A Semantic-Guided Multimodal Image Segmentation Framework with Graph-Structured Reasoning

1st ziyang Tong
School of Automation
Wuhan University of Technology
Wuhan, China
Email: 348917@whut.edu.cn

2nd yixin Su
School of Automation
Wuhan University of Technology
Wuhan, China
Email: suyixin@whut.edu.cn

***Abstract*—Image segmentation driven by complex natural language instructions is a significant challenge in computer vision, with wide applications in scenarios such as street-scene analysis and autonomous driving. Existing methods often fall short in handling semantic ambiguity, modeling inter-object relationships, and achieving fine-grained segmentation. To address this, we propose a novel image segmentation framework that fuses multimodal semantic guidance with graph-structured reasoning. Our method first leverages the LLaMA 3-V multimodal large language model to jointly encode the image and language instruction, generating a task-relevant semantic guidance vector. Subsequently, a cascade of GroundingDINO and SAM generates semantically-aware candidate masks, whose visual features are extracted using DINOv2. The core innovation is a graph-based mask selection module that models candidates as nodes and leverages a graph convolutional network to perform collaborative reasoning based on spatial and semantic relationships, significantly improving the differentiation of similar objects. To reduce training costs, we employ a parameter-efficient modular fine-tuning strategy, inserting a QLoRA structure into LLaMA 3-V and a lightweight Adapter into the selection module, with tunable parameters accounting for less than 1% of the total. Experiments on street-scene datasets demonstrate that our method excels in complex instruction comprehension, fine-grained segmentation accuracy, and computational efficiency, offering a viable solution for practical applications.**

***Index Terms*—Multimodal learning, image segmentation, graph neural networks, large language models, parameter-efficient fine-tuning**

I. Introduction

Referring Expression Segmentation (RES) aims to segment a specific object or region in an image based on a guiding natural language description. With the increasing complexity of real-world applications like autonomous driving and robotic interaction, there is a growing demand for systems that can interpret nuanced, long-form instructions involving multiple attributes, spatial relationships, and actions. However, conventional methods often struggle with semantic ambiguity and fail to effectively model the contextual relationships between different object instances, particularly in cluttered scenes like street views.

Recent advances in large multimodal models (LMMs) and foundation models for vision, such as LLaMA 3-V [1] and the Segment Anything Model (SAM) [2], have opened new avenues for this task. Yet, directly applying these models is suboptimal. LMMs may not capture the fine-grained visual details required for precise segmentation, while SAM, being class-agnostic, requires explicit and accurate prompts to function effectively. Furthermore, the immense computational cost associated with fine-tuning these large models poses a significant barrier to their adaptation for specific downstream tasks [3].

To overcome these challenges, we propose a novel framework that synergizes the strengths of LMMs, open-set detectors, and graph-based reasoning. Our key contributions are threefold:

- **Multimodal Semantic Guidance:** We employ LLaMA 3-V to jointly process the image and a complex language instruction, generating a rich semantic vector that guides the entire segmentation process. This allows for a deeper understanding of user intent, including object attributes and spatial constraints.
- **Graph-Structured Reasoning for Mask Selection:** We introduce a dynamic graph-structured module to model the relationships between candidate masks generated by a GroundingDINO-SAM pipeline. By propagating information through a graph convolutional network [4], our model performs collaborative reasoning to disambiguate and select the most relevant masks, enhancing fine-grained accuracy.
- **Parameter-Efficient Fine-Tuning:** We design a modular fine-tuning strategy that freezes the majority of the pretrained model weights. By inserting lightweight, trainable components like QLoRA and Adapters into the language and selection modules, we drastically reduce the number of trainable parameters to less than 1% of the total, enabling efficient adaptation without sacrificing performance.

