Efficient Segmentation with Multimodal Large Language Model via Token Routing

Changsong Wen, Zelin Peng, Yu Huang, Wei Shen[†]

MoE Key Lab of Artificial Intelligence, AI Institute, School of Computer Science, Shanghai Jiao Tong University
{changsong, zelin.peng, yellowfish, wei.shen}@sjtu.edu.cn

Abstract

Recent advances in multimodal large language models (MLLMs) have demonstrated strong capabilities in addressing open-world segmentation tasks. However, the substantial computational cost of the LLM components presents a significant challenge, especially in segmentation tasks, where efficiency has long been a central concern. Existing efficient MLLM approaches typically reduce computation cost by pruning visual tokens in the early layers, as they account for the majority of the input sequence. Despite their efficiency, this is incompatible with dense prediction tasks such as segmentation, since removing visual tokens leads to the loss of essential object parts and spatial details. To better understand the roles of visual tokens in segmentation, we analyze the attention weights of both image and mask tokens within LLM. We find that image tokens are important throughout all layers, whereas mask tokens only attend to image tokens at deeper layers. Based on the observation, we build an efficient segmentation framework based on MLLMs by introducing a sophisticated token routing strategy. This strategy dynamically determines when and how different tokens participate in computation: For mask tokens, they are only inserted at deeper layers of the LLM to reduce redundant computation, since they rarely attend to image tokens in early layers; For image tokens, only a small number of them, named proxies, are updated via full feedforward network (FFN) computation, while the update of the remaining tokens is guided by these proxies, i.e., efficiently computed through a lightweight projector applied on the difference of the proxies during their update. Our method achieves a 1.5× acceleration over the original LLM process by reducing its FLOPs to 56%, while maintaining the same segmentation performance.

Code — https://github.com/downdric/Token-Routing

Introduction

Driven by the significant progress of LLMs (Achiam et al. 2023; Touvron et al. 2023a,b; Li et al. 2023b), recent research has increasingly focused on extending LLMs to handle visual modalities, enabling more comprehensive understanding and reasoning across multimodal inputs. As a representative multimodal large language model (MLLM), LLaVA (Liu et al. 2023a, 2024) integrates a pretrained vision encoder with a powerful language model to support a wide range of vision-language tasks, such as detailed image description and image-grounded dialogue. However, MLLMs like LLaVA can only produce textual outputs, which makes it inherently challenging for such models to address pixel-level vision tasks such as segmentation.

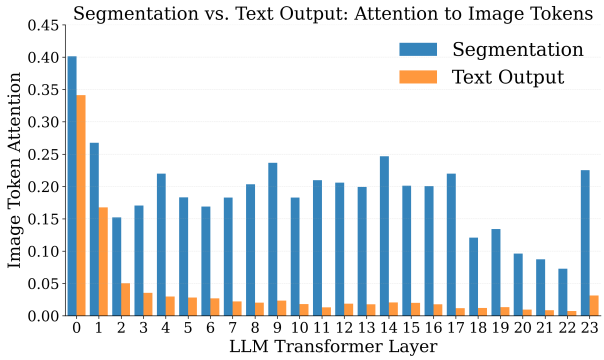

Figure 1: Visualization of attention weights. Image tokens receive little attention in the deeper layers of the LLM during text output tasks, whereas in segmentation tasks, they receive notable attention throughout all layers.

Recent studies (Xia et al. 2024; Wei et al. 2025; Zhu et al. 2025; Gong et al. 2025) investigate extending MLLMs to support segmentation tasks. One notable approach is LISA (Lai et al. 2024), which introduces a special token into the output of LLaVA to represent the target region in an auto-regressive manner. This token is subsequently decoded by SAM (Kirillov et al. 2023) to segment image regions according to the textual description. More recently, PSALM (Zhang et al. 2024e) introduces a unified framework that can address more generalized segmentation tasks, such as referring segmentation, panoptic (generic) segmentation, and interactive segmentation, achieving superior performance with a more compact model. In this framework, image tokens, instruction prompts[1], and mask tokens[2] are jointly processed by an LLM backbone

[†] Corresponding author.

[1] Instruction prompts provide high-level textual guidance tailored to different segmentation tasks.
[2] Mask tokens attend to image tokens to encode visual information, and can be regarded as segmentation-specific visual tokens.

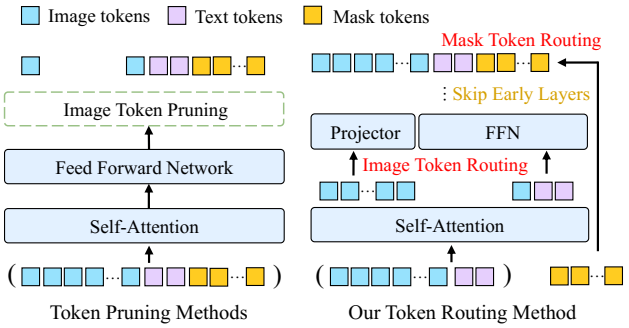

Figure 2: Illustration of token pruning methods used in text output tasks and our proposed token routing method, including image and mask tokens. Our token routing strategy significantly improves computational efficiency while preserving important visual information for segmentation.

and a Mask2Former decoder (Cheng et al. 2022), allowing it to generate segmentation masks in a single forward pass.

