LLAMASEG: IMAGE SEGMENTATION VIA AUTOREGRESSIVE MASK GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present **LlamaSeg**, a visual autoregressive framework that unifies multiple image segmentation tasks via natural language instructions. By reformulating segmentation as visual generation, LlamaSeg encodes masks as visual tokens and uses a LLaMA-style Transformer for direct next-token prediction, naturally fitting segmentation into autoregressive architectures. To support large-scale training, we introduce a data annotation pipeline and construct the **SA-OVRS** dataset, which contains **2M** segmentation masks annotated with over **5,800** open-vocabulary labels or diverse textual descriptions, spanning diverse real-world scenarios. This enables our model to localize objects in images based on text prompts and to generate fine-grained masks. We further introduce the composite metric average Hausdorff Distance (d_{AHD}) to evaluate mask contour fidelity for generative models better. Experiments show that LlamaSeg consistently outperforms existing generative approaches on multiple segmentation benchmarks and delivers finer, more accurate segmentation masks.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have demonstrated the potential of unifying diverse vision-language tasks under an autoregressive framework (Alayrac et al., 2022; Dai et al., 2023; Xu et al., 2024; Zhang et al., 2024). However, extending this paradigm to dense segmentation remains challenging, as it demands precise pixel-level generation (Wang et al., 2023; Wu et al., 2024). As illustrated in Figure 1, existing autoregressive segmentation approaches can be broadly categorized into three paradigms. As illustrated in Figure 1, existing autoregressive segmentation approaches can be broadly grouped into three paradigms. Embedding-based representations (Lai et al., 2024; Xia et al., 2024) predict a latent mask token such as <SEG> and then delegate pixel-level mask generation to a specialist model like SAM (Kirillov et al., 2023). Although effective, this reliance on external segmentation experts introduces task-specific components and undermines the goal of a unified end-to-end framework. Coordinate-driven modeling (Wang et al., 2023; Pramanick et al., 2024) encodes each segmentation mask as a sequence of polygon vertices. While naturally aligned with next-token prediction, it requires long coordinate sequences to capture intricate boundaries, reducing efficiency and scalability for dense semantic segmentation. Text-conditioned generation (Wang et al., 2024a; Lan et al., 2024) formulates segmentation as a text generation task by emitting category tokens or free-form descriptions that must later be converted into pixel masks, which inevitably limits spatial precision. These limitations highlight that current autoregressive methods remain inadequate for fine-grained, end-to-end segmentation within a single unified framework.

However, the progress of autoregressive image generation (Esser et al., 2021; Razavi et al., 2019; Sun et al., 2024) shows that appropriate representations and objectives enable these frameworks to model complex visuals. These methods, which typically condition image synthesis on category labels, text prompts, or visual cues, highlight the expressive power of token-based modeling. Their potential remains underexplored in structured prediction tasks with deterministic ground truth, such as segmentation. **Building on these insights, we aim to bridge this gap by revisiting autoregressive modeling for segmentation and devising a representation that predicts high-fidelity masks as discrete visual tokens within a unified next-token prediction framework.**

In this paper, we propose **LlamaSeg**, which reformulates image segmentation as a visual generation task, as shown in panel (d) of Figure 1. We treat segmentation masks as image-like structures, which