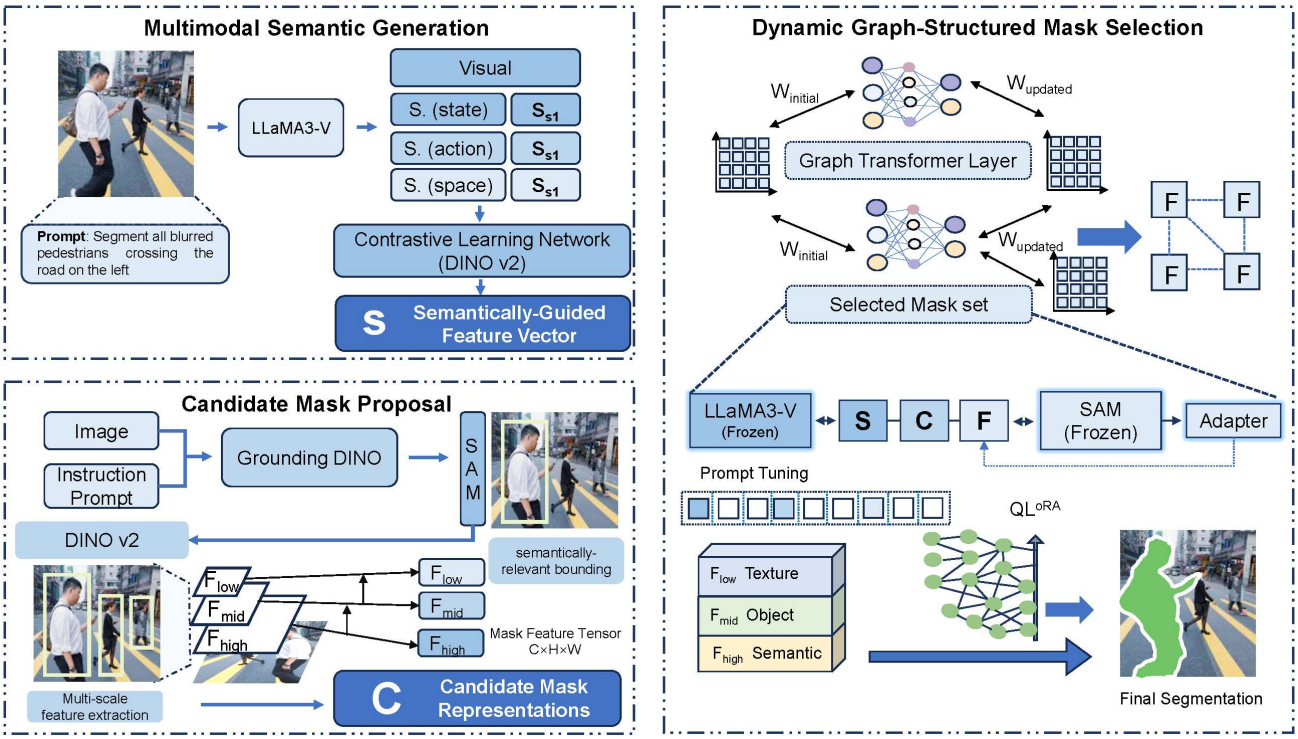

Fig. 1. The pipeline of the proposed framework comprises four core stages. Stage 1: Multimodal Semantic Generation. LLaMA-3V parses the input image and the natural language instruction to generate a structured semantic guidance vector. Stage 2: Candidate Mask Generation. Grounding DINO is employed to produce target bounding boxes, which are subsequently fed into SAM to generate candidate masks. Multi-scale visual features are then extracted from these masks using DINOv2. Stage 3: Graph-Structured Mask Selection. Candidate masks are constructed as nodes in a graph. A Graph Transformer performs reasoning by fusing spatial and semantic relationships among nodes, ultimately selecting the final masks under the guidance of the semantic vector. Stage 4: Parameter-Efficient Fine-Tuning. A modular fine-tuning strategy is adopted, where the backbone networks are frozen. Trainable QLoRA modules and a lightweight adapter are inserted to enable efficient model adaptation with minimal parameter updates.

II. PROPOSED METHOD

A. Multimodal Semantic Generation

This module distills high-level intent from an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and a language instruction T. We employ instruction decomposition, where LLaMA 3-V parses T into K semantic attributes $\{T_k\}_{k=1}^{K}$. For each attribute, the model outputs a semantic vector $\mathbf{v}_s^{(k)}$, forming a set $V_s = \{\mathbf{v}_s^{(k)}\}_{k=1}^{K}$. These vectors are projected into the mask feature space by a contrastive mapping network $\mathcal{M}_\theta$, yielding aligned guidance vectors $G_s = \{\mathbf{g}_s^{(k)}\}_{k=1}^{K}$ trained via a contrastive loss.

$$\mathbf{g}_s^{(k)} = \mathcal{M}_\theta(\mathbf{v}_s^{(k)}) \in \mathbb{R}^d \tag{1}$$

