepted: 28 June 2024

Abstract

Open-vocabulary (OV) semantic segmentation has attracted increasing attention in recent years, which aims to recognize objects in an open class set for real-world applications. While prior OV semantic segmentation approaches have relied on additional semantic knowledge derived from vision-language (VL) pre-training, such as the popular CLIP model, this paper introduces a novel paradigm by harnessing the unprecedented capabilities of large language models (LLMs). Inspired by recent breakthroughs in LLMs that provide a richer knowledge base compared to traditional vision-language pre-training, our proposed methodology capitalizes on the vast knowledge embedded within LLMs for OV semantic segmentation. Particularly, we partition LLM knowledge into object, attribute, and relation priors, and propose three novel attention modules-semantic, scaled visual, and relation attentions, to utilize the LLM priors. Extensive experiments are conducted on common benchmarks including ADE20K (847 classes) and Pascal Context (459 classes). The results show that our model outperforms previous state-of-the-art (SoTA) methods by up to 7.2% absolute. Moreover, unlike previous VL-pre-training-based works, our method can even predict OV segmentation results without target candidate classes.

Keywords Open-vocabulary · Semantic segmentation · Large language model

1 Introduction

Semantic segmentation aims to extract masks for all objects in an image, which serves as a fundamental and vital step for many real-world applications, such as object extraction (Fan & Zhang, 2023; Zhang et al., 2023; Lin et al., 2023), robot navigation (Hu et al., 2023; Li et al., 2023) and multimedia retrieval (Ma et al., 2023; Shi et al., 2022; Wang et al., 2023; Shi et al., 2023). Prior segmentation methods (Li et al., 2017; He et al., 2020) are usually trained on a limited dataset and focus on recognizing a fixed number of object categories, as shown in Fig. 1a. They cannot handle real-world applications that require segmenting diverse novel objects. This leads to the recent trend in open-vocabulary (OV) semantic segmentation that aims to generate masks for objects in an open-category set.

To recognize open-set objects, OV semantic segmentation methods (Liang et al., 2023; Ding et al., 2022; Xu et al., 2023; Yu et al., 2023) usually leverage extra knowledge to extend their semantic spaces, as illustrated in Fig. 1b, and they can generally be categorized into two groups: one- and two-stage methods. Early works (Xu et al., 2021; Ghiasi et al., 2022) often adopt a two-stage strategy, which decouples OV semantic segmentation into two sub-tasks: mask proposal generation and classification. The proposal generation step uses pre-trained segmentation models to generate class-agnostic masks to capture as many objects as possible. The mask classification step often employs vision-language (VL) pre-training feature extractors such as CLIP (Radford et al., 2021) to classify class-agnostic mask proposals to recognize OV objects. Although two-stage approaches are intuitive, they highly rely on well-trained mask proposal generators. To address the issue, recent studies (Xu et al., 2023; Yu et al., 2023) move to the one-stage architecture, which directly

Communicated by Zhun Zhong.

✉ Hengcan Shi
shihengcan@gmail.com

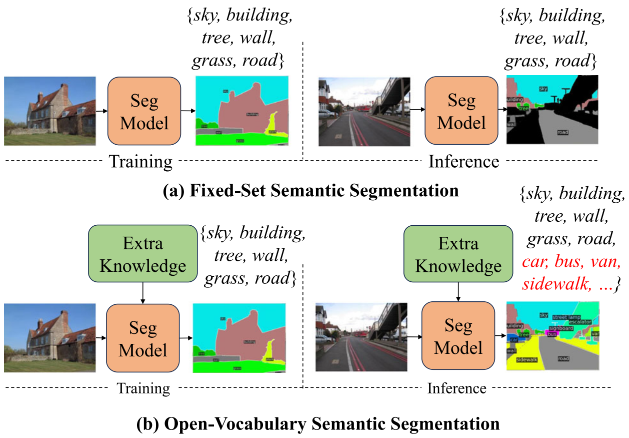
Son Duy Dao
duy.dao@monash.edu

Jianfei Cai
jianfei.cai@monash.edu

[1] The College of Electrical and Information Engineering, Hunan University, Changsha, China

[2] Department of Data Science & AI, Monash University, Melbourne, Australia

Published online: 16 August 2024



(a) Fixed-Set Semantic Segmentation

(b) Open-Vocabulary Semantic Segmentation

Fig. 1 Comparison between fixed-set and open-vocabulary semantic segmentation. **a** Fixed-set methods are trained with segmentation data and only recognize limited objects in their training data during inference. **b** Open-vocabulary (OV) semantic segmentation also uses segmentation data for training, but is expected to segment diverse novel objects (red) during inference, which are typically addressed by leveraging extra knowledge

leverages VL pre-training to simultaneously predict masks and their classes. Despite notable progress, most of these existing OV semantic segmentation methods mainly utilize VL pre-training models to extract embeddings, which only provides limited implicit semantic information.

On the other hand, in the past few years, we have witnessed the huge success of large language models (LLMs), which can provide more comprehensive understanding of scenes. This motivates us to consider: *Can we leverage LLM knowledge to address the challenges in OV semantic segmentation?* Particularly, we observe that three types of knowledge in LLM descriptions are useful for OV semantic segmentation. Firstly, object names (e.g., *glass ceiling*, *statues* and *chairs* in Fig. 2) in LLM descriptions indicate potential objects, which can help OV object discovery, mask prediction and classification. Secondly, object attributes often provide diverse cues for segmentation, such as object sizes (e.g., *large*), numbers (e.g., *several*) and appearances (e.g., *naked*). Finally, object relations (e.g., *in* and *near*) can provide valuable context.

Based on these observations, in this paper, we propose a novel LLMFormer that exploits LLM knowledge as priors to improve OV semantic segmentation. Specifically, we extract three useful knowledge from LLMs for OV semantic segmentation, i.e., object names, attributes and relations. Three attention models are introduced to leverage these priors. Firstly, a semantic attention mechanism is proposed to incorporate object name and attribute priors into mask embeddings for OV object discovery and classification. Secondly, considering many attributes indicate object sizes, we propose a scaled visual attention module to segment objects of different sizes based on attribute priors. Finally, we introduce relation attention to incorporate LLM relation priors for better OV semantic segmentation.

Our key contributions can be summarized as follows.

- This is a pioneering work in proposing the idea of exploiting the comprehensive knowledge of LLMs, beyond the conventional object names, for OV semantic segmentation.
- We propose LLMFormer, which consists of three novel attention modules: semantic, scaled visual and relation attentions, to leverage different LLM knowledge (objects, attributes and relations) for OV semantic segmentation.
- Extensive experiments on ADE20K, Pascal Context and Pascal VOC show that our method significantly outperforms the state-of-the-art solutions. Moreover, our method also shows the ability to predict OV results without pre-defined candidate classes, which is more practical in real-world applications.

2 Related Work

2.1 Fixed-Set Semantic Segmentation

Early methods use CNN-based architectures to deal with the fixed-set semantic segmentation task. FCN (Long et al., 2015) designs an encoder-decoder structure, where a CNN encoder extracts image features and a CNN decoder classifies each pixel in the feature maps. Nevertheless, the vanilla FCN (Long et al., 2015) is easy to lose image details, due to too many down-sampling layers in the encoder. To capture more details, Chen et al. (2018) and Yu and Koltun (2016) replace a number of downsampling layers with atrous convolutions and dilated convolutions in the encoder, respectively. Noh et al. (2015) design a deconvolutional decoder to gradually restore more details, mirroring the CNN encoder. Although these approaches make significant progress for recognizing objects at fixed size ranges, they struggle to segment objects of diverse sizes. To segment variable sized objects, many prior works (Lin et al., 2017; Zhao et al., 2017; Chen et al., 2018; Shi et al., 2018; Chen et al., 2016; Li et al., 2020) use multi-scale combinations, such as by pyramid pooling module (PPM), pyramid atrous convolutions and feature pyramid network (FPN). Some approaches (Lin et al., 2018; Liu et al., 2015; Shi et al., 2019; Chen et al., 2018; Zhang et al., 2018; Ding et al., 2018; Shi et al., 2018) model global context of objects to better understand the whole scene.

With the development of transformers in recent years, many works leverage ViTs to model long-range dependencies to improve the semantic segmentation performance. DPT (Ranftl et al., 2021) and Zheng et al. (2021) use transformers as encoders to extract feature maps and use

CNN-based decoders to generate semantic segmentation results. Strudel et al. (2021) and Xie et al. (2021) present transformer decoders that take object classes as queries to classify image regions. Many methods such as P2T (Wu et al., 2021), PVT (Wang et al., 2021), Focal Transformer (Yang et al., 2021) and Liu et al. (2021) design pyramid encoders to capture multi-scale features for better segmentation. SenFormer (Bousselham et al., 2022), PFT (Qin et al., 2022) and TSG (Shi et al., 2023) further propose multi-scale decoders. MaskFormer (Cheng et al., 2021) and Mask2Former (Cheng et al., 2022) develop instance-level transformer decoders and combine them with pixel-level decoders to predict segmentation results. Based on the Mask2Former (Cheng et al., 2022) decoder, Oneformer (Jain et al., 2023) presents a multi-dataset training approach, which allows joint training on semantic, instance and panoptic segmentation datasets. Nonetheless, these methods are hard to leverage the powerful knowledge from large pre-trained models, due to different network architectures. ViT-Adapter (Chen et al., 2023) employs the adapter mechanism to transfer knowledge from large pre-trained ViTs. These works are foundations of our OV segmentation method. However, they can only recognize fixed-set objects, while our method focuses on exploiting LLM knowledge to improve the OV segmentation ability.

2.2 Open-Vocabulary Semantic Segmentation

OV semantic segmentation is an emerging problem that requires a trained model to segment any arbitrary concepts during testing without the need for retraining or adaptation. Previous OV semantic works can be mainly categorized into two types: one- and two-stage. Two-stage methods leverage separate models for mask proposal generation and classification. They first train a model or employ a pre-trained method to generate mask proposals, which are then fed into a VL pre-training model for classification. Xu et al. (2021); Ding et al. (2022) train mask proposal generation models with segmentation datasets to extract class-agnostic masks, and then these mask proposals are classified by CLIP (Radford et al., 2021). OpenSeg (Ghiasi et al., 2022) introduces image-text pairs to train the classification network. DeOP (Han et al., 2023) optimizes network connection by a decoupled and single-pass framework. CEL (Dao et al., 2023) proposes background learning to improve the OV training. Qi et al. (2022) only generate masks without their classes. Jaus et al. (2023) presents a contrastive-learning-based unsupervised method to recognize OV panoptic objects in panoramic images. These methods are intuitive. Nevertheless, they highly rely on well-trained mask proposal generation models. Moreover, since two-stage methods involve two heavy models, they usually require high computational costs.