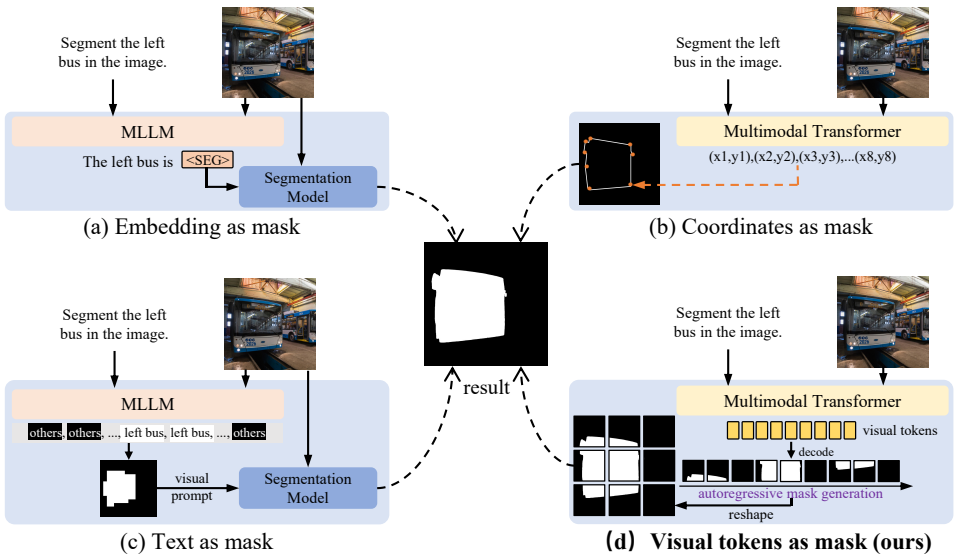

Figure 1: Comparison of autoregressive segmentation paradigms. (a) Embedding as mask. (b) Coordinates as mask. (c) Text as mask. (d) Visual tokens as mask (ours), which directly autoregresses visual tokens to capture fine-grained structures and yields precise pixel-level masks.

are discretized into token sequences via VQGAN (Esser et al., 2021). A LLaMA-based autoregressive model (Touvron et al., 2023a;b) then generates mask tokens conditioned on the encoded visual inputs, maintaining the standard next-token prediction paradigm, thereby facilitating the integration of segmentation tasks into general autoregressive frameworks. Our framework is scalable across different model sizes and image resolutions, and supports training from scratch as well as fine-tuning on pretrained MLLMs. To address the lack of suitable training data and evaluation tools for this paradigm, we introduce a large-scale dataset, SA-OVRS, containing 2 million segmentation instances annotated with open-vocabulary labels or rich textual descriptions derived from SA-1B (Kirillov et al., 2023), and propose the average Hausdorff Distance (d_{AHD}) under multi-level IoU thresholds as a new metric to assess contour fidelity in generated masks. Experiments on multiple segmentation benchmarks demonstrate that our model produces high-quality, fine-grained masks and outperforms existing visual generative approaches. Our main contributions are as follows:

1. We propose **LlamaSeg**, a visual autoregressive segmentation framework that shares the core next-token generation paradigm of Large Language Models (LLMs), allowing segmentation to be natively unified within the LLMs architecture while effectively modeling visual context and producing high-quality, fine-grained masks.
2. We develop a data annotation pipeline and construct the **SA-OVRS** dataset comprising 2M masks with over 5,800 open-vocabulary or richly descriptive labels. The dataset and the associated annotation pipeline are released as reusable tools to advance future segmentation research and to facilitate the creation of new high-quality datasets.
3. We introduce a composite evaluation metric d_{AHD} to assess mask-contour fidelity and demonstrate that LlamaSeg surpasses most existing visual generative models on diverse semantic and referring segmentation benchmarks, delivering more precise pixel-level masks.

2 Related Work

2.1 Multimodal Large Language Models

The research on MLLMs is closely related to cross-modal representation learning. With the rapid advancement of LLMs (Vaswani et al., 2017; Radford et al., 2018; Brown et al., 2020; Chowdhery et al., 2023), recent efforts have increasingly focused on aligning visual representations with the language embedding space. BLIP-2 (Li et al., 2023) and InstructBLIP (Dai et al., 2023) incorporate lightweight Querying Transformer modules to inject visual context. The LLaVA series (Liu et al., 2023b; 2024a; Li et al., 2024) employs simple linear projection layers to map image features. The Qwen-VL family (Bai et al., 2023; Wang et al., 2024b) leverages cross-attention mechanisms with

learnable queries for tighter vision-language integration. These approaches collectively enable large language models to interpret and reason over visual content effectively. Beyond image-level understanding, recent studies have begun to extend vision-language models toward dense prediction tasks such as image segmentation, which require fine-grained spatial reasoning and localized visual grounding.