Although these methods have achieved promising results, the substantial parameters and computational cost of LLMs pose a significant challenge. Current efforts in efficient MLLMs (Chen et al. 2024a; Xing et al. 2025; Zhang et al. 2024c; Shang et al. 2024) primarily focus on text output tasks. These approaches reduce computational cost through visual token reduction (e.g., pruning), which shortens the context length and thereby improves computational efficiency. This design is primarily motivated by two empirical observations: the number of visual tokens typically surpasses that of textual tokens, and these visual tokens receive significantly less attention in the deeper layers (Chen et al. 2024a). While such strategies perform well in tasks like image captioning that rely on global semantics, they are suboptimal for dense prediction tasks such as segmentation. This can be attributed to the fact that each visual token corresponds to a specific region of the image, and pruning them results in the loss of object parts or spatial details, which are crucial for generating complete and accurate segmentation masks, particularly in the multi-object scenes.

To illustrate the necessity of preserving visual tokens in dense prediction tasks, we analyze the layer-wise image attention weights in the MLLM-based segmentation model to understand how image tokens participate in the prediction process. Specifically, we visualize the attention weights assigned to image tokens by PSALM (Zhang et al. 2024e) on the RefCOCO dataset for segmentation, and compare results with those observed in text output tasks (Chen et al. 2024a). The results are shown in Fig. 1. Unlike text output tasks, where image tokens receive attention primarily in the early layers (Chen et al. 2024a; Zhang et al. 2025a), image tokens in segmentation consistently receive notable attention across all layers. This reveals the essential role of image tokens in segmentation tasks and indicates that reducing image tokens leads to the loss of critical information.

To reduce the computational burden of the LLM in segmentation tasks, we propose a token routing strategy. This strategy is designed to route both mask and image tokens, with the overall framework illustrated in Fig. 2. For the mask tokens, they begin to attend to image tokens primarily in the deeper layers of the LLM, suggesting that the segmentation model progressively updates them at these stages to encode visual information relevant to the target regions (detailed quantitative results are presented in the analysis section below). Mask tokens are largely underutilized in the early layers and are routed to skip them, reducing redundant computation. For the image tokens, they have already acquired hierarchical visual representations from the vision encoder and thus retain discriminative semantic information, passing them through computation-intensive feed-forward networks (FFNs) results in substantial computational costs and unnecessary processing. Inspired by this, proxy-guided image token routing updates only a small subset of image tokens using the original FFN as proxies. The differences before and after their update are then used to guide the adaptation of the remaining tokens through a lightweight projector, enabling efficient layer-wise feature propagation throughout the network.

We evaluate our model on a wide range of public datasets from various segmentation tasks, including referring segmentation, panoptic segmentation (covering semantic and instance segmentation), and interactive segmentation. Compared to the LLM baseline, our method requires only 56% of the original computational cost and improves inference speed by 1.5×, without sacrificing performance. These results demonstrate the potential of our approach as an efficient and versatile model for general segmentation tasks.

Related Work

Multimodal Large Language Model

The advances of large language models (LLMs), such as GPT-series (Achiam et al. 2023; Brown et al. 2020), Qwen (Yang et al. 2024), and LLaMA (Touvron et al. 2023b) provide a strong foundation for general-purpose language understanding. To extend these capabilities to multi-modal tasks, multimodal large language models (MLLMs) (Li et al. 2023a; Alayrac et al. 2022; Liu et al. 2023a, 2024) incorporate a vision encoder (Radford et al. 2021; Zhai et al. 2023) to perform joint reasoning over visual and textual inputs. In representative MLLMs (Liu et al. 2023a, 2024; Chen et al. 2024c), visual features are projected into the language embedding space through a lightweight projection layer, enabling effective alignment between visual and textual modalities. This unified architecture supports a wide range of vision-language tasks (Alayrac et al. 2022; Liu et al. 2023a; Shao et al. 2024; Zhao et al. 2024; Zhang et al. 2025b,c).

However, these models typically involve a large number of parameters in the LLM component and result in substantial computational costs, making them resource-intensive for training and deployment. Recent studies investigate more efficient approaches, including smaller model architectures (Gunasekar et al. 2023; Li et al. 2023b; Zhang et al. 2024a), efficient training strategies (Zhang et al. 2025a; Xing et al. 2025), and low-cost inference techniques (Chen et al. 2024a) to alleviate the computational burden while maintaining strong multi-modal capabilities. FastV (Chen et al. 2024a) identifies

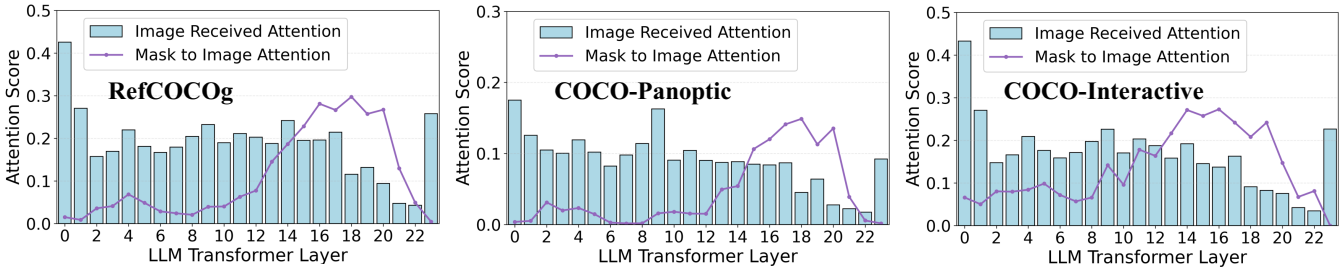

Figure 3: Layer-wise attention analysis. We visualize the average attention received by image tokens (blue bars, including self-attention) and the attention from the mask to image tokens (purple lines).