Therefore, one-stage methods Xu et al. (2022); Shi et al. (2024); Liang et al. (2023); Xu et al. (2023) are proposed, which generate masks and their classes simultaneously by a single model. Xu et al. (2022) trains a ViT model with image-text pairs and leverages group tokens to generate segmentation results. ODISE Xu et al. (2023) employs a fixed-set semantic segmentation network (Cheng et al., 2022), and uses pre-trained Stable Diffusion (Takagi & Nishimoto, 2023) as the image encoder for OV segmentation. (Xu et al., 2023) introduces a side adaptation network that simultaneously learns mask proposals and mask classification from the pre-trained CLIP image encoder. FC-CLIP (Yu et al., 2023) improves input resolutions to achieve finer segmentation results. AttrSeg (Ma et al., 2023) adds category attribute descriptions to reduce ambiguous categories and recognize indescribable categories. HIPIE (Wang et al., 2023) unifies semantic-, instance- and part-level segmentation tasks. Some works (Zhang et al., 2023; Xu et al., 2023; Liang et al., 2023) propose new training strategies. OpenSeed (Zhang et al., 2023) combines detection and segmentation data to boost the training. OVSegmentor (Xu et al., 2023) adopts web training data and presents cross-modal as well as cross-image consistency to improve OV training. OV-Seg (Liang et al., 2023) proposes visual prompt tuning to improve the OV training. To reduce the reliance of visual supervisions, FOSSIL (Barsellotti et al., 2024) leverages pre-trained diffusion models to generate text-conditioned visual embeddings. CLIP-DIY (Wysoczanska et al., 2024) employs pre-trained CLIP to classify multi-scale image patches as coarse segmentation results, and refines them by existing segmentation technologies. However, most of these methods only extract knowledge from VL pre-training models or self-trained models, which only provide implicit semantic information. Unlike them, we exploit more comprehensive knowledge from LLMs, and we propose three types of attention modules to guide OV semantic segmentation based on LLM priors.

2.3 Large Language Model

LLMs provide powerful knowledge for many real-world applications. Different from early VL pre-training models (Radford et al., 2021), which only learn embeddings to model semantic spaces, LLMs (e.g., GPT-4 OpenAI (2023), MiniGPT-4 Zhu et al. (2023) and LLAMA Touvron et al. (2023)) are usually trained on a mass of language data to obtain comprehensive and complex reasoning abilities. Some recent models such as Liu et al. (2023) and Dai et al. (2023) employ vision-language data and instruction learning to allow LLMs to understand diverse visual content, namely multi-modal large language models (MLLMs). These LLMs are based on question answering (QA) or visual question answering (VQA) mechanisms that flexibly generate answers for various tasks. Based on visual input and language questions, Peng et al. (2023), Zhang et al. (2023) and SoM Yang et al. (2023) further extend LLMs to the region level to achieve

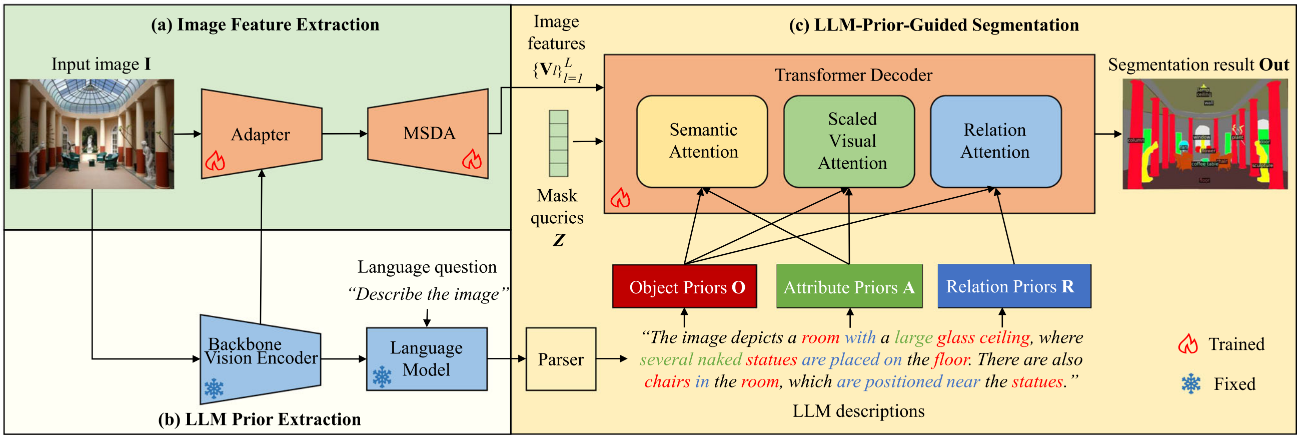


Fig. 2 Illustration of our proposed LLMFormer for OV semantic segmentation. Our model contains three parts: **a** image feature extraction extracts multi-scale image features; **b** LLM prior extraction extracts prior knowledge from the LLM for segmentation; **c** LLM-prior-guided segmentation divides LLM knowledge into three types: objects (red), attributes (green) and relations (blue), and utilizes them via semantic, scaled visual and relation attention modules

more fine-grained visual understanding and reasoning. Lai et al. (2023) proposes a reasoning segmentation task and leverages LLMs for complex reasoning. Our work is built upon these LLMs, but different from them. LLMs aim at general representations and predictions, while our method focuses on leveraging LLMs to boost the OV semantic segmentation performance.

3 Our Method

In this section, we first describe the definition of the OV semantic segmentation problem and the overall architecture of our proposed LLMFormer in Sect. 3.1. Then, we discuss the details of our method in Sect. 3.2–3.4. Finally, the training strategy of our model is presented in Sect. 3.5.

3.1 Problem Definition and Overview

Consider an input RGB image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, where H and W are the height and width. OV semantic segmentation expects to predict a pixel-wise classification map $\mathbf{Out} \in \mathbb{R}^{H\times W}$, in which each element indicates the class of the corresponding pixel in the image. OV methods typically generate segmentation results by mask-text alignment, where N masks are generated and aligned with C candidate classes on the target application.

Figure 2 shows the overall architecture of our proposed LLMFormer, which consists of three parts: *Image Feature Extraction*, *LLM Prior Extraction* and *LLM-Prior-Guided Segmentation*. (a) The *Image Feature Extraction* part contains a vision encoder from multi-modal large language models (MLLMs), an adapter and an MSDA (multi-scale deformable attention) to capture multi-scale image features. (b) The *LLM Prior Extraction* part is to extract comprehensive prior knowledge from LLMs. (c) The *LLM-Prior-Guided Segmentation* part introduces novel semantic, scaled visual and relation attentions to decouple LLM priors and guide OV semantic segmentation. Next, we introduce each module in detail.

3.2 Image Feature Extraction

Our image feature extraction module takes the image $\mathbf{I}$ as input, and generates image feature maps, as shown in Fig. 3. To better segment objects of different sizes, we generate multi-scale feature maps. We use the vision encoder of an MLLM (such as LLAVA Liu et al. (2023)), while our model can use any image encoder. ViT Adapter Chen et al. (2023) is leveraged to transform the feature maps from the general vision encoder to our target domain. Inspired by fixed-set segmentation methods (Cheng et al., 2021, 2022), an MSDA (multi-scale deformable attention) module is also adopted to refine multi-scale feature maps. Let $\{\mathbf{V}_l \in \mathbb{R}^{H_l\times W_l\times D_l}\}_{l=1}^{L}$ denote the refined feature maps, where L is the number of scales, and H_l, W_l and D_l are the height, width and channel number for the l-th feature map, respectively.

3.3 LLM Prior Extraction

In this subsection, we extract comprehensive knowledge of LLMs for OV semantic segmentation. Current MLLMs are usually based on the visual question answering (VQA) architecture. They contain a vision encoder to obtain general features of the input image, and a language model to generate answers from a question and the image features, as

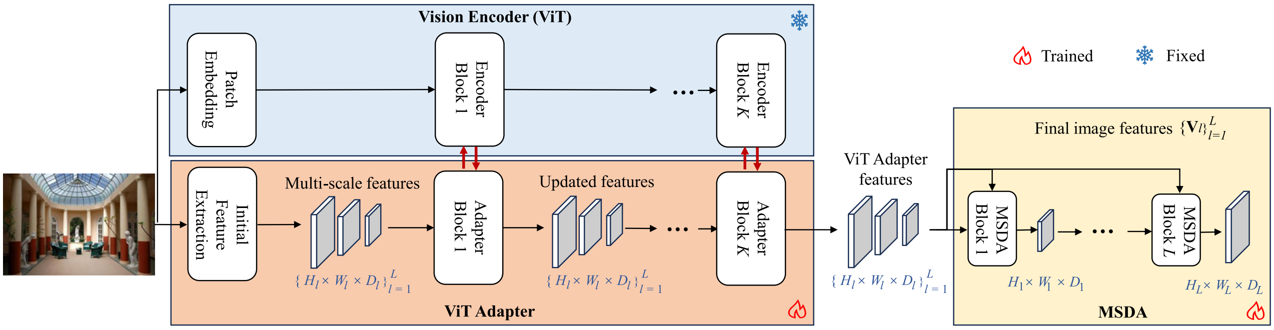

Vision Encoder (ViT)
Patch Embedding
Encoder Block 1
Encoder Block K
Trained
Fixed
Initial Feature Extraction
Multi-scale features
$\{H_l \times W_l \times D_l\}_{l=1}^{L}$
Adapter Block 1
Updated features
$\{H_l \times W_l \times D_l\}_{l=1}^{L}$
Adapter Block K
ViT Adapter
ViT Adapter features
$\{H_l \times W_l \times D_l\}_{l=1}^{L}$
Final image features $\{\mathbf{V}_l\}_{l=1}^{L}$
MSDA Block 1
$H_1 \times W_1 \times D_1$
MSDA Block L
$H_L \times W_L \times D_L$
MSDA

Fig. 3 The flow of our image feature extraction. We adopt ViT Adapter Chen et al. (2023) to extract knowledge from the pre-trained vision encoder (ViT) to discover open-vocabulary objects. Specifically, an initial feature extraction module is first used to generate multi-scale image features. Then, K adapter blocks progressively embed knowledge from K vision encoder blocks into our multi-scale image features. Finally, an MSDA module (Cheng et al., 2022) is leveraged to gradually refine the feature map on each scale, and we take these refined features as our final image features $\{\mathbf{V}_l\}_{l=1}^{L}$

shown in Fig. 2. While any VQA-based MLLM can be used in our model, here we adopt Liu et al. (2023). We input questions such as 'Describe the image' to obtain comprehensive descriptions of an image.

As discussed in Sect. 1, we want to extract object, attribute and relation priors from the MLLM. To do that, we employ language parsing tools (e.g., Schuster et al. (2015)) to extract nouns as *object priors*, adjectives related to an object as its *attribute priors*, and verbs or prepositions related to multiple objects as their *relation priors*. Textual embeddings of these words are also extracted from the MLLM. Let $\mathbf{O} \in \mathbb{R}^{M \times D}$, $\mathbf{A} \in \mathbb{R}^{M \times D}$ and $\mathbf{R} \in \mathbb{R}^{K \times D}$ represent the embeddings of object, attribute and relation priors, respectively, where M is the number of objects, K is the count of object relations, and D is the dimension of textual embeddings. If a prior contains multiple words, we average their embeddings. Moreover, we generate M attribute prior embeddings to align with object priors. Similarly, if an object is related to multiple attributes, we use the average of the embeddings to represent the object attribute prior. If there is no attribute for an object, we set an all-zero vector as its corresponding attribute prior. Note that LLMs usually cannot describe all objects as well as their attributes and relations in an image. Thus, we use LLM priors as guidance, but do not completely rely on them.