B. Candidate Mask Generation

This module generates a rich set of mask proposals in a three-stage cascade. First, GroundingDINO uses I and T for open-vocabulary detection, yielding bounding boxes $\{b_j\}$. Second, each b_j prompts SAM to produce high-quality masks $\{m_{ji}\}$, which are aggregated into a candidate pool $\mathcal{C} = \{m_1, \ldots, m_N\}$. Finally, a DINOv2 encoder extracts a multi-scale feature vector $\mathbf{h}_i \in \mathbb{R}^d$ for each mask m_i via masked average pooling. The primary goal of this module is to extract a task-specific semantic guidance vector from the input image and natural language instruction. We leverage the powerful joint reasoning capabilities of the LLaMA 3-V model. This vector encapsulates the high-level user intent, including object categories, attributes, and spatial relationships.

C. Dynamic Graph-Structured Mask Selection

This module is responsible for generating a set of high-quality, semantically-aware mask proposals. We employ a two-stage approach. First, the GroundingDINO model [5] takes the image I and instruction T as input to produce a set of bounding boxes corresponding to potential objects of interest. These bounding boxes serve as semantic prompts for the subsequent stage. Next, each predicted bounding box is fed as a prompt to SAM [2], which generates multiple high-quality, class-agnostic segmentation masks for the specified region. This cascade leverages the open-vocabulary detection power of GroundingDINO and the strong segmentation capabilities of SAM. For each resulting candidate mask m_i, we extract its corresponding visual feature vector $\mathbf{f}_i$ using a pretrained DINOv2 [6] vision encoder, followed by feature pooling guided by the mask's spatial location.

The core of our framework is a dynamic graph $\mathcal{G} = (V, E)$ that models the N candidates. Each node $v_i \in V$ is initialized with its feature vector $\mathbf{h}_i$. The edge weights w_{ij} between nodes

are computed adaptively based on spatial proximity (IoU), semantic affinity (cosine similarity), and an uncertainty score $\mathcal{U}(m_i)$ from SAM. We employ a Graph Transformer (GTr) for contextual reasoning. At each layer l, a node's representation $\mathbf{h}_i^{(l)}$ is updated via a multi-head attention mechanism over its neighborhood $\mathcal{N}_i$, and is subsequently refined via cross-attention with the guidance vectors G_s:

$$\mathbf{h}_{i,\text{attn}}^{(l)} = \text{Attention}(\mathbf{W}_Q \mathbf{h}_i^{(l)}, \mathbf{W}_K \mathbf{H}_{\mathcal{N}_i}^{(l)}, \mathbf{W}_V \mathbf{H}_{\mathcal{N}_i}^{(l)}) \quad (2)$$

$$\mathbf{h}_i^{(l+1)} = \text{CrossAttention}(\mathbf{h}_{i,\text{attn}}^{(l)}, G_s) \quad (3)$$

After L layers, a final classification head selects the masks that best align with the user's intent.

The edges E in the graph are constructed to capture both spatial and semantic relationships between masks. An edge is established between two nodes v_i and v_j if their corresponding masks m_i and m_j satisfy a predefined condition. The edge weight e_{ij} is a function of their spatial overlap, measured by the Intersection over Union (IoU), and their semantic similarity, computed as the cosine similarity between their feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$. This graph structure enables collaborative reasoning. We employ a graph convolutional network to update the node representations by aggregating information from their neighbors. The updated feature for node v_i is computed as:

$$\mathbf{f'}_i = \sigma\left(\sum_{j \in \mathcal{N}(i)} e_{ij} \mathbf{W} \mathbf{f}_j\right) \quad (4)$$

where $\mathcal{N}(i)$ is the set of neighbors of node i, $\mathbf{W}$ is a learnable weight matrix, and σ is a non-linear activation function. After several layers of graph convolution, the refined node features $\{\mathbf{f'}_1, \ldots, \mathbf{f'}_N\}$ incorporate contextual information from related masks. Finally, a binary classification head, conditioned on the semantic guidance vector $\mathbf{g}_s$, determines whether each mask should be included in the final output set. This is achieved by computing a matching score for each mask and applying a threshold.

D. Modular Fine-Tuning Strategy

To ensure efficiency, we employ a Parameter-Efficient Fine-Tuning (PEFT) strategy where foundation model backbones are frozen. We insert trainable QLoRA matrices and Prompt Tuning vectors into LLaMA 3-V, and a lightweight Adapter layer into the GTr module. This hybrid approach ensures that the total number of updated parameters is less than 0.5% of the entire model, enabling fine-tuning on a single GPU. Training large-scale models from scratch is computationally prohibitive. Therefore, we employ a parameter-efficient, modular fine-tuning strategy. The backbones of LLaMA 3-V, GroundingDINO, SAM, and DINOv2 are kept frozen. We introduce a small number of trainable parameters at strategic locations.