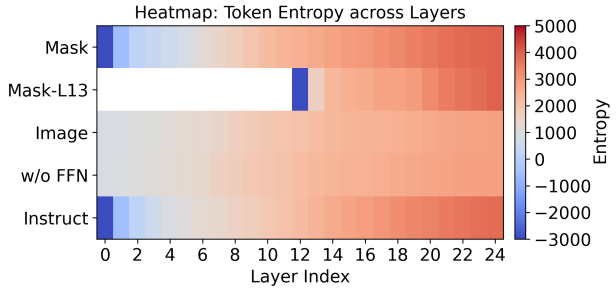

Figure 4: Layer-wise analysis of entropy variation of different tokens, where Mask-L13 indicates that mask tokens are inserted starting from layer 13 of the LLM, and w/o FFN denotes that FFN layers are skipped for image tokens.

the redundancy of using a large amount of visual tokens in deeper layers of MLLMs, and proposes a plug-and-play solution that adaptively prunes visual tokens to improve computational efficiency. PyramidDrop (Xing et al. 2025) addresses token redundancy by progressively dropping image tokens across model stages based on token similarity, improving efficiency with minimal performance degradation. LLaVA-Mini (Zhang et al. 2025a) and Libra-Merging (Yang et al. 2025) adopt token-merging strategies to compress visual information into fewer tokens, reducing context length and overall computational cost. While these efficient MLLMs demonstrate impressive performance on concept-level vision-language tasks, they are primarily optimized for text output tasks. Thus, they are not inherently designed for dense prediction tasks such as image segmentation.

Segmentation with MLLMs

Image segmentation is a fundamental pixel-level understanding task that assigns a semantic label to each pixel and serves as a critical component in a wide range of applications requiring detailed scene understanding (Long, Shelhamer, and Darrell 2015; Zhang et al. 2024d; Peng et al. 2025). Some recent works (He et al. 2024; Li et al. 2024; Rasheed et al. 2024; Shindo et al. 2024) extend the reasoning abilities of LLM to image segmentation. LISA (Lai et al. 2024) first propose to leverage LLMs for reasoning segmentation, enabling the interpretation of complex and implicit query texts. The image and text tokens are jointly processed by the LLM (e.g., LLaVA-7B), which outputs a special [SEG] token that is subsequently utilized by the SAM decoder to predict the segmentation mask. GSVA (Xia et al. 2024), GLaMM (Rasheed et al. 2024), and PixelLM (Ren et al. 2024) further extend this line of work by incorporating multi-object segmentation and the ability to handle queries with no target objects. SegLLM (**?**) performs multi-round interactive segmentation by integrating conversational memory into a mask-aware multimodal LLM, enabling context-aware object segmentation in a chat-like manner. Different from previous works that primarily focus on specific segmentation tasks, PSALM (Zhang et al. 2024e) and HyperSeg (Wei et al. 2025) adopt a unified architecture based on Mask2Former and incorporate large language models, enabling them to support a wider range of segmentation tasks.

In practical applications of image segmentation, model efficiency is of paramount importance. Therefore, researchers develop various efficient segmentation frameworks that reduce computational complexity while maintaining acceptable performance in both CNN-based (Yu et al. 2018, 2021; Fan et al. 2021; Zhao et al. 2018) and transformer-based architectures (Zhang, Cai, and Han 2024; Xie et al. 2021; Xiong et al. 2024). Recently, segmentation models built upon LLMs demonstrate remarkable reasoning capabilities. However, these models typically involve high computational cost and large model sizes. Motivated by this, our work aims to enhance the efficiency of LLM-based segmentation models, achieving competitive performance with lower computational requirements.

Observation and Analysis

Since the mask and image tokens account for a substantial portion of the input (exceeding 80% on RefCOCO), they contribute significantly to the overall computational cost. We provide analysis to support the design of our mask and image token routing strategies.

Analysis of mask token participation. Mask tokens are input as content-free embeddings and progressively learn to locate segmentation regions by attending to image features in the LLM. To investigate how mask tokens integrate visual cues, we compute the average attention that mask tokens assign to image tokens across different lay-

ers. Let $\mathbf{A}^{(l)} \in \mathbb{R}^{N\times N}$ denote the attention matrix at layer l, where N is the total number of tokens. The set of image token indices is denoted by $\mathcal{I} \subset \{0, 1, \ldots, N_{\text{img}} - 1\}$, and $\mathcal{M} \subset \{0, 1, \ldots, N_{\text{mask}} - 1\}$ denote the indices of mask tokens, the mask-to-image token attention can be formulated as:

$$\text{Attn}^{(l)}_{\text{mask2img}} = \frac{1}{|\mathcal{M}|} \sum_{q\in\mathcal{M}} \frac{\sum_{k\in\mathcal{I}} \mathbf{A}^{(l)}_{qk}}{\sum_{k=1}^{N} \mathbf{A}^{(l)}_{qk}} \tag{1}$$

Experiments are conducted on three representative segmentation datasets: RefCOCOg (Kazemzadeh et al. 2014; Nagaraja, Morariu, and Davis 2016), COCO-Panoptic (Lin et al. 2014), and COCO-Interactive (Zhang et al. 2024e). The results are illustrated by the purple lines in Fig. 3. We observe that across all three datasets, mask tokens begin to attend to image tokens primarily in the deeper layers, indicating that the model defers its decision on which regions to segment until the deeper stages. This motivates us to delay the participation of mask tokens until deeper stages of LLM for efficiency.

Analysis of image token participation. We compute the average attention weights assigned to image tokens to analyze their relative importance across layers. The average image attention received at layer l is calculated as:

$$\text{Attn}^{(l)}_{\text{img}} = \frac{1}{N} \sum_{q=1}^{N} \frac{\sum_{k\in\mathcal{I}} \mathbf{A}^{(l)}_{qk}}{\sum_{k=1}^{N} \mathbf{A}^{(l)}_{qk}} \tag{2}$$

As shown by the bars in Fig. 3, image tokens consistently receive attention and actively participate in the computation across all datasets, highlighting their critical role in segmentation throughout the entire LLM. Our following experiments show that pruning image tokens results in a drop in performance. Therefore, we turn to explore image token routing with the goal of simplifying their transformations within the LLM.