3.4 LLM-Prior-Guided Segmentation

Here, we build a transformer-based decoder to generate segmentation results guided by LLM priors. As illustrated in Fig. 4, we set N learnable mask embeddings $\mathbf{Z} \in \mathbb{R}^{N \times D_Z}$ as queries in our decoder, where each embedding is a D_Z-dimensional vector. $\mathbf{Z}$ is randomly initialized and can be learned during training. Our transformer decoder contains L_{dec} blocks, and each block consists of three main components: semantic, scaled visual and relation attentions.

Semantic Attention. Semantic attention embeds object and attribute priors into mask embeddings, to leverage object classes and appearances to enhance OV object discovery, mask prediction and classification. It also captures the correspondences between object priors and masks for subsequent attention modules. Our semantic attention is a multi-head cross-attention model:

$$\begin{aligned}\mathbf{Z}^S, \mathbf{Att}^S = MHCA(&query = \mathbf{Z},\\ &key = \mathbf{O} + \mathbf{A},\\ &value = \mathbf{O} + \mathbf{A})\end{aligned} \tag{1}$$

where $MHCA(\cdot, \cdot, \cdot)$ is multi-head cross-attention with addition and normalization (Vaswani et al., 2017). We take mask embeddings $\mathbf{Z}$ as queries, and the sum of object and attribute prior embeddings as keys and values in this attention. The outputs $\mathbf{Z}^S \in \mathbb{R}^{N \times D_S}$ are updated mask embeddings, which incorporate object and attribute priors from LLMs. $\mathbf{Att}^S \in \mathbb{R}^{N \times M}$ is the average of attention maps from all heads, which captures the relationships between N masks and M LLM object priors. Each element $a_{n,m}^S$ in $\mathbf{Att}^S$ is from 0 to 1, and a high $a_{n,m}^S$ means that the n-th mask is highly related to the m-th LLM object.

Scaled Visual Attention. The updated mask embeddings $\mathbf{Z}^S$ are then input into scaled visual attention to embed visual information $\{\mathbf{V}_l\}_{l=1}^{L}$. We find that many attributes indicate object sizes. For example, the attribute *large* in Fig. 2 directly describes the object size, and the word *several* indicates that the corresponding object class involves many image regions. Therefore, we propose to leverage attribute priors for scale selection to better segment objects in different sizes.

Concretely, we expect to select suitable visual feature maps from $\{\mathbf{V}_l\}_{l=1}^{L}$ for each mask based on attribute priors $\mathbf{A}$. To this end, we first generate attribute embeddings for every mask:

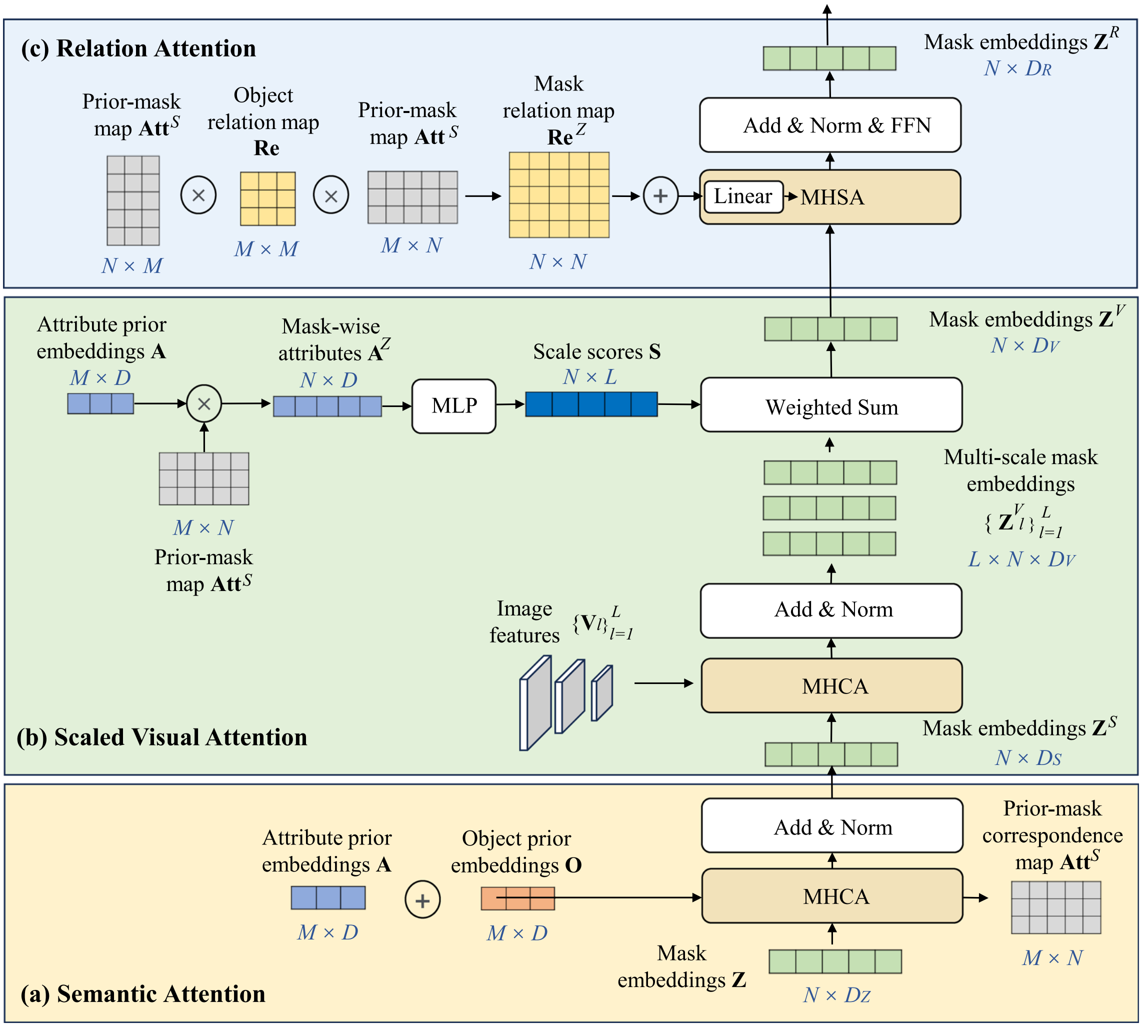

(c) Relation Attention
Prior-mask map $\mathbf{Att}^S$
$N \times M$
Object relation map $\mathbf{Re}$
$M \times M$
Prior-mask map $\mathbf{Att}^S$
$M \times N$
Mask relation map $\mathbf{Re}^Z$
$N \times N$
Linear
MHSA
Add & Norm & FFN
Mask embeddings $\mathbf{Z}^R$
$N \times D_R$
(b) Scaled Visual Attention
Attribute prior embeddings $\mathbf{A}$
$M \times D$
Mask-wise attributes $\mathbf{A}^Z$
$N \times D$
MLP
Scale scores $\mathbf{S}$
$N \times L$
Weighted Sum
Mask embeddings $\mathbf{Z}^V$
$N \times D_V$
$M \times N$
Prior-mask map $\mathbf{Att}^S$
Multi-scale mask embeddings
$\{\mathbf{Z}^V_l\}_{l=1}^L$
$L \times N \times D_V$
Add & Norm
Image features $\{\mathbf{V}_l\}_{l=1}^L$
MHCA
Mask embeddings $\mathbf{Z}^S$
$N \times D_S$
(a) Semantic Attention
Attribute prior embeddings $\mathbf{A}$
$M \times D$
Object prior embeddings $\mathbf{O}$
$M \times D$
Add & Norm
MHCA
Prior-mask correspondence map $\mathbf{Att}^S$
$M \times N$
Mask embeddings $\mathbf{Z}$
$N \times D_Z$

Fig. 4 Our semantic, scaled visual and relations attentions. **a** Semantic attention incorporates object name and attribute priors into mask embeddings for OV object discovery and classification. **b** Scaled visual attention is to segment objects of different sizes based on attribute priors. **c** Relation attention incorporates LLM relation priors into attention learning to better model object relationships for OV semantic segmentation

$$\mathbf{A}^Z = \mathbf{Att}^S \times \mathbf{A} \tag{2}$$

where × denotes matrix multiplication. We leverage the prior-mask correspondence matrix $\mathbf{Att}^S$ to transform object-prior-wise attributes $\mathbf{A}$ into mask-wise ones $\mathbf{A}^Z \in \mathbb{R}^{N \times D}$.

Next, a two-layer MLP (multilayer perceptron) with Softmax is used to predict scale selection scores for every mask

$$\mathbf{S} = Softmax(MLP(\mathbf{A}^Z)) \tag{3}$$

where $\mathbf{S} \in \mathbb{R}^{N \times L}$ are scale selection scores. In $\mathbf{S}$, each element $s_{n,l}$ is from 0 to 1 and means the confidence of choosing the l-th image feature map to segment the n-th mask.

On the other hand, we embed each image feature map $\mathbf{V}_l$ into mask embedding by a multi-head cross-attention with addition and normalization as

$$\begin{aligned} \mathbf{Z}_l^V = MHCA(&query = \mathbf{Z}^S, \\ &key = \mathbf{V}_l \\ &value = \mathbf{V}_l) \end{aligned} \tag{4}$$

where mask embeddings $\mathbf{Z}^S$ are queries, the image feature map $\mathbf{V}_l$ is used as keys and values, and $\mathbf{Z}_l^V \in \mathbb{R}^{N \times D_V}$ are the output mask embeddings which encode visual information from the l-th image feature map. In this way, we generate L updated mask embeddings $\{\mathbf{Z}_l^V\}_{l=1}^L$, where each vector $\mathbf{z}_{n,l}^V \in \mathbb{R}^{D_V}$ denotes the embedding of the n-th mask in the l-th scale.