Semantic Generation Module: We insert QLoRA [7] structures into select Transformer layers of LLaMA 3-V. This involves training only the low-rank adaptation matrices, allowing the model to adapt its language understanding capabilities to our specific task.

Mask Selection Module: We introduce a lightweight Adapter layer within the graph convolutional network. This small, learnable module enhances the fusion of semantic guidance and graph features, improving the discriminative power of the selection process.

This approach ensures that the extensive knowledge encoded in the pretrained models is preserved while enabling efficient adaptation to the target domain. The total number of updated parameters is less than 1% of the entire model, significantly reducing memory and computational requirements.

E. Training Objective and Loss Functions

The end-to-end training is guided by a composite loss function $\mathcal{L}_{\text{total}}$, a weighted sum of three components which correspond to the loss curves in our analysis:

$$\mathcal{L}_{\text{total}} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{align}\mathcal{L}_{align} + \lambda_{reg}\mathcal{L}_{reg} \quad (5)$$

$$\mathcal{L}_{align} = \sum_{i=1}^{N} \max(0, D(\mathbf{g}_s, \mathbf{h}_i^+) - D(\mathbf{g}_s, \mathbf{h}_i^-) + \alpha) \quad (6)$$

$$\mathcal{L}_{reg} = \frac{1}{N} \sum_{i=1}^{N} \text{smooth}_{L1}(v_i - \hat{v}_i) \quad (7)$$

where $D(\cdot,\cdot)$ is a distance function and α is a margin. $\mathcal{L}_{reg}$ provides a fine-grained signal, for instance, by predicting the IoU of each mask. We use the Smooth L1 loss. v_i is the predicted IoU and $\hat{v}_i$ is its ground-truth value.

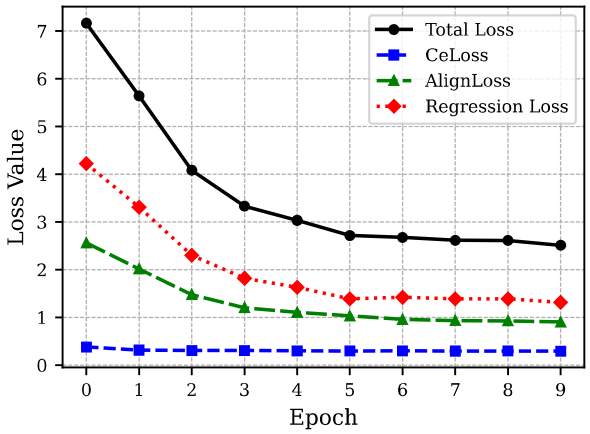

Fig. 2. Validation of the training stability. The Total Loss, a weighted sum of the CeLoss, AlignLoss, and Regression Loss, converges smoothly, confirming the effectiveness of our multi-task learning objective.

III. EXPERIMENTS & RESULTS

A. Dataset

The dataset used is the Complex Urban Referring Segmentation Dataset, which collected street views from 478 streets in 20 cities in China. It contains 20000 images with fine-grained linguistic instructions. As this work aims to complete the street view segmentation task for specific scenes, each street view image corresponds to a complex actual scene.