Image tokens are generated by a pretrained vision encoder, which already captures hierarchical semantics before they enter the LLM. Accordingly, applying intensive FFN transformations to image tokens provides limited benefit. To empirically validate this, we follow (Lin et al. 2024) and compute the entropy of different tokens at each layer to quantify the information content of each token type throughout the network. Specifically, given hidden states $\mathbf{z}_v^{(l)} \in \mathbb{R}^{N_v\times D}$, we compute the token entropy at layer l as:

$$\mathcal{H}^{(l)} = \sum_{d=1}^{D} \log\left(\sigma_d^{(l)} + \epsilon\right), \tag{3}$$

where $\sigma_d^{(l)}$ denotes the standard deviation of the d-th channel across tokens, and ϵ is a small constant to avoid numerical instability. As illustrated in Fig. 4, image tokens begin with a relatively high entropy, reflecting rich semantic content extracted by the vision encoder. Moreover, their entropy increases more slowly than that of other token types across layers, indicating that their representations remain stable and reducing the need for intensive transformation. When image tokens are allowed to skip FFN layers, their entropy remains relatively consistent, and subsequent ablation studies show no substantial drop in segmentation performance. This suggests that most of the information gain for image tokens arises from attention-based interactions, rather than FFN-based transformations. This motivates our proxy-guided update strategy, which applies full FFN computation only to a small set of representative image tokens, significantly reducing computational cost without sacrificing performance.

Methodology

Unified Segmentation with MLLMs

MLLM-based unified segmentation models typically consist of four components: a visual encoder f_{vis}, a modality alignment projector f_{proj}, a LLM backbone f_{LLM}, and a segmentation decoder f_{seg}. Given an input image $\mathbf{I}$, the visual encoder f_{vis} first extracts visual features. These features are then projected into the LLM feature space via a light-weight modality alignment projector:

$$\mathbf{v} = f_{\text{vis}}(\mathbf{I}), \mathbf{v}_a = f_{\text{proj}}(\mathbf{v}). \tag{4}$$

The LLM backbone f_{LLM} takes the aligned visual features $\mathbf{v}_a$, the instruction prompt $\mathbf{p}$, and mask tokens $\mathbf{m}$ as input to facilitate unified multi-modal comprehension for segmentation:

$$\mathbf{z}_v, \mathbf{z}_p, \mathbf{z}_m = f_{\text{LLM}}(\mathbf{v}_a, \mathbf{p}, \mathbf{m}), \tag{5}$$

where $\mathbf{z}_v, \mathbf{z}_p, \mathbf{z}_m$ are the LLM's output hidden states corresponding to the visual features, instruction prompts, and mask tokens, respectively. Finally, the segmentation decoder f_{seg} generates the segmentation map based on the image features and output of f_{LLM}:

$$\hat{\mathbf{M}}, \hat{\mathbf{C}} = f_{\text{seg}}(\mathbf{v}, \mathbf{z}_p, \mathbf{z}_m), \tag{6}$$

where $\hat{\mathbf{M}}$ and $\hat{\mathbf{C}}$ are the predicted mask and category. To enable the model to handle various segmentation tasks, we follow the architecture of Mask2Former (Cheng et al. 2022) and PSALM (Zhang et al. 2024e), incorporating unified instruction prompts to guide different segmentation objectives. The instruction prompt consists of two components: the task-specific prompt and the conditional prompt. The task-specific prompt provides high-level guidance tailored to different segmentation tasks. For example, a general instruction such as *"You need to segment all objects. These are all the candidate categories."* is used for panoptic segmentation. The conditional prompt further guides the model with detailed information such as class names, descriptive text, and visual cues. These task-specific and conditional components are jointly encoded and fed into the LLM to enable a unified and flexible segmentation. In the training process, we compute the mask loss using a combination of pixel-level binary cross-entropy and dice loss, and apply an autoregressive cross-entropy loss for vision-language instruction-following tasks.

Late-stage Mask Token Routing

Based on the analysis above, we propose a token routing strategy that dynamically regulates the participation of mask and image tokens in the segmentation process. The overall pipeline of our method is illustrated in Fig. 5.