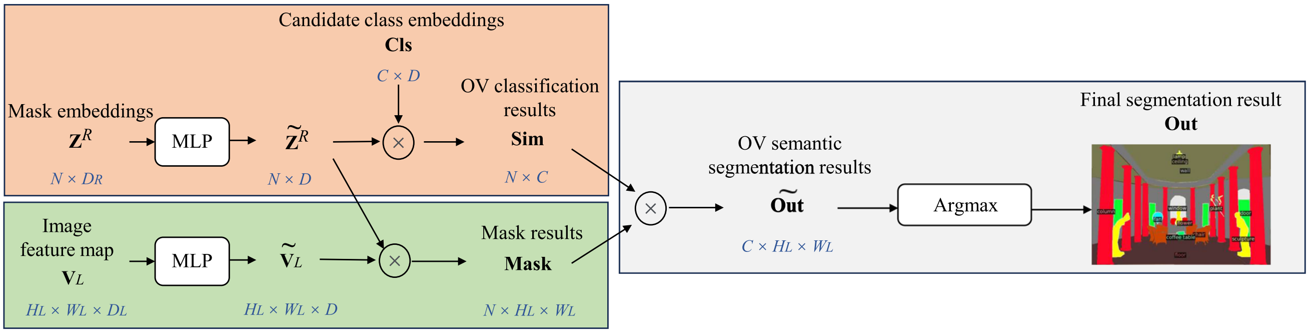

Candidate class embeddings
Cls
C × D
Mask embeddings
$\mathbf{Z}^R$
$N \times D_R$
MLP
$\widetilde{\mathbf{Z}}^R$
$N \times D$
OV classification results
Sim
$N \times C$
Image feature map
$\mathbf{V}_L$
$H_L \times W_L \times D_L$
MLP
$\widetilde{\mathbf{V}}_L$
$H_L \times W_L \times D$
Mask results
Mask
$N \times H_L \times W_L$
OV semantic segmentation results
$\widetilde{\mathbf{Out}}$
$C \times H_L \times W_L$
Argmax
Final segmentation result
Out

Fig. 5 Pipeline of generating OV semantic segmentation results. We generate mask results **Mask** based on mask embeddings $\mathbf{Z}^R$ and image features $\mathbf{V}_L$, while classifying each mask by calculating the similarities **Sim** between mask embeddings $\mathbf{Z}^R$ and candidate class embeddings **Cls**. The final segmentation results **Out** is the combination of **Mask** and **Sim**

Finally, we leverage scale selection scores **S** to choose mask embeddings of suitable scales as

$$\mathbf{z}_n^V = \sum_{l=1}^{L} s_{n,l}\, \mathbf{z}_{n,l}^V \,. \tag{5}$$

For each mask n, we calculate the weighted sum of its embeddings in different feature maps $\{\mathbf{Z}_l^V\}_{l=1}^L$, and the weights are corresponding scale selection scores $s_{n,l}$ in **S**. $\mathbf{z}_n^V$ is the final embedding for the n-th mask, and $\mathbf{Z}^V \in \mathbb{R}^{N\times D_V}$ is the embedding map for all masks.

Relation Attention. We input $\mathbf{Z}^V$ to our relation attention to learn mask relationships with LLM relation prior guidances. We generate a 0-1 object relation map $\mathbf{Re} \in \mathbb{R}^{M\times M}$ from relation prior **R**. If two objects are related in LLM descriptions, the corresponding value in **Re** is 1; otherwise, the value is 0. Then, we map **Re** to mask-level as

$$\mathbf{Re}^Z = \mathbf{Att}^S \times \mathbf{Re} \times (\mathbf{Att}^S)^T \tag{6}$$

where $\mathbf{Re}^Z \in \mathbb{R}^{N\times N}$ indicates the LLM relation prior for every mask pairs, and each value in $\mathbf{Re}^Z$ is in [0, 1].

These priors is encoded into mask embeddings by a multi-head self-attention as follows,

$$\begin{aligned} \mathbf{Z}^R = MHSA(&query = \mathbf{Z}^V, \\ &key = \mathbf{Z}^V, \\ &value = \mathbf{Z}^V, \\ &attention = \widetilde{\mathbf{Att}}) \end{aligned} \tag{7}$$

$$\widetilde{\mathbf{Att}} = Softmax(Linear(\mathbf{Att} + \mathbf{Re}^Z)) \tag{8}$$

where $MHSA(\cdot,\cdot,\cdot)$ is multi-head self-attention with addition, normalization and FFN in transformers (Vaswani et al., 2017). We take $\mathbf{Z}^V$ as queries, keys and values to model the relationships among masks. $\mathbf{Att} \in \mathbb{R}^{N\times N}$ is the original attention map. We use a linear layer to integrate the transformer attention map **Att** and LLM relation priors $\mathbf{Re}^Z$. A Softmax is leveraged to normalize the integrated attention map. $\widetilde{\mathbf{Att}} \in \mathbb{R}^{N\times N}$ denotes the integrated and normalized attention, and we use it to embed both visual and prior relations into mask embeddings. $\mathbf{Z}^R \in \mathbb{R}^{N\times D_R}$ is the output of each decoder block, where D_R is its dimension. $\mathbf{Z}^R$ in the final block is the output of our entire decoder, which embeds mask relations, scaled visual information, and object, attribute and relation priors for OV semantic segmentation.

Segmentation Result Generation. Similar to fixed-set methods (Cheng et al., 2021, 2022), mask embeddings $\mathbf{Z}^R$ and the largest pixel-level image feature map $\mathbf{V}_L$ are used to generate mask results, as shown in Fig. 5. We first leverage two two-layer MLPs to convert the dimensions of $\mathbf{Z}^R$ and $\mathbf{V}_L$ into the same. $\widetilde{\mathbf{Z}}^R \in \mathbb{R}^{N\times D}$ and $\widetilde{\mathbf{V}}_L \in \mathbb{R}^{H_L\times W_L\times D}$ are the transformed features, and the dimension D is as the same as textual embeddings.

Mask results are generated by

$$\mathbf{Mask} = \widetilde{\mathbf{Z}}^R \times (\widetilde{\mathbf{V}}_L)^T \tag{9}$$

where $\mathbf{Mask} \in \mathbb{R}^{N\times H_L\times W_L}$ is N segmentation masks, and each mask is an $H_L \times W_L$ segmentation map.

For OV classification, We generate a text embedding matrix $\mathbf{Cls} \in \mathbb{R}^{C\times D}$ for all the C candidate classes on the target application, where the text embedding for each class can be generated by our LLM. If there is no pre-defined candidate class, we can use all LLM object priors **R** as candidates to realize OV segmentation. Then, we calculate mask-text similarities:

$$\mathbf{Sim} = \widetilde{\mathbf{Z}}^R \times \mathbf{Cls}^T \tag{10}$$

where $\mathbf{Sim} \in \mathbb{R}^{N\times C}$ denotes similarities between each mask and each candidate class. For each mask, the class with the highest similarity can be selected as the class of this mask.

Finally, we convert mask-wise predictions into pixel-level classification maps by

$$\widetilde{\mathbf{Out}} = \mathbf{Sim}^T \times \mathbf{Mask} \tag{11}$$

where $\widetilde{\mathbf{Out}} \in \mathbb{R}^{C \times H_L \times W_L}$ is a pixel-wise semantic segmentation map. The final segmentation output **Out** is generated by selecting the class with the highest score for each pixel.

3.5 Training

Similar to fixed-set semantic segmentation (Cheng et al., 2022), our entire training loss contains two parts as follows:

$$Loss = L_{mask} + \lambda L_{cls} \tag{12}$$

where L_{mask} is the mask loss for mask predictions **Mask**, including a binary cross-entropy loss and a dice loss. L_{cls} is a cross-entropy loss for classification results based on mask-class similarities **Sim**. λ is a weight to control the loss ratio. We only train the adapter, MSDA and transformer decoder, while fixing other parts, as shown in Fig. 2. During training, LLM descriptions for each image only need to be extracted once to reduce computational redundancies.

4 Experiments

4.1 Datasets and Metrics

Following previous OV semantic segmentation methods (Liang et al., 2023; Xu et al., 2022, 2023), we train our model on the COCO-Stuff dataset (Caesar et al., 2018), while testing on ADE20K (Zhou et al., 2017), Pascal Context (Mottaghi et al., 2014) and Pascal VOC (Everingham et al., 2010) to evaluate the OV performance.

COCO-Stuff (Caesar et al., 2018) is a comprehensive segmentation dataset that includes mask annotations for 171 classes. These classes encompass both things (such as *dogs* and *cats*) and stuffs (e.g., *grass* and *sky*). The dataset comprises more than 118,000 training images and 5,000 validation images.

Pascal Context (Mottaghi et al., 2014) contains 5105 natural images for validation. There are two class settings in this dataset: the 59 most frequent classes (PC-59) and all 459 classes (PC-459). P-459 is harder and can better estimate the OV ability.

ADE20K (Zhou et al., 2017) comprises a set of 2000 validation photographs. We utilize two variations: one including the top 150 most common classes (A-150) and the other encompassing a broader range of 847 classes (A-847).

The Pascal VOC 2012 , as described in Everingham et al. (2010), consists of 20 classes. Most of the classes overlap with COCO-Stuff. Therefore, OV methods on this dataset usually achieve high accuracy. The dataset consists of 11,185 training images and 1,449 validation images.

Evaluation metrics. We use the common semantic segmentation metric, mean Intersection over Union (mIoU), to evaluate the performance. All reported mIoU scores are in a percentage format.

4.2 Implementation Details

Our method can use any VQA-based MLLM as the backbone. Here, we use LLAVA$-$1.5-7B Liu et al. (2023) as an example, which includes an image encoder (ViT-L trained on CLIP Radford et al. (2021)) and a language model (Vicuna-7B Vicuna (2023)). Textual embeddings are extracted from the final embedding layer. The maximum numbers M and K of object and relation priors are both set to 50. The number N of mask queries is 100. We use four image feature maps, i.e., $L = 4$. We set $\lambda = 2.0$ in our loss function. The input image size is 640×640 and the number of iterations is 120K. Other network and training settings are the same as Cheng et al. (2022). The model is trained on 8 Nvidia V100 GPUs.

4.3 Comparisons with State-of-the-art Methods

We report OV semantic segmentation results on the Pascal Context dataset in Table 1. FC-CLIP (Yu et al., 2023) shows the best results in previous methods, which proposes a network with higher resolutions. Compared with FC-CLIP (Yu et al., 2023), our method achieves gains of 7.2% on PC-459 and 5.8% on PC-59. SAN (Xu et al., 2023) is the second-best existing method. It leverages a side adapter to adapt CLIP knowledge for OV semantic segmentation. Our method outperforms it by 8.3% and 4.0% on PC-459 and PC-59, respectively. We achieve this superior performance because various LLM knowledge are exploited by our method, and our semantic, scaled visual as well as relation attentions effectively leverage these knowledge to guide OV semantic segmentation.

Table 2 shows the results on the ADE20K dataset. Our proposed method also achieves state-of-the-art performance on both A-847 and A-150. In particular, our method exceeds FC-CLIP (Yu et al., 2023), the second-best method, by 1.7% on A-847 and 4.4% on A-150. When comparing with SAN (Xu et al., 2023) using the VIT-L backbone, we yield improvements of 2.8% on A-847 and 5.2% on A-150.