TABLE I
QUANTITATIVE COMPARISON ON CATEGORY AND CLASS SEGMENTATION. WE REPORT INTERSECTION OVER UNION (IOU) AND INSTANCE IOU (IIOU). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	Classes		Categories		Use LLM	Zero-Shot
	IoU	iIoU	IoU	iIoU		
Mainstream Semantic Segmentation Models						
DeepLabv3+ [8]	58.5	35.1	78.2	55.4	✗	✗
PSPNet [9]	57.9	34.5	77.5	54.1	✗	✗
HRNetV2-W48 [10]	60.2	37.8	80.1	60.5	✗	✗
SegFormer [11]	61.5	39.2	81.9	63.3	✗	✗
Mask2Former [12]	63.4	41.5	83.5	66.8	✗	✗
Fast-SCNN [13]	51.3	29.8	69.7	48.2	✗	✗
Referring Segmentation Baselines						
CRIS [14]	59.8	36.4	80.5	61.2	✗	✓
SEEM [15]	61.3	38.2	82.2	65.4	✗	✓
LLM-based Methods (Fine-tuned)						
LISA [16]	61.9	33.6	81.6	60.9	✓	✗
GLaMM [17]	58.3	37.4	83.4	67.2	✓	✗
LLM-Seg [18]	58.0	31.8	78.2	58.4	✓	✗
Ablation of Our Method (Fine-tuned)						
Baseline (LLaMA3-V + SAM)	56.1	34.2	79.8	66.4	✓	✗
+ Multimodal Semantic Generation (MSG)	57.0	32.0	79.1	61.9	✓	✗
+ Candidate Mask Generation (CMG)	59.1	28.1	79.5	57.9	✓	✗
+ Dynamic Graph-Structured Mask Selection (DGSMS)	62.5	34.4	82.7	66.0	✓	✗
Comparison of Fine-tuning Strategies (Ours)						
Full Model w/o Modular Fine-Tuning Strategy	63.1	34.5	81.2	58.7	✓	✗
Full Model w/o QLoRA on LLaMA3-V	64.8	34.9	81.3	58.7	✓	✗
Full Model w/o Adapter on Selection Module	66.4	**46.7**	82.8	67.4	✓	✗
Ours (Full Model w/ MFTS)	**67.1**	42.0	**86.5**	**71.1**	✓	✗

B. Implementation Details

This project employs PyTorch 2.0+ as the primary deep learning framework, leveraging its torch.compile feature for dynamic graph compilation optimization. Multimodal model processing is based on the Transformers 4.35+ library, which integrates pre-trained model interfaces such as LLaMA and SAM. For the graph neural network component, PyTorch Geometric 2.4+ is utilized to provide efficient graph operations. To handle large-scale model training, DeepSpeed/FSDP is introduced as the distributed training framework.

The development environment requires Python 3.9 or higher, with CUDA 11.8 or 12.1 together with cuDNN 8.9+. Core dependencies include torch 2.1.0, transformers 4.36.0, torch-geometric 2.4.0, and deepspeed 0.12.0. QLoRA-based quantization training requires bitsandbytes 0.41.0, while experiment management is performed using wandb 0.16.0.

The minimum hardware configuration suitable for single-GPU fine-tuning scenarios consists of one NVIDIA A100 40GB or RTX 4090 24GB GPU, paired with a 32-core AMD EPYC or Intel Xeon processor, 128 GB of DDR4 RAM, and 2 TB of NVMe SSD storage. This setup supports QLoRA 4-bit quantized fine-tuning with a batch size of 4-8.

C. Quantitative Results

We conducted a comprehensive evaluation of our framework on challenging street-view datasets, reporting Intersection over Union (IoU) and instance IoU (iIoU) for both broad **Categories** and fine-grained **Classes**. As detailed in Table I, our proposed framework, **Ours (w/ MFTS)**, establishes a new state-of-the-art across all metrics. It significantly outperforms strong mainstream baselines like **Mask2Former** (Categories IoU 86.5 vs. 83.5) and specialized referring segmentation models like **SEEM** (86.5 vs. 82.2). Crucially, even when compared to other fine-tuned LLM-based methods, our model demonstrates a clear performance advantage, surpassing **LISA** by a margin of nearly 5 points in Categories IoU. This highlights the architectural benefits of our dynamic graph reasoning module.

D. Ablation Study

Our ablation studies further validate the efficacy of our design. An incremental analysis of our model's components reveals that the **Dynamic Graph-Structured Mask Selection (DGSMS)** module provides the most substantial performance increase, underscoring the importance of our graph-based reasoning approach. The fine-tuning strategy comparison further solidifies our methodology. Our proposed **MFTS** not only surpasses a fully fine-tuned model (`Full Model w/o MFTS`) by a significant margin (Categories IoU of 86.5 vs. 81.2), but we also show that the removal of either the **QLoRA** or **Adapter** components leads to a notable drop in accuracy. This dual result confirms that our parameter-efficient strategy is not only more efficient but also more effective than traditional full fine-tuning for this task.

	Input	LLM-based Methods	LISA	GLAMM	Ours	GT
Where can we see cars in this street scene? Please output the segmentation mask.						
Where can we see pedestrians in this street scene? Please output the segmentation mask.						
Where can we see traffic lights in this street scene? Please output the segmentation mask.						
Where can we see the derivable road area in this street scene? Please output the segmentation mask.						