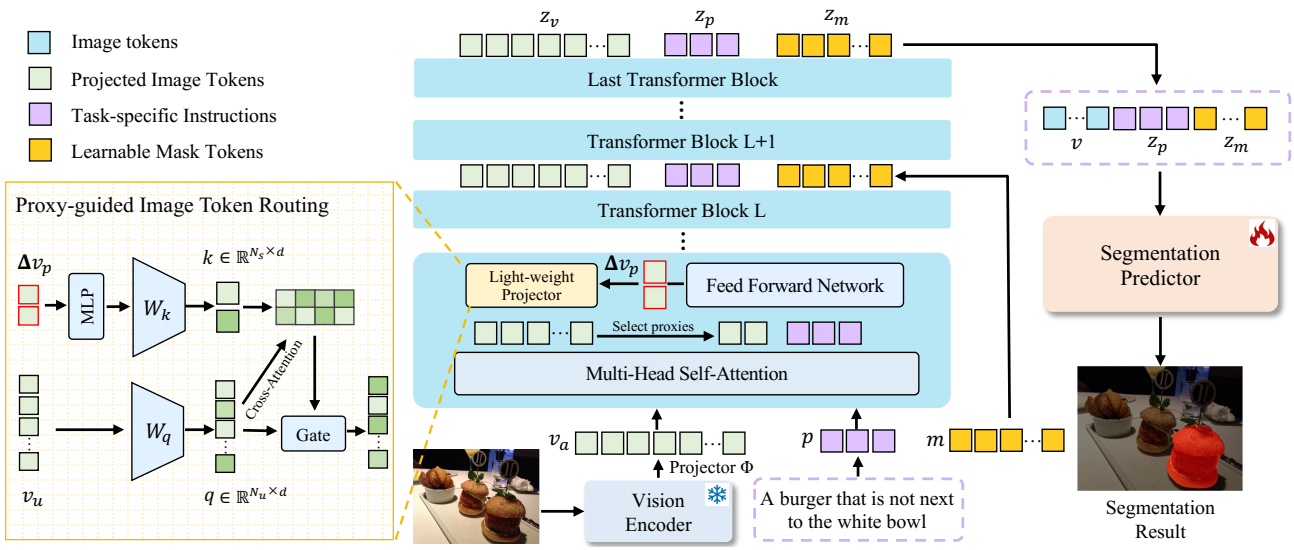

Figure 5: Architecture of the proposed efficient MLLM-based segmentation pipeline with late-stage mask token routing and proxy-guided image token routing.

Motivated by the observation that mask tokens aggregate visual information primarily in deeper layers, we design a mask token routing strategy that explicitly delays the injection of mask tokens into the model. Specifically, given the visual features $\mathbf{v}_a$ and instruction prompts $\mathbf{p}$, LLM f_{LLM} only process these two inputs in the early layers:

$$[\mathbf{z}_p^{(l)}, \mathbf{z}_v^{(l)}] = f_{\text{LLM}}^{(l)}([\mathbf{p}; \mathbf{v}_a]), \quad \text{for } l < L_{\text{mask}}, \tag{7}$$

where L_{mask} is the LLM layer at which the learnable mask tokens $\mathbf{m}$ start to get involved. Starting from this layer, the LLM processes the full token sequence consisting of instruction prompts, visual features, and mask tokens:

$$[\mathbf{z}_p^{(l+1)}, \mathbf{z}_v^{(l+1)}, \mathbf{z}_m^{(l+1)}] = f_{\text{LLM}}^{(l)}([\mathbf{z}_p^{(l)}; \mathbf{z}_v^{(l)}; \mathbf{m}]), \tag{8}$$

where $l = L_{\text{mask}}$. For the remaining layers $l > L_{\text{mask}}$, the LLM iteratively refines the hidden states of all token types by jointly processing the full token sequence.

Proxy-guided Image Token Routing

To improve computational efficiency, we propose a proxy-guided image token routing mechanism that enables effective updates of visual tokens. In LLMs, FFN typically follows a two-layer structure with an intermediate expansion from D to a higher dimension D_{e}, and then back to D, where D_{e} is commonly set to $4D$. Due to this dimensional expansion, FFNs are the most computationally intensive part of the model (Wei et al. 2024b), especially when applied to a large number of image tokens. To mitigate this, we apply the full FFN computation only to a small set of selected proxy tokens. The residual differences are subsequently propagated to the remaining image tokens through a lightweight, attention-guided mechanism, enabling efficient representation updating across layers. Let $\mathbf{v}_u \in \mathbb{R}^{N_u \times D}$ denote the image tokens not selected for full FFN computation, and $\mathbf{v}_p \in \mathbb{R}^{N_s \times D}$ represents the randomly selected proxies for each FFN layer, respectively. The representation shift is computed as:

$$\Delta\mathbf{v}_p = \phi(\mathbf{FFN}(\mathbf{v}_p) - \mathbf{v}_p), \tag{9}$$

To transfer the update to unselected tokens, we compute attention-based similarities between unselected tokens $\mathbf{v}_u$ (as queries) and selected proxy tokens $\mathbf{v}_p$ (as keys). Both are first projected into a lower-dimensional space via learnable matrices $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{D \times d_p}$, where $d_p \ll D$. Then, the attention weights between each unselected token and the selected proxies can be computed using scaled dot-product attention:

$$\alpha_{ij} = \frac{\exp(\langle q_i, k_j \rangle)}{\sum_{j'=1}^{N_s} \exp(\langle q_i, k_{j'} \rangle)}, \tag{10}$$

where α_{ij} denotes the attention weight between unselected token i and proxy token j. The aggregated update for each unselected token $\mathbf{x}_{u,i}$ is obtained by applying the attention weights to the residuals:

$$\Delta\mathbf{v}_{u,i} = \sum_{j=1}^{N_s} \alpha_{ij} \cdot \Delta\mathbf{v}_{p,j}. \tag{11}$$

To regulate the influence of the propagated update, we further introduce a learnable gate $\mathbf{g}_i$, with values computed by applying a sigmoid activation to the output of an MLP. The final updated representation is then given by:

$$\mathbf{v}'_{u,i} = \mathbf{v}_{u,i} + \sigma(\text{MLP}(\mathbf{v}_{u,i})) \odot \Delta\mathbf{v}_{u,i}, \tag{12}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, $\odot$ denotes element-wise multiplication. This proxy-guided update strategy significantly reduces the overall computation while preserving the expressive capacity of the model for image token propagation.