In Table 3, we evaluate the effectiveness of our LLM-Former on Pascal VOC 2012. Since this dataset only contains 20 classes and most of them are included in COCO-Stuff training data, all methods show high performance. Despite that, we also achieve the best mIoU, and significantly outperform the previous SOTA method FC-CLIP by 1.4%. We also report results under the Open IoU metric (Zhou et al., 2023)

Table 1 OV semantic segmentation results on pascal context

Method	Backbone	Training Dataset	PC-459 (mIoU (%))	PC-59 (mIoU (%))
LSeg+ Li et al. (2022)	R101	COCO-Panoptic	5.2	36.0
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic	6.5	36.9
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic + Loc. Narr	7.9	40.1
GroupVIT Xu et al. (2022)	VIT-S	GCC + YFCC	4.9	25.9
Zegformer Ding et al. (2022)	R101	COCO-Stuff	10.4	45.5
Simple Xu et al. (2022)	R101c	COCO-Stuff	8.7	47.7
DeOP Han et al. (2023)	R101c	COCO-Stuff 156	9.4	48.8
MaskCLIP Ding et al. (2023)	R50	COCO-Panoptic	10.0	45.9
OV-Seg Liang et al. (2023)	R101c	COCO-Stuff + COCO-Caption	11.0	53.3
OV-Seg Liang et al. (2023)	Swin-B	COCO-Stuff + COCO Caption	12.4	55.7
HIPIE Wang et al. (2023)	ViT-H	O365,COCO,RefCOCO,PACO	14.4	59.3
ODISE Xu et al. (2023)	ViT-B	COCO-Panoptic	14.5	57.3
SAN Xu et al. (2023)	VIT-B	COCO-Stuff	12.7	53.4
SAN Xu et al. (2023)	VIT-L	COCO-Stuff	17.1	60.2
FC-CLIP Yu et al. (2023)	ConvNeXt-L	COCO-Panoptic	18.2	58.4
LLMFormer (Ours)	ViT-L	COCO-Stuff	**25.4**	**64.2**

Bold values indicate the best results
'PC-459' and 'PC-59' mean 459 and 59 classes on pascal context, respectively

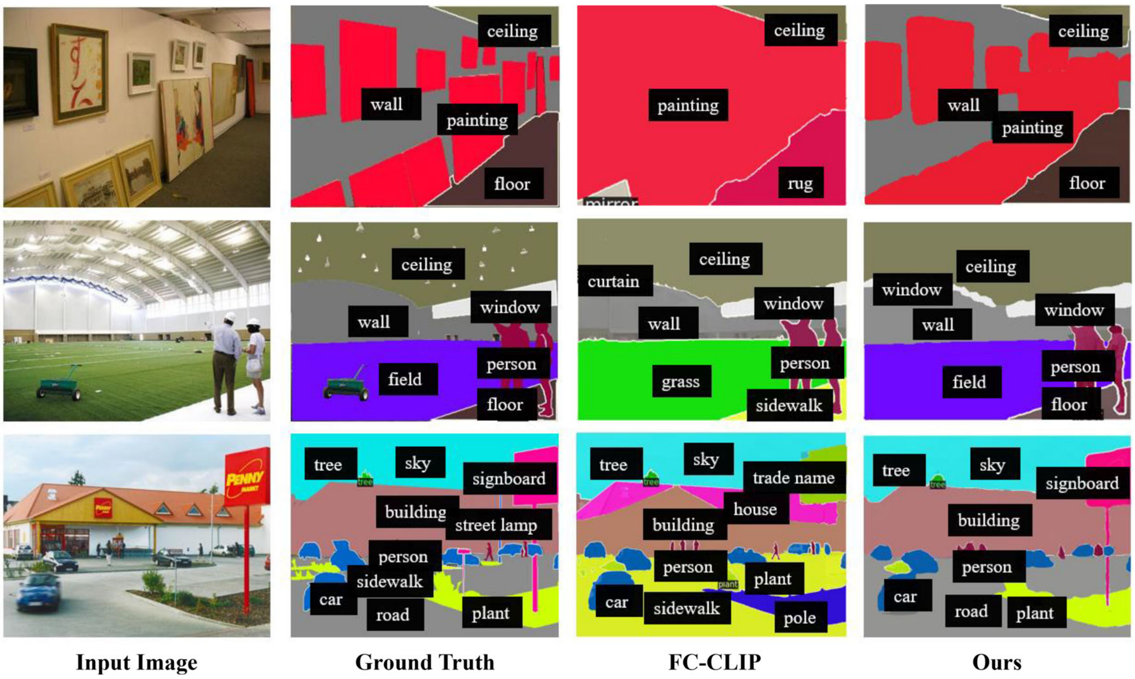


Fig. 6 OV semantic segmentation results on A-150. Left to right: input images, ground truths, results of FC-CLIP Yu et al. (2023) and ours. Previous methods fail to segment and classify some objects, such as *wall*, *floor* and *field* in the first and second images. Prior works also over-segment some large objects (e.g., *plant* and *building* in the third image), and miss small objects like the pole of the *signboard* object in the third image. Our method avoids these errors, because we extract object, attribute and relation priors from LLM, and leverage them for OV object segmentation, classification as well as scale selection

Table 2 OV semantic segmentation results on ADE20K

Method	Backbone	Training dataset	A-847 (mIoU (%))	A-150 (mIoU (%))
LSeg+ Li et al. (2022)	R101	COCO-Panoptic	2.5	13.0
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic	4.0	15.3
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic + Loc. Narr	4.4	17.5
GroupVIT Xu et al. (2022)	VIT-S	GCC + YFCC	4.3	10.6
Zegformer Ding et al. (2022)	R101	COCO-Stuff	5.6	18.0
Simple Xu et al. (2022)	R101c	COCO-Stuff	7.0	20.5
DeOP Han et al. (2023)	R101c	COCO-Stuff 156	7.1	22.9
MaskCLIP Ding et al. (2023)	R50	COCO-Panoptic	8.2	23.7
OV-Seg Liang et al. (2023)	R101c	COCO-Stuff + COCO-Caption	7.1	24.8
OV-Seg Liang et al. (2023)	Swin-B	COCO-Stuff + COCO Caption	9.0	29.6
HIPIE Wang et al. (2023)	ViT-H	O365,COCO,RefCOCO,PACO	9.7	29.0
ODISE Xu et al. (2023)	ViT-B	COCO-Panoptic	11.1	29.9
SAN Xu et al. (2023)	VIT-B	COCO-Stuff	10.2	27.6
SAN Xu et al. (2023)	VIT-L	COCO-Stuff	13.7	33.3
FC-CLIP Yu et al. (2023)	ConvNeXt-L	COCO-Panoptic	14.8	34.1
LLMFormer (Ours)	ViT-L	COCO-Stuff	**16.5**	**38.5**

Bold values indicate the highest results
'A-847' and 'A-150' represent 847 and 150 classes on ADE20K, respectively

Table 3 OV semantic segmentation results on Pascal VOC 2012

Method	Backbone	Training dataset	mIoU (%)
LSeg+ Li et al. (2022)	R101	COCO-Panoptic	59.0
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic	60.0
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic + Loc. Narr	63.8
GroupVIT Xu et al. (2022)	VIT-S/16	GCC + YFCC	50.7
Zegformer Ding et al. (2022)	R101	COCO-Stuff	89.5
Simple Xu et al. (2022)	R101c	COCO-Stuff	88.4
DeOP Han et al. (2023)	R101c	COCO-Stuff 156	91.7
OV-Seg Liang et al. (2023)	R101c	COCO-Stuff + COCO-Caption	92.6
OV-Seg Liang et al. (2023)	Swin-B	COCO-Stuff + COCO Caption	94.5
SAN Xu et al. (2023)	CLIP VIT-B/16	COCO-Stuff	94.0
SAN Xu et al. (2023)	CLIP VIT-L/14	COCO-Stuff	95.5
FC-CLIP Yu et al. (2023)	ConvNeXt-L	COCO-Panoptic	95.4
LLMFormer (Ours)	ViT-L	COCO-Stuff	**96.8**

Bold value indicates the best results

Table 4 OV semantic segmentation results under the Open IoU metric Zhou et al. (2023), which can better verify the OV capability

Method	Backbone	Training dataset	A-847	A-150	PC-459	PC-59	Pascal VOC
Simple Xu et al. (2022)	R101c	COCO-Stuff	12.9	29.0	14.5	50.8	90.2
OV-Seg Liang et al. (2023)	Swin-B	COCO-Stuff + COCO Caption	15.7	36.9	17.4	59.6	95.6
SAN Xu et al. (2023)	CLIP VIT-L/14	COCO-Stuff	19.2	39.0	19.9	60.4	96.0
FC-CLIP Yu et al. (2023)	ConvNeXt-L	COCO-Panoptic	20.6	40.4	16.3	61.2	96.2
LLMFormer (Ours)	ViT-L	COCO-Stuff	**23.2**	**44.8**	**28.5**	**68.2**	**97.3**

Bold values indicate the best results

Table 5 The effects of main components in our method on the ADE20K dataset

Model	Semantic attention	Scaled visual attention	Relation attention	A-847	A-150
Baseline				7.3	23.4
Model A	✓			11.9	30.5
Model B	✓	✓		14.8	36.2
Model C	✓		✓	13.3	34.0
Full model	✓	✓	✓	16.5	38.5

The baseline model is a modification of the fixed-set segmentation method Mask2Former (Cheng et al., 2022), where we replace the fixed-set classification head with a mask-text alignment head for OV recognition. The text embeddings are extracted by our language model. We also replace the image encoder with our encoder (i.e., ViT, adapter and MSDA)

in Table 4, which can better evaluate the OV ability. It can be observed that our method achieves significant improvements, compared with previous works. These results demonstrate the effectiveness of our LLMFormer and attention modules.

We visualize our results in Fig. 6. It can be seen that FC-CLIP misclassifies some classes due to their similarity. For instance, *rug* and *grass* in the first and second images are misclassified as *floor* and *field*, respectively. In addition, FC-CLIP also does not segment out the *wall* object in the first image. Our method exploits LLM object, attribute and relation priors for OV semantic segmentation, and thus reduces these mistakes. In addition, in the third image in Fig. 6, FC-CLIP fails to segment small objects such as the pole of the *signboard* object, despite high-resolution inputs. Meanwhile, the *plant* object in this image is over-segmented by FC-CLIP. Our method successfully segments out these objects, because we leverage LLM attribute priors for segmentation scale selection.

4.4 Discussion

In this section, we conduct extensive ablation studies to further verify the effectiveness of our proposed methods. All models are trained with COCO-Stuff while being tested on ADE20K.