Fig. 3. Visual comparison of ours with SOTA methods. Our model shows high-quality segmentation results even for multiple instances.

E. Qualitative Results

We visualize some cases of the ReasonSeg validation split Figure 3 shows some success cases of our method. In order to highlight the performance of this framework in practical scenarios, we selected four representative natural language instructions for visual comparative analysis.We compare it with mainstream image segmentation methods and demonstrate real annotation (GT) as a reference.

The experimental results show that our proposed framework excels in understanding complex instructions and performing fine-grained mask selection. For car segmentation, our method accurately identifies and segments multiple vehicles, maintaining high completeness even under partial occlusion or varying lighting conditions, whereas LISA and GLaMM exhibit missed detections or blurred boundaries.For pedestrian segmentation, our approach effectively distinguishes individuals in crowded scenes and captures contour details under dynamic poses, outperforming the compared methods significantly.For traffic light segmentation, our model not only correctly locates traffic lights but also distinguishes them from similar objects (e.g., street lamps, billboards) in cluttered backgrounds, demonstrating the effectiveness of semantic guidance and graph-structured reasoning.For drivable road area segmentation, our framework shows a deep understanding of scene structure, accurately separating roads from sidewalks and green belts, while maintaining smooth and continuous boundaries that align with real driving scenarios.In summary, the visual results validate the advantages of our framework in semantic understanding accuracy, mask selection consistency, and scene adaptability, particularly in handling semantic ambiguity, multi-object interactions, and fine-grained structural segmentation tasks.

IV. CONCLUSION

We introduced a novel framework for language-driven segmentation that achieves state-of-the-art performance by fusing the semantic guidance of LLaMA3-V with dynamic graph-based reasoning. Our modular, parameter-efficient fine-tuning strategy proves both highly effective and practical for adapting large foundation models to specialized domains. This work paves the way for more accurate and context-aware visual perception systems. We hope our work can shed new light on the direction of Street scene segmentation and the field of autonomous driving specific tasks.

REFERENCES

[1] M. AI, "The llama 3 vision model," https://ai.meta.com/research/publications/llama-3/, 2024, accessed: 2025-09-01.

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[3] T. Tang, S. Xu, Y. Wu, and Z. Lu, "Causal-sam-llm: Large language models as causal reasoners for robust medical segmentation," 2025. [Online]. Available: https://arxiv.org/abs/2507.03585

[4] Y. Li, Z. Lu, F. Tang, S. Lai, M. Hu, Y. Zhang, H. Xue, Z. Wu, I. Razzak, Q. Li, and J. Su, "Rhythm of opinion: A hawkes-graph framework for dynamic propagation analysis," 2025. [Online]. Available: https://arxiv.org/abs/2504.15072

[5] S. Liu, Z. Zeng, T. Ren, H. Li, F. Zhang, X. Li, and J. Sun, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[6] M. Oquab, T. Darcet, T. X. Moutakanni, D. Mane, D. Haziza, F. Massa, M. Szafraniec, V. Birodkar, W.-Y. Lo, A. Haziza, I. Misra, P. Bojanowski, A. Joulin, I. Laptev, A. Vedaldi, Y. LeCun, and H. Jegou, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] B. Cheng, I. Misra, R. Girdhar, A. G. Schwing, and A. Kirillov, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.

[14] D.-J. Wang, L. Zhang, G.-H. Chen, Z.-H. Liu, X.-D. Zhang, H. Lu, and X.-R. Wang, "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 616–11 625.

[15] X. Zou, Z. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Sandeep, L. Yuan, V. Le, Y. Liu, L. Wang, Y. J. Liu, J. Wang, and J. Gao, "Segment everything everywhere all at once," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6053–6065.

[16] X. Lai, Z. Tian, Y. Liu, Y. Chen, G. Chen, L. Wang, Y. Li, and J. Jia, "Lisa: Reasoning segmentation with large language models," *arXiv preprint arXiv:2308.00692*, 2023.

[17] H. Seo, D.-h. Lee, G. Kim, T. Kim, S. Lee, and B. Han, "Glamm: Grounded large multimodal model for attentive multi-object comprehension," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 13 328–13 337.

[18] J. Wang and L. Ke, "Llm-seg: Bridging image segmentation and large language model reasoning," 2024. [Online]. Available: https://arxiv.org/abs/2404.08767