Method	LLM Type	RefCOCO			RefCOCO+			RefCOCOg		gRefCOCO		
		val	testA	testB	val	testA	testB	val	test	val	testA	testB
Segmentation Specialist												
SEEM-L (Zou et al. 2023)	-	-	-	-	-	-	-	65.6	-	-	-	-
CRIS (Wang et al. 2022)	-	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	55.3	63.8	51.0
LAVT (Yang et al. 2022)	-	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	57.6	65.3	55.0
PolyFormer-B (Liu et al. 2023b)	-	74.8	76.6	71.1	67.6	72.9	59.3	67.8	69.1	-	-	-
UNINEXT-L (Yan et al. 2023)	-	80.3	82.6	77.8	70.0	74.9	62.6	73.4	73.7	-	-	-
MLLM-based Segmentation												
LISA (Lai et al. 2024)	Vicuna-7B	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	38.7	52.6	44.8
GLaMM (Rasheed et al. 2024)	Vicuna-7B	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	-	-	-
GSVA (Xia et al. 2024)	Vicuna-7B	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	61.7	69.2	60.3
LaSagnA (Wei et al. 2024a)	Vicuna-7B	76.8	78.7	73.8	66.4	70.6	60.1	70.6	71.9	38.1	50.4	42.1
SAM4MLLM (Chen et al. 2024b)	Qwen-VL-7B	79.6	82.8	76.1	73.5	77.8	65.8	74.5	75.6	66.3	70.1	63.2
OMG-LLaVA (Zhang et al. 2024b)	InterLM2-7B	78.0	80.3	74.1	69.1	73.1	63.0	72.9	72.9	-	-	-
POPEN (Zhu et al. 2025)	Llama2-7B	79.3	82.0	74.1	73.1	77.0	65.1	75.4	75.6	-	-	-
PerceptionGPT (Pi et al. 2024)	Vicuna-13B	75.3	79.1	72.1	68.9	74.0	61.9	70.7	71.9	-	-	-
PixelLM (Ren et al. 2024)	Llama2-13B	77.7	79.9	74.2	68.0	71.5	61.5	73.2	73.9	-	-	-
PSALM (Zhang et al. 2024e)	Phi-1.5 (1.3B)	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	42.0	52.4	50.6
Efficient MLLM-based Methods												
FastV (Chen et al. 2024a)	Phi-1.5 (1.3B)	82.3	82.4	79.7	70.5	73.4	67.5	72.0	71.8	40.0	49.4	47.2
LLaVA-Mini (Zhang et al. 2025a)	Phi-1.5 (1.3B)	81.5	82.2	79.8	69.6	72.4	66.7	71.7	71.5	39.3	49.2	46.5
PyramidDrop (Xing et al. 2025)	Phi-1.5 (1.3B)	82.7	82.5	80.1	71.1	73.5	68.0	72.4	72.1	40.2	50.5	47.7
Token Routing (Ours)	Phi-1.5 (1.3B)	84.2	84.4	82.7	73.5	76.6	69.4	75.2	74.5	42.4	53.4	50.4
Δ *compared to PSALM baseline (56% FLOPs ↓)*		+0.6	-0.3	+1.1	+0.6	+1.1	-0.7	+1.4	+0.1	+0.4	+1.0	-0.2

Table 1: Comparison with the state-of-the-art methods on the referring segmentation datasets with cIoU. The gray numbers indicate results obtained using gRefCOCO for training, which is not utilized in our method.

Method	Backbone	PQ	mAP	mIoU
Mask2Former (Cheng et al. 2022)	Swin-B	55.1	45.2	65.1
OMG-Seg (Li et al. 2024)	ConvNeXt	55.4	-	-
PSALM (Zhang et al. 2024e)	Swin-B	55.9	45.7	66.6
PyramidDrop (Xing et al. 2025)	Swin-B	54.5	44.6	65.2
Token Routing (Ours)	Swin-B	55.7	45.7	66.6

Table 2: Comparison with the state-of-the-art methods on Panoptic COCO-val.

Method	Backbone	Point	Box	Scrib	Mask
SAM-L (Kirillov et al. 2023)	Swin-L	51.8	-	76.6	-
PSALM (Zhang et al. 2024e)	Swin-B	64.3	67.3	66.9	67.6
FastV (Chen et al. 2024a)	Swin-B	63.4	63.8	63.5	64.6
PyramidDrop (Xing et al. 2025)	Swin-B	63.6	64.5	64.3	65.0
Token Routing (Ours)	Swin-B	64.7	66.8	66.6	67.4

Table 3: Comparison with the state-of-the-art methods on COCO-Interactive.

Method	FLOPs (G) ↓	Latency (ms) ↓	cIoU ↑
PSALM (baseline)	1126.98	123	69.2
w/ Mask Routing	998.32	109	69.2
w/ Image Routing	758.75	97	69.3
FastV (prune 50%)	677.28	89	66.9
PyramidDrop	694.21	92	67.3
Ours	**630.09**	**83**	**69.7**

Table 4: Comparison of FLOPs, latency, and cIoU on the RefCOCO series. All methods are evaluated with a fixed number of 450 tokens (common in the RefCOCO dataset).

Experiment

Datasets

We train our model on three segmentation tasks: referring segmentation (RefCOCO, RefCOCO+, RefCOCOg (Nagaraja, Morariu, and Davis 2016; Yu et al. 2016)), generic segmentation (COCO Panoptic Segmentation (Lin et al. 2014)), and interactive segmentation (COCO-Interactive (Zhang et al. 2024e)). In addition, we also train the model using the LLAVA-1.5 dataset (Liu et al. 2024) to preserve its reasoning capabilities.