The effects of main components. The effects of every proposed component are reported in Table 5. We modify several components of the fixed-set segmentation method Mask2Former (Cheng et al., 2022) as our baseline. We place the fixed-set classification head with a mask-text alignment head for OV recognition. The text embeddings are extracted by our language model. Its image encoder is also replaced with our encoder for a fair comparison. When comparing Model A with the baseline, we observe significant performance gains of 4.6% and 7.1% on A-847 and A-150, respectively. These findings highlight the significance of our semantic attention module, which embeds LLM object and attribute priors into the segmentation model. Moreover, both Model B and Model C show improvements compared to

Table 6 The effects of our *semantic attention* on the ADE20K dataset

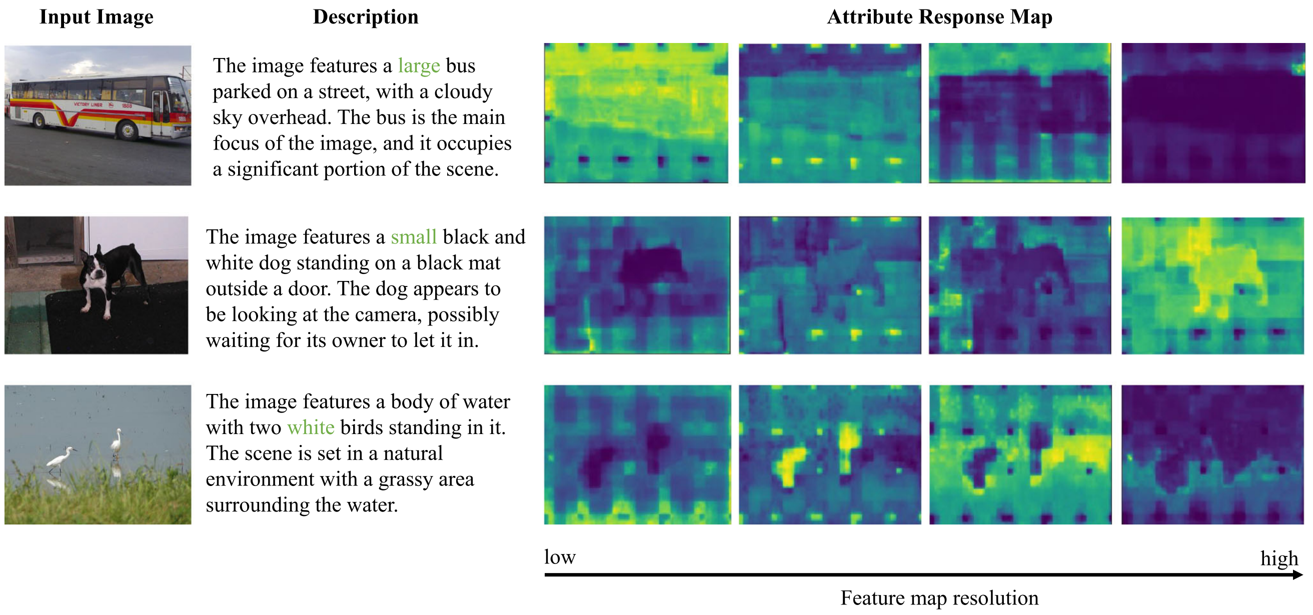
Model	Object priors	Attribute priors	A-847	A-150
Baseline			7.3	23.4
Embedding enhance				
Model D	✓		8.8	26.3
Model E	✓	✓	9.4	26.9
Embedding enhance and self-attention				
Model F			8.7	26.5
Model G	✓		9.6	27.8
Model H	✓	✓	10.4	28.3
Cross-attention				
Model I	✓		11.0	29.2
Model A	✓	✓	11.9	30.5

Scaled visual and relation attention are not used in this experiment. ‘Embedding enhance’ means that we do not use the MHCA in our semantic attention. Instead, we directly sum all object and attribute embeddings as a single vector, and add the sum to each mask embedding for enhancement. ‘Embedding enhance and self-attention’ means that we sum the priors and add them to mask embeddings, and then input enhanced mask embeddings into MHSA. ‘Cross-attention’ means that we leverage cross-attention to enhance mask embeddings based on our priors, as described in Sect. 3.4

Table 7 The effects of scaled visual attention on ADE20K

Model	Scale strategies	A-847	A-150
Model A	Multi-scale progress	11.9	30.5
Model J	Single scale (large)	10.5	27.9
Model K	Multi-scale fusion	11.7	30.8
Model L	Scale selection	13.5	35.1
Model B	Scale selection based on attributes	14.8	36.2

Here, we use semantic attention. ‘Multi-scale progress’ means that multi-scale feature maps are progressively input into different blocks in the transformer decoder as in Cheng et al. (2022). In Model J, we only use the large feature map to capture more visual details. In Model K, all image feature maps are input into the MHCA and we average the output embeddings. Model L leverages a vision-based scale selection method (Shi et al., 2023). ‘Model B’ is our method that selects segmentation scales based on LLM attribute priors



The image features a large bus parked on a street, with a cloudy sky overhead. The bus is the main focus of the image, and it occupies a significant portion of the scene.

The image features a small black and white dog standing on a black mat outside a door. The dog appears to be looking at the camera, possibly waiting for its owner to let it in.

The image features a body of water with two white birds standing in it. The scene is set in a natural environment with a grassy area surrounding the water.

Fig. 7 Visualization of scale selection based on attributes. We show how a certain attribute (the green word in the description) selects multi-scale feature maps. To this end, we first set other attributes to 0 in our attribute priors **A** and predict scale selection scores **S**. Then, **S** in the final decoder block is multiplied with the mask predictions **Mask** to generate attribute response maps. In the first image, the *large* bus area is highlighted in low-resolution feature maps to avoid over-segmentation. High-resolution feature maps are selected in the second image to better segment the *small* dog. Other attributes (such as *white* in the third image) are also helpful to highlight objects, but only in middle-resolution feature maps

Table 8 The effects of relation attention on ADE20K

Model	Relation strategies	A-847	A-150
Model B	Transformer attention only	14.8	36.2
Model M	Transform	15.1	36.6
Full model	Sum and transform	16.5	38.5

Here,we use semantic and scaled visual attention. Model B uses original attention maps in the MHSA. Model M only uses the linear layer with softmax. Our full model adds our relation priors to original attention maps and leverages a linear layer with softmax to transform the sum

Model A, suggesting that scale selection based on LLM attributes and attention enhancement by LLM relations can further improve the performance. Therefore, we incorporate all three modules into our full model, yielding the best performance of 16.5% and 38.5% on A-847 and A-150, respectively.

Semantic attention. Table 6 shows different setups of our semantic attention. We explore the effectiveness of object priors, attribute priors, and various methods to incorporate these priors. Overall, the utilization of object priors consistently improves the OV segmentation performance. For example, all models involving object priors, such as Models D, G & I, outperform models without object priors, such as the baseline and Model F. Attribute priors are employed in Models E, H & A to further enhance the results. Among different methods of prior integration, cross-attention exhibits the most superior performance. Meanwhile, cross-attention allows us to calculate prior-mask correspondences, which is crucial for our following scaled visual and relation attentions.

Table 9 The effects of different attention methods on ADE20K

Model	A-847	A-150
Masked attention Cheng et al. (2022) (Baseline)	7.3	23.4
TSG attention Shi et al. (2023)	9.4	30.6
Ours	16.5	38.5

Masked attention (Cheng et al., 2022) adds additional masks to cross-attention. It is equal to our baseline model, where we change the classification head in Mask2Former (Cheng et al., 2022) into the OV classification head, and replace the image encoder with our encoder. TSG Attention (Shi et al., 2023) generates scale gates to improve attentions, where we add this module to our baseline

Table 10 The number of trainable parameters and FLOPs

Model	Params (M)	FLOPs (T)	A-847	A-150
Zegformer Ding et al. (2022)	430	10.5	8.0	23.5
Simple Xu et al. (2022)	64	9.6	9.9	28.3
OV-Seg Liang et al. (2023)	321	9.6	12.5	33.1
SAN Xu et al. (2023)	11	0.9	13.7	33.3
Ours (LLAVA−1.5-7B)	129	11.8	16.5	38.5
Ours (LLAVA-Phi2-2.7B)	129	6.3	15.2	36.7

We reproduce Zegformer Ding et al. (2022), Simple Xu et al. (2022), OV-Seg Liang et al. (2023) and SAN Xu et al. (2023) with CLIP ViT-L/14. The size of input images is 640x640

Scaled visual attention. We compare different scale strategies in Table 7. In general, multi-scale methods are superior to the single-scale approach. Furthermore, scale selection approaches such as Models L & B outperform other non-selection methods. Our proposed scale selection by attribute priors outperforms Model L by 1.3% and 1.1% on A-847 and A-150, respectively. These results demonstrate the effectiveness of our scaled visual attention based on LLM attribute priors.

We visualize the relationships between object attributes and multi-scale feature maps in Fig. 7. It can be observed that scale-aware attributes are helpful for scale selection. For example, in the first image, low-resolution feature maps are selected for the *large* bus to reduce over-segmentation. In contrast, in the second image, the *small* dog is highlighted in high-resolution feature maps to generate finer masks. Other attributes (color, texture, etc.) can also highlight objects, such as *white* birds in the third image. Nevertheless, such attributes mainly highlight objects in middle-resolution feature maps, while scale-aware attributes are able to select very high-/low-resolution ones.

Relation attention. Table 8 reports the effects of our relation priors. Compared with Model B, our Full Model achieves increases of 1.7% and 2.3% on A-847 and A-150. In our Full Model, we incorporate LLM relation priors into self-attention maps, and use a linear layer to integrate them. Only using this linear layer (Model M) yields slight improvements compared to Model B. These results demonstrate that the gains come from our relation priors.

Different attention methods. We show the results of different attentions in Table 9. Our attention methods significantly outperform masked attention (Cheng et al., 2022) and TSG Attention (Shi et al., 2023), because they can not recognize OV objects, while our attention modules exploit LLM priors to improve OV semantic segmentation.

Cost comparison. Table 10 reports the computational costs of state-of-the-art methods and our model. Compared with Zegformer (Ding et al., 2022) and OV-Seg (Liang et al., 2023), our model includes fewer trainable parameters while achieving higher accuracy, because we only have trainable parameters in the adapter, MSDA and decoder. We can also use LLAVA-Phi2-2.7B (Zhu et al., 2024) to further reduce the computational overhead, while only slightly decreasing the performance.

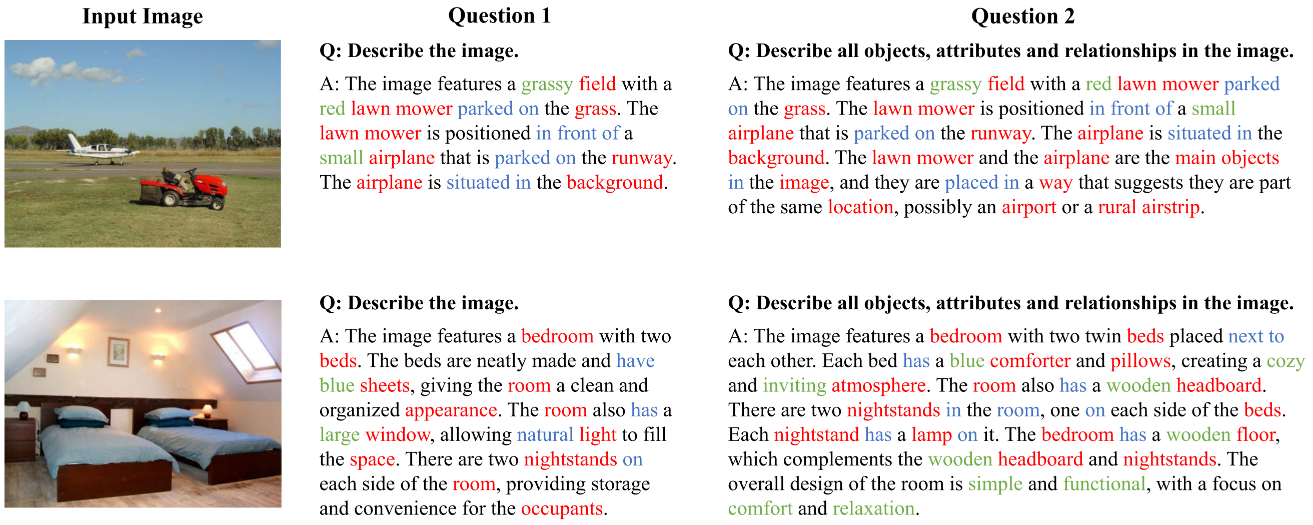
Table 11 The effects of different questions on ADE20K. ‘Question 1’ only use simple prompts, while ‘Question 2’ includes more detailed prompts

Model	A-847	A-150
Question 1: describe the image	16.5	38.5
Question 2: describe all objects, attributes and relationships in the image	17.8	39.9

Different prompts. In Table 11, we compare questions with simple prompts ‘describe the image’ as well as detailed prompts ‘describe all objects, attributes and relationships in the image’. Detailed prompts are able to further increase the segmentation performance. Figure 8 shows some examples of the generated descriptions. With detailed prompts, more objects, attributes and relationships are generated, such as *lamp* and *wooden* in the second image. Since we focus on visual predictions, we mainly use simple prompts in other experiments as a running example.

Cross-domain zero-shot semantic segmentation. To further evaluate the generalization ability, we conduct cross-domain zero-shot semantic segmentation experiments. We train all models on COCO-Stuff while testing on the Pascal Context dataset. During training, we remove annotations whose classes belong to 59 Pascal Context classes. The results are shown in Table 12. It can be observed that our method outperforms SAN (Xu et al., 2023) and FC-CLIP (Yu et al., 2023) by 5.5% and 9.8%, respectively. We depict qualitative results in Fig. 9. FC-CLIP (Yu et al., 2023) fails to segment some objects, such as *ground* in the first image, and *ceiling*, *wall* as well as *floor* in the third image. Our method avoids such errors. These results further demonstrate the generalization ability of our method.

Failure cases. Figure 10 shows failure cases of our method. Firstly, some complicated object boundaries are not well segmented, such as *tree* in the first image and *lamp* in the second image. Enlarging input image sizes can alleviate these



Q: Describe the image.

A: The image features a grassy field with a red lawn mower parked on the grass. The lawn mower is positioned in front of a small airplane that is parked on the runway. The airplane is situated in the background.

Q: Describe all objects, attributes and relationships in the image.

A: The image features a grassy field with a red lawn mower parked on the grass. The lawn mower is positioned in front of a small airplane that is parked on the runway. The airplane is situated in the background. The lawn mower and the airplane are the main objects in the image, and they are placed in a way that suggests they are part of the same location, possibly an airport or a rural airstrip.

Q: Describe the image.

A: The image features a bedroom with two beds. The beds are neatly made and have blue sheets, giving the room a clean and organized appearance. The room also has a large window, allowing natural light to fill the space. There are two nightstands on each side of the room, providing storage and convenience for the occupants.

Q: Describe all objects, attributes and relationships in the image.

A: The image features a bedroom with two twin beds placed next to each other. Each bed has a blue comforter and pillows, creating a cozy and inviting atmosphere. The room also has a wooden headboard. There are two nightstands in the room, one on each side of the beds. Each nightstand has a lamp on it. The bedroom has a wooden floor, which complements the wooden headboard and nightstands. The overall design of the room is simple and functional, with a focus on comfort and relaxation.

Fig. 8 Image descriptions generated by different questions. 'Question 1' only contains simple prompts 'describe the image'. 'Question 2' uses more detailed prompts 'describe all objects, attributes and relationships in the image', and thus predicts more image information

Table 12 Cross-domain zero-shot semantic segmentation results on pascal context

Model	Backbone	Training data	P-59
SAN Xu et al. (2023)	VIT-L	COCO-Stuff w/o P-59	45.3
FC-CLIP Yu et al. (2023)	ConvNeXt-L	COCO-Stuff w/o P-59	41.0
Ours	ViT-L	COCO-Stuff w/o P-59	50.8

'COCO-Stuff w/o P-59' means that we remove 59 pascal context classes from the COCO-Stuff dataset during training

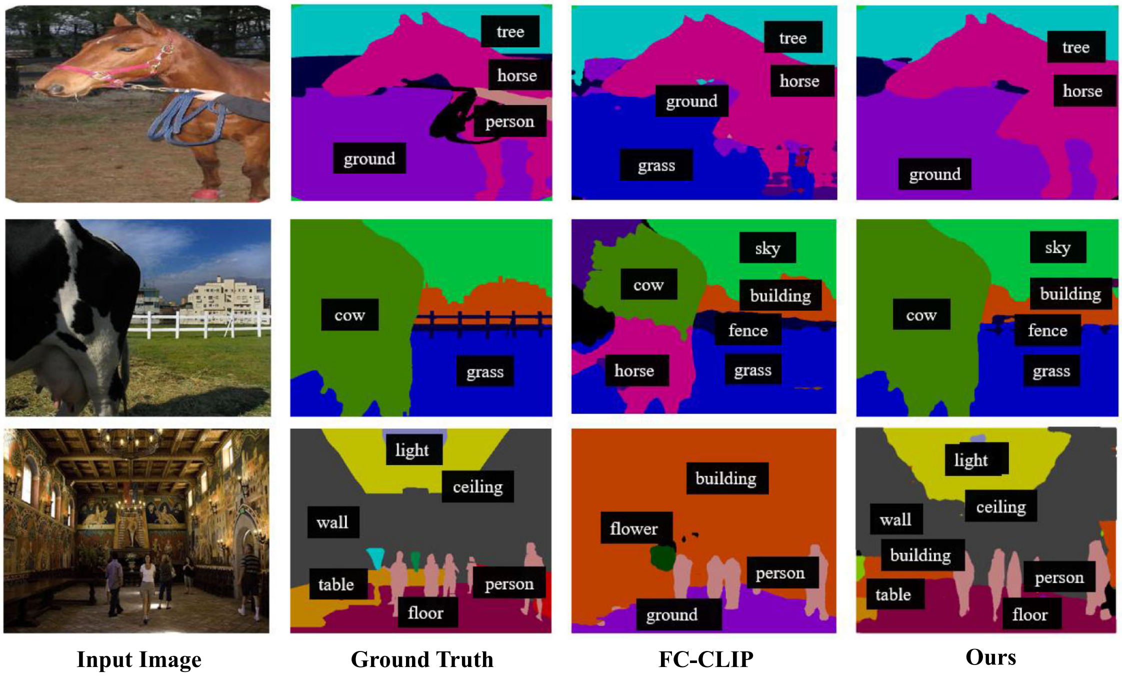


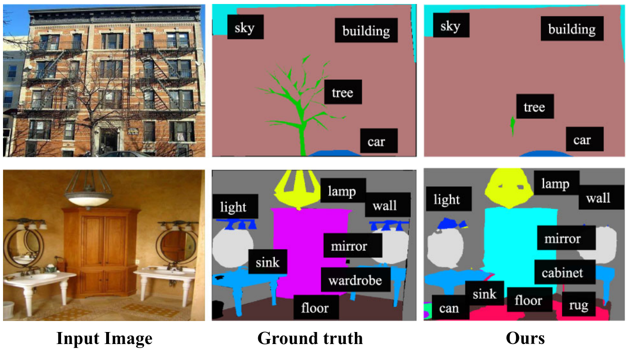
Fig. 9 Visualized cross-domain zero-shot semantic segmentation results on Pascal Context. Left to right: input images, ground truths, results of FC-CLIP Yu et al. (2023) and ours. FC-CLIP Yu et al. (2023) generates several mistakes. For example, *ceiling* and *wall* in the third image are predicted as *building*. *Cow* in the second image is over-segmented. Our method correctly segments these objects

Table 13 OV semantic segmentation results without pre-defined candidate classes on ADE20K

Model	Backbone	Training data		A-847	A-150
With candidate classes					
LSeg+ Li et al. (2022)	R101	COCO-Panoptic		2.5	13.0
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic		4.0	15.3
OpenSeg Ghiasi et al. (2022)	R101	COCO-Panoptic + Loc. Narr		4.4	17.5
GroupVIT Xu et al. (2022)	VIT-S	GCC + YFCC		4.3	10.6
Zegformer Ding et al. (2022)	R101	COCO-Stuff		5.6	18.0
Simple Xu et al. (2022)	R101c	COCO-Stuff		7.0	20.5
DeOP Han et al. (2023)	R101c	COCO-Stuff 156		7.1	22.9
LLMFormer (Ours)	ViT-L	COCO-Stuff		16.5	38.5
	Semantic	Scaled visual	Relation		
Model	attention	attention	attention	A-847	A-150
Without candidate classes					
Model A	✓			4.9	9.6
Model B	✓	✓		6.2	13.8
Full model	✓	✓	✓	**7.3**	**16.4**

Bold values indicate the highest results

errors. Secondly, our method confuses several very similar classes. For instance, the *wardrobe* object in the second image is misclassified as *cabinet*. We will develop fine-grained OV methods in the future to address this issue.



Input Image Ground truth Ours

Fig. 10 Failure cases on A-150. Left to right: input images, ground truths and our results. Our method does not well segment complicated object boundaries, such as the *tree* object in the first image and the *lamp* object in the second image. Fine-grained classes also fail to be distinguished. For example, in the second image, *wardrobe* is misclassified as *cabinet*

4.5 OV Semantic Segmentation without Pre-defined Candidate Classes

In this section, we evaluate our ability for OV semantic segmentation without pre-defined candidate classes. During inference, previous methods require a set of candidate classes. However, in real-world applications, candidate classes are usually not provided. Different from prior works, our method is able to generate OV semantic segmentation without pre-defined candidate classes, because LLMs provide object class priors.

Table 13 shows the results. Since previous works cannot generate such results, we compare different settings of our method. As introduced in Sect. 3.4, when there is no pre-defined class, we use object classes in LLM descriptions as candidate classes. Nevertheless, there is a problem during evaluation. Due to the language diversity, LLMs may generate various names for an object, while the target dataset only labels one or several names. For example, LLM predicts a *student* object, while the label on the target dataset is *person*. To solve this problem, we leverage CLIP textual similarity for evaluation. If the textual similarity between the predicted and ground truth object names is higher than a threshold Thr, we think this object is correctly predicted. We set Thr to 70 during evaluation. Our full model achieves 7.3% mIoU on A-857. Although this result is lower than ours with pre-defined classes, because LLMs miss some objects and the language diversity reduces evaluation accuracy, it still outperforms many previous methods based on pre-defined classes. These results show our ability to segment OV objects without pre-defined candidate classes. Meanwhile, the comparisons among Model A, Model B and Full Model further demonstrate the effectiveness of our proposed priors and attention modules.