Implementation Details

Our architecture employs Swin-Base (Liu et al. 2021) for visual encoding and Phi-1.5 (1.3B) (Li et al. 2023b) as the LLM backbone due to its relatively small size, with the vision-language aligned projector initialized following PSALM (Zhang et al. 2024e). Our segmentation decoder is based on the Mask2Former (Cheng et al. 2022) and is initialized with its pretrained weights. We train our model for 56K iterations on 8 RTX 3090 GPUs with a batch size of 64 and an initial learning rate of 4e-5. We use AdamW as the optimizer and adopt a cosine learning rate decay schedule with a warm-up ratio of 0.03. The number of proxy tokens N_s is set to 8, and the dimension d_p of the query and key projection matrices $\mathbf{W}_q$ and $\mathbf{W}_k$ is 128. To reduce computational cost, the intermediate dimension d_g of the MLP in the gating module is set to 512, which is only 1/16 of that in the standard FFN. In addition, the mask tokens are inserted starting from layer $L_{mask} = 13$. During training, we freeze the

Ablations	RefCOCO	COCO-Pan	COCO-Point
A1. Mask Token Routing			
Baseline	81.56	55.19	61.26
Mask Routing(L=6)	81.55	55.20	61.31
Mask Routing(L=18)	80.48	55.12	60.54
Skip LLM	79.93	54.88	60.06
Ours (L=13)	**81.65**	**55.15**	**61.34**
A2. Image Token Routing			
Baseline	81.56	55.19	61.26
w/o FFN	80.40	54.85	60.37
w/o Gate	81.02	55.10	60.63
More Proxies (N_s=32)	81.65	55.14	61.53
Ours (Proxy-Guided)	**81.71**	**55.14**	**61.52**

Table 5: Ablation studies on token routing strategies. The model is trained by 14K iterations for practical efficiency.

vision encoder and update all other parameters in the model. For practical efficiency, we reduce the training iterations to 14K for ablation studies.

Main Results

We first compare our approach with SOTA methods on the referring expression segmentation task. The evaluation is conducted on four widely used benchmarks: RefCOCO, RefCOCO+, RefCOCOg, and gRefCOCO. Notably, our model is not trained on gRefCOCO, ensuring a fair comparison against PSALM. We categorize existing methods into three groups: (1) segmentation specialists that rely on dedicated segmentation architectures. (2) MLLM-based segmentation methods that incorporate MLLMs into segmentation tasks. (3) Efficient MLLM-based methods that aim to reduce computational cost by reducing visual tokens. To ensure a fair comparison, we reproduce these efficient MLLM-based methods using the same LLM backbone, following their original implementation protocols.

As shown in Table 1, our method achieves comparable performance to PSALM (69.7 vs. 69.2 on average) while significantly reducing computational cost. Under the average context length of the RefCOCO dataset, our approach achieves similar segmentation accuracy with approximately 56% FLOPs. In addition, our method also performs better than several larger models built on 7B or 13B LLMs. Token pruning-based MLLM methods, such as FastV and LLaVA-Mini, show a clear drop in performance on segmentation tasks, with average scores falling to 66.9 and 66.4 across the evaluated datasets.

We present the results of our method on both panoptic segmentation using the COCO-Panoptic dataset and interactive segmentation on the COCO-Interactive benchmark, as shown in Table 2 and Table 3, respectively. Compared to previous segmentation methods, our approach demonstrates promising performance across different benchmarks and achieves results comparable to PSALM, particularly in panoptic segmentation, which involves multiple object instances and requires fine-grained visual understanding. These results highlight the effectiveness of our method in improving computational efficiency while maintaining competitive segmentation performance.

Complexity Analysis

We conduct computational complexity analysis to quantify the efficiency of our method, as demonstrated in Table 4. We first compute the FLOPs by summing the major components of a standard transformer block: QKV projection ($6ND^2$), self-attention computation ($4N^2D$), attention output projection ($2ND^2$), and the feed-forward network ($16ND^2$). In this context, N denotes the sequence length of the input tokens, and D represents the hidden dimension of the LLM. The mask token routing skips the early L_{mask} layers for mask tokens, thereby saving FLOPs by reducing the sequence length to $N - N_{mask}$, where N_{mask} is usually set as 100. Second, the proxy-guided image token routing updates only a small number of image tokens via the full FFN, while routing the remaining tokens through a lightweight projector. This reduces the FFN FLOPs from $16N_{img}D^2$ to approximately $16N_sD^2 + 4(N_{img} - N_s)D(d_p + d_g)$, where $N_s = 8$ is the number of selected proxies and $d_p = 128$, $d_g = 512$ are the reduced dimension. We summarize the results in Table 4, illustrating the impact of our full method, its two key components, and token pruning methods on computational cost and average performance. Our method significantly reduces FLOPs in the LLM component to 56% and accelerates inference by 1.5 times without sacrificing performance.

Ablation Studies

We conduct experiments on three representative segmentation datasets: RefCOCO, COCO-Panoptic, and COCO-Interactive (Point), as shown in Table 5. We investigate the impact of injecting mask tokens at different layers of the LLM. The results indicate that skipping early LLM layers has a limited effect on performance, whereas delaying injection (e.g., L=18) or completely skipping the LLM results in a notable performance drop. For image tokens, we observe that skipping FFN layers has a limited impact on performance, resulting in only a minor drop of 0.79 points on average. This suggests that the image features do not require substantial updates through FFNs. We also evaluate the effect of the gating mechanism and find that removing it reduces the model's representational capacity. Moreover, using more proxies results in redundant computational overhead, with negligible improvement in performance.

Conclusion

We analyze how image and mask tokens participate in the computation within LLMs for segmentation tasks and propose a token routing strategy that inserts mask tokens only in deeper layers and uses proxy-guided updates for image tokens. This method effectively reduces computation by avoiding redundant updates while maintaining segmentation accuracy. Experiments show our approach achieves comparable performance with only 56% of the original LLM's FLOPs, enabling more efficient and practical deployment of MLLM-based segmentation models.

Acknowledgments

This work was supported by the NSFC under Grant 62322604 and 62576207.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35.