5 Conclusion

In this paper, we have presented LLMFormer, a novel approach exploiting LLM knowledge for OV semantic segmentation. Three types of attention modules are proposed to leverage object, attribute and relation priors from LLMs for segmentation. Firstly, object and attribute priors are used to improve OV mask prediction and classification by semantic attention. Secondly, scaled visual attention is introduced to select suitable segmentation scales for each mask based on attribute priors. Thirdly, our relation attention enhances visual long-range dependency learning by LLMs relation priors. Extensive experiments demonstrate the effectiveness of our LLMFormer and each attention module. Moreover, our model can predict OV segmentation results without pre-defined candidate classes, which is more practical for real-world applications.

Although the current work explores LLM knowledge to boost OV semantic segmentation performance, the huge number of parameters in LLMs reduce the speed. Meanwhile, this work focuses on improving the OV recognition ability, while there are still fine-grained classification and boundary segmentation problems. Therefore, in the future, we will effort to study the efficiency, fine-grained classification and segmentation problems in the OV field.

Data availability The data supporting the findings of this study are based on public databases (https://github.com/nightrome/cocostuff Caesar et al. (2018), https://groups.csail.mit.edu/vision/datasets/ADE20K/ Zhou et al. (2017), https://cs.stanford.edu/~roozbeh/pascal-context/ Mottaghi et al. (2014) and http://host.robots.ox.ac.uk/pascal/VOC/ Everingham et al. (2010)) and are available from the corresponding author upon request.

References

Barsellotti, L., Amoroso, R., Baraldi, L., & Cucchiara, R. (2024). FOSSIL: free open-vocabulary semantic segmentation through synthetic references retrieval. In *IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1453–1462). IEEE

Bousselham, W., Thibault, G., Pagano, L., Machireddy, A., Gray, J., Chang, Y. H., & Song, X. (2022). Efficient self-ensemble for semantic segmentation. arXiv:2111.13280.

Caesar, H., Uijlings, J., & Ferrari, V. (2018). Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1209–1218).

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., & Qiao, Y. (2023). Vision transformer adapter for dense predictions. in *International Conference on Learning Representations*

Chen,L.-C., Yang,Y., Wang,J., Xu,W. & Yuille,A. L.(2016). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 3640–3649).

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision* (pp. 801–818).

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R.(2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1290–1299).

Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems, 34*, 17864–17875.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834–848.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv:2305.06500

Dao, S. D., Shi, H., Phung, D., & Cai, J. (2023). Class enhancement losses with pseudo labels for open-vocabulary semantic segmentation. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2023.3330102

Ding, H., Jiang, X., Shuai, B., Liu, A. Q., & Wang, G. (2018). Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2393–2402).

Ding, Z., Wang, J., & Tu, Z.(2023). Open-vocabulary panoptic segmentation with maskclip. arXiv:2208.08984

Ding, J., Xue, N., Xia, G.-S., & Dai, D. (2022). Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11583–11592).

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Fan, J., & Zhang, Z. (2023). Toward practical weakly supervised semantic segmentation via point-level supervision. *International Journal of Computer Vision, 131*(12), 3252–3271.

Ghiasi,G., Gu,X., Cui,Y., & Lin,T.-Y.(2022). Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision* (pp. 540–557). Springer.

Han, C., Zhong, Y., Li, D., Han, K., & Ma, L. (2023). Open-vocabulary semantic segmentation with decoupled one-pass network. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1086–1096).

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2020). Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(02), 386–397.

Hu, S., Zhao, X., & Huang, K. (2023). SOTVerse: A user-defined task space of single object tracking. *International Journal of Computer Vision, 132*(2), 872–930.

Jain, J., Li, J., Chiu, M. T., Hassani, A., Orlov, N., & Shi, H. (2023). Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2989–2998).

Jaus, A., Yang, K., & Stiefelhagen, R. (2023). Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning. *IEEE Transactions on Intelligent Transportation Systems, 24*(4), 4438–4453.

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., & Jia, J. (2023). Lisa: Reasoning segmentation via large language model. arXiv:2308.00692.

Li, Y., Qi, H., Dai, J., Ji, X., & Wei, Y. (2017). Fully convolutional instance-aware semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition.* (pp. 4438–4446).

Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., & Ranftl, R. (2022). Language-driven semantic segmentation. In *The International Conference on Learning Representations*

Li,X., Zhao,H., Han,L., & Tong,Y.(2020). Gated fully fusion for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11418–11425)

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., & Marculescu, D. (2023). Open-vocabulary semantic segmentation with mask-adapted clip. In *The IEEE / CVF Conference on Computer Vision and Pattern Recognition.* (pp. 7061–7070).

Lin, G., Milan, A., Shen, C., & Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1925–1934).

Lin, F., Hu, W., Wang, Y., Tian, Y., Lu, G., Chen, F., Xu, Y., & Wang, X. (2023). Universal object detection with large vision model. *International Journal of Computer Vision, 132*(4), 1258–1276.

Lin, G., Shen, C., Van Den Hengel, A., & Reid, I. (2018). Exploring context with deep structured models for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(6), 1352–1366.

Liu, H., Li, C., Li, Y., & Lee, Y. J.(2023). Improved baselines with visual instruction tuning. In*NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*

Liu, H., Li, C., Wu, Q., & Lee, Y. J.(2023). Visual instruction tuning. arXiv:2304.08485.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. arvix:1506.04579

Li, X., Zhang, J., Yang, Y., Cheng, G., Yang, K., Tong, Y., & Tao, D. (2023). Sfnet: Faster and accurate semantic segmentation via semantic flow. *International Journal of Computer Vision, 132*(2), 466–489.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

Ma, C., Yang, Y., Ju, C., Zhang, F., Zhang, Y., & Wang, Y. (2023). Open-vocabulary semantic segmentation via attribute decomposition-aggregation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*

Ma, J., Liu, J., Chai, Q., Wang, P., & Tao, J. (2023). Diagram perception networks for textbook question answering via joint optimization. *International Journal of Computer Vision, 132*, 1578–1591.

Mottaghi, R., Chen, X., Liu,X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., & Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 891–898).

Noh, H., Hong, S., & Han, B.(2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1520–1528).

OpenAI, (2023). Gpt-4 technical report.

Peng ,Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., & Wei, F. (2023). Kosmos-2: Grounding multimodal large language models to the world. arXiv:2306.14824.

Qi, L., Kuen, J., Wang, Y., Gu, J., Zhao, H., Torr, P., Lin, Z., & Jia, J. (2022). Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(7), 8743–8756.

Qin, Z., Liu, J., Zhang, X., Tian, M., Zhou, A., Yi, S.,& Li, H. (2022). Pyramid fusion transformer for semantic segmentation. arXiv:2201.04019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P.,& Clark, J., et al., (2021). Learning transferable visual models from natural language supervision. arXiv:2103.00020.

Ranftl, R., Bochkovskiy, A., & Koltun,V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 12179–12188).

Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., & Manning, C. D. (2015). Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language* (pp. 70–80). Citeseer

Shi, H., Hayat, M., & Cai, J.(2023). Open-vocabulary object detection via scene graph discovery. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 4012–4021).

Shi, H., Hayat, M., & Cai, J.(2023). Transformer scale gate for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3051–3060).

Shi,H., Hayat,M., & Cai,J.(2024) . Unified open-vocabulary dense visual prediction. *IEEE Transactions on Multimedia*

Shi, H., Hayat, M., Wu, Y., & Cai, J. (2022). Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9611–9620).

Shi, H., Li, H., Wu, Q. & Song, Z.(2019). Scene parsing via integrated classification model and variance-based regularization. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5307-5316).

Shi, H., Li, H., Wu, Q., Meng, F., & Ngan, K. N. (2018). Boosting scene parsing performance via reliable scale prediction. In *2018 ACM Multimedia Conference on Multimedia Conference ACM* (pp. 492–500).

Shi, H., Li, H., Meng, F., Wu, Q., Xu, L., & Ngan, K. N. (2018). Hierarchical parsing net: Semantic scene parsing from global scene to objects. *IEEE Transactions on Multimedia, 20*(10), 2670–2682.

Strudel, R., Garcia, R., Laptev, I., & Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7262–7272).

Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14453–14463).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Roziere, B., Goyal, N., Hambro, E.,& Azhar, F., et al., (2023). Llama: Open and efficient foundation language models. arXiv:2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*arXiv:1706.03762.

Vicuna, (2023) Vicuna: An open-source chatbot impressing gpt-4 with 90 quality. [Online]. Available: https://vicuna.lmsys.org/,

Wang, X., Li, S., Kallidromitis, K., Kato, Y., Kozuka, K., & Darrell, T.(2023). Hierarchical open-vocabulary universal image segmentation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 568–578).

Wang, W., Wang, R., Shan, S., & Chen, X. (2023). Importance first: Generating scene graph of human interest. *International Journal of Computer Vision, 131*(10), 2489–2515.

Wu, Y.-H., Liu, Y., Zhan, X., Cheng, M.-M.(2021). P2t: Pyramid pooling transformer for scene understanding. arXiv:2106.12011

Wysoczanska, M., Ramamonjisoa, M., Trzcinski, T., & Siméoni, O. (2024). CLIP-DIY: CLIP dense inference yields open-vocabulary semantic segmentation for-free. In *IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1392–1402). IEEE

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems, 34*, 12077–12090.

Xu,J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., & Wang, X. (2022). Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18134–18144).

Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., & Xie,W. (2023). Learning open-vocabulary semantic segmentation models from natural language supervision. Un *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2935–2944).

Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., & De Mello, S. (2023). Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., & Bai, X. (2021). A simple baseline for zero-shot semantic segmentation with pretrained vision-language model. arXiv:2112.14757 .

Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., & Bai, X. (2022). A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX.* (pp. 736–753). Springer

Xu, M., Zhang, Z., Wei, F., Hu, H., & Bai, X. (2023). SAN: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(12), 15546–15561.

Yang,J., Li,C., Zhang,P., Dai,X., Xiao,B., Yuan,L., Gao,J.(2021). Focal self-attention for local-global interactions in vision transformers. arXiv:2107.00641

Yang, J., Zhang, H., Li, F., Zou, X., Li, C., & Gao, J. (2023). Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv:2310.11441.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122v3

Yu, Q., He, J., Deng, X., Shen, X., & Chen, L.-C. (2023). Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 32215–32234). Curran Associates.

Zhang, H., Li, F., Zou, X., Liu, S., Li, C., Yang, J., & Zhang, L. (2023). A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1020–1031).

Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K. & Luo, P. (2023). Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv:2307.03601.

Zhang, H.,Dana, K., Shi, J., Zhang, Z., Wang, X., Tyagi, A., & Agrawal, A. (2018). Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7151-7160).

Zhang, D., Lin, Y., Tang, J., & Cheng, K. T. (2023). CAE-GReaT: Convolutional-auxiliary efficient graph reasoning transformer for dense image predictions. *International Journal of Computer Vision, 132*, 1502–1520.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2881–2890).

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al., (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6881–6890).

Zhou, H., Shen, T., Yang, X., Huang, H., Li, X., Qi, L., & Yang, M.-H. (2023). Rethinking evaluation metrics of open-vocabulary segmentaion. arXiv:2311.03352.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592.

Zhu, Y. , Zhu, M., Liu, N., Ou, Z., Mou, X., & Tang, J. (2024). Llava-phi: Efficient multi-modal assistant with small language model. arXiv:2401.02330.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.