Chen, Y.-C.; Li, W.-H.; Sun, C.; Wang, Y.-C. F.; and Chen, C.-S. 2024b. SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. In *European Conference on Computer Vision*, 323–340. Springer.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; and Wei, X. 2021. Rethinking bisenet for real-time semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9716–9725.

Gong, S.; Zhuge, Y.; Zhang, L.; Yang, Z.; Zhang, P.; and Lu, H. 2025. The Devil is in Temporal Token: High Quality Video Reasoning Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Del Giorno, A.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

He, J.; Wang, Y.; Wang, L.; Lu, H.; He, J.-Y.; Lan, J.-P.; Luo, B.; and Xie, X. 2024. Multi-modal instruction tuned llms with fine-grained visual perception. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13980–13990.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742.

Li, X.; Yuan, H.; Li, W.; Ding, H.; Wu, S.; Zhang, W.; Li, Y.; Chen, K.; and Loy, C. C. 2024. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 27948–27959.

Li, Y.; Bubeck, S.; Eldan, R.; Del Giorno, A.; Gunasekar, S.; and Lee, Y. T. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Lin, S.; Lyu, P.; Liu, D.; Tang, T.; Liang, X.; Song, A.; and Chang, X. 2024. Mlp can be a good transformer learner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19489–19498.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R. K.; Mahadevan, V.; and Manmatha, R. 2023b. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18653–18663.

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

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision*, 792–807.

Peng, Z.; Xu, Z.; Zeng, Z.; Wen, C.; Huang, Y.; Yang, M.; Tang, F.; and Shen, W. 2025. Understanding Fine-tuning

CLIP for Open-vocabulary Semantic Segmentation in Hyperbolic Space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4562–4572.

Pi, R.; Yao, L.; Gao, J.; Zhang, J.; and Zhang, T. 2024. Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27124–27133.

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. In *International conference on machine learning*, 8748–8763.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26374–26383.

Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2024. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. *arXiv preprint arXiv:2403.15388*.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37: 8612–8642.

Shindo, H.; Brack, M.; Sudhakaran, G.; Dhami, D. S.; Schramowski, P.; and Kersting, K. 2024. Deisam: Segment anything with deictic prompting. *Advances in Neural Information Processing Systems*, 37: 52266–52295.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11686–11695.

Wei, C.; Tan, H.; Zhong, Y.; Yang, Y.; and Ma, L. 2024a. Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*.

Wei, C.; Zhong, Y.; Tan, H.; Liu, Y.; Zhao, Z.; Hu, J.; and Yang, Y. 2025. HyperSeg: Towards Universal Visual Segmentation with Large Language Model. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Wei, X.; Moalla, S.; Pascanu, R.; and Gulcehre, C. 2024b. Building on Efficient Foundations: Effectively Training LLMs with Structured Feedforward Layers. In *Advances in neural information processing systems*.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3858–3869.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xing, L.; Huang, Q.; Dong, X.; Lu, J.; Zhang, P.; Zang, Y.; Cao, Y.; He, C.; Wang, J.; Wu, F.; et al. 2025. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Xiong, Y.; Varadarajan, B.; Wu, L.; Xiang, X.; Xiao, F.; Zhu, C.; Dai, X.; Wang, D.; Sun, F.; Iandola, F.; et al. 2024. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16111–16121.

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15325–15336.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, L.; Shen, D.; Cai, C.; Chen, K.; Yang, F.; Gao, T.; Zhang, D.; and Li, X. 2025. Libra-Merging: Importance-redundancy and Pruning-merging Trade-off for Acceleration Plug-in in Large Vision-Language Model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 9402–9412.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18155–18165.

Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; and Sang, N. 2021. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International journal of computer vision*, 129: 3051–3068.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024a. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Zhang, S.; Fang, Q.; Yang, Z.; and Feng, Y. 2025a. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. In *International Conference on Learning Representations*.

Zhang, T.; Li, X.; Fei, H.; Yuan, H.; Wu, S.; Ji, S.; Loy, C. C.; and Yan, S. 2024b. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37: 71737–71767.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; et al. 2024c. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Zhang, Z.; Cai, H.; and Han, S. 2024. Efficientvit-sam: Accelerated segment anything model without performance loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7859–7863.

Zhang, Z.; Chen, S.; Wang, Z.; and Yang, J. 2024d. PlaneSeg: Building a Plug-In for Boosting Planar Region Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8): 11486–11500.

Zhang, Z.; Ma, Y.; Zhang, E.; and Bai, X. 2024e. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, 74–91.

Zhang, Z.; Wang, W.; Zhu, Y.; Qin, W.; Wan, P.; Zhang, D.; and Yang, J. 2025b. VidEmo: Affective-Tree Reasoning for Emotion-Centric Video Foundation Models. *arXiv preprint arXiv:2511.02712*.

Zhang, Z.; Xia, W.; Zhao, C.; Yan, Z.; Liu, X.; Zhu, Y.; Qin, W.; Wan, P.; Zhang, D.; and Yang, J. 2025c. Moda: Modular duplex attention for multimodal perception, cognition, and emotion understanding. *arXiv preprint arXiv:2507.04635*.

Zhao, C.; Wang, Y.; Jiang, X.; Shen, Y.; Song, K.; Li, D.; and Miao, D. 2024. Learning Domain Invariant Prompt for Vision-Language Models. *IEEE Transactions on Image Processing*, 33: 1348–1360.

Zhao, H.; Qi, X.; Shen, X.; Shi, J.; and Jia, J. 2018. Icnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)*, 405–420.

Zhu, L.; Chen, T.; Xu, Q.; Liu, X.; Ji, D.; Wu, H.; Soh, D. W.; and Liu, J. 2025. POPEN: Preference-Based Optimization and Ensemble for LVLM-Based Reasoning Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36: 19769–19782.