2.2 Image Segmentation

Transformer-based segmentation method. Segmentation models based on the Transformer (Vaswani et al., 2017) architecture typically include MaskFormer (Cheng et al., 2021), SegFormer (Xie et al., 2021), and SAM (Kirillov et al., 2023). Since language models such as BERT (Devlin et al., 2019) also adopt Transformer architectures, this shared foundation facilitates cross-modal information fusion. Several approaches (Ding et al., 2021; Wang et al., 2022; Liu et al., 2023a) have explored referring image segmentation built upon this architecture. Recent studies have increasingly focused on MLLMs. For example, LISA (Lai et al., 2024) integrates SAM and LLaVA (Liu et al., 2023b), leveraging token embeddings to guide the segmentation model. Subsequent methods (Ren et al., 2024; Xia et al., 2024) build upon this framework with further enhancements. Despite their strong performance, these approaches typically adopt a pipeline architecture, which increases model complexity and training difficulty.

Generative segmentation method. Generative methods recast segmentation as a mask-generation problem conditioned on the input image, providing a streamlined alternative to traditional pipeline-based approaches. GSS (Chen et al., 2023), for example, links the mask posterior distribution with the latent prior of the input image for semantic segmentation. Despite adopting the generative paradigm, these approaches remain confined to vision-only tasks. Unified-IO (Lu et al., 2022) and Unified-IO2 (Lu et al., 2024) extend to multiple modalities and include segmentation, but neither explores it deeply. Unified-IO struggles with complex language understanding, while Unified-IO2 employs a two-stage procedure dependent on bounding boxes. Such reliance on task-specific modules or intermediate representations hinders seamless end-to-end segmentation and limits generalization.

2.3 Autoregressive Image Generation

Autoregressive modeling enables end-to-end image synthesis pipelines, offering an appealing alternative to traditional modular designs. Recent models (Tian et al., 2024; Team) discretize images into 1D sequences with image tokenizers for next-token prediction. Subsequent work (Yu et al., 2021; Sun et al., 2024) refines tokenizer design and benchmarks against continuous models such as VAE (Kingma et al., 2013), showing competitive generation quality. Despite these advances, applying autoregressive models to dense prediction tasks like segmentation remains largely unexplored.

3 LlamaSeg

3.1 Overview

The overall framework of LlamaSeg is illustrated in Fig. 2. For the LLaMA-based framework, we employ image and text encoders to extract input features, which are projected to align with the embedding dimension of LLaMA (Touvron et al., 2023a;b) and concatenated with learnable separator tokens to form the input sequence. The LLaMA model then autoregressively generates mask tokens as output. We adopt VQGAN (Esser et al., 2021) as the mask tokenizer. During training, the mask tokenizer encodes the ground truth mask into code indices, which serve as the training targets for the LLaMA model. During inference, the mask tokenizer receives the mask tokens generated by the LLaMA model, retrieves the corresponding codes from the codebook, and decodes them to produce the segmentation mask. For the MLLM-based framework utilizing a pretrained MLLM, the text encoder is omitted, while the remaining procedure remains unchanged.

3.2 Mask Tokenizer

We use an image tokenizer to transform segmentation masks into discrete tokens. Specifically, we use a VQGAN from LlamaGen (Sun et al., 2024) with a downsample rate of 16 and a codebook $Z \in \mathbb{R}^{K \times d_{vq}}$ containing K vectors, where d_{vq} is the vector dimension. We consider the segmentation mask as a special RGB image composed only of black and white pixels. During training, the encoder $\mathcal{E}$ transforms the normalized mask $M_I \in \mathbb{R}^{H \times W \times 3}$ into features $f_m \in \mathbb{R}^{h \times w \times d_{vq}}$, and the quantizer $\mathcal{Q}$ converts features into discrete tokens $q \in [K]^{h \times w}$. The quantizer matches each vector in f_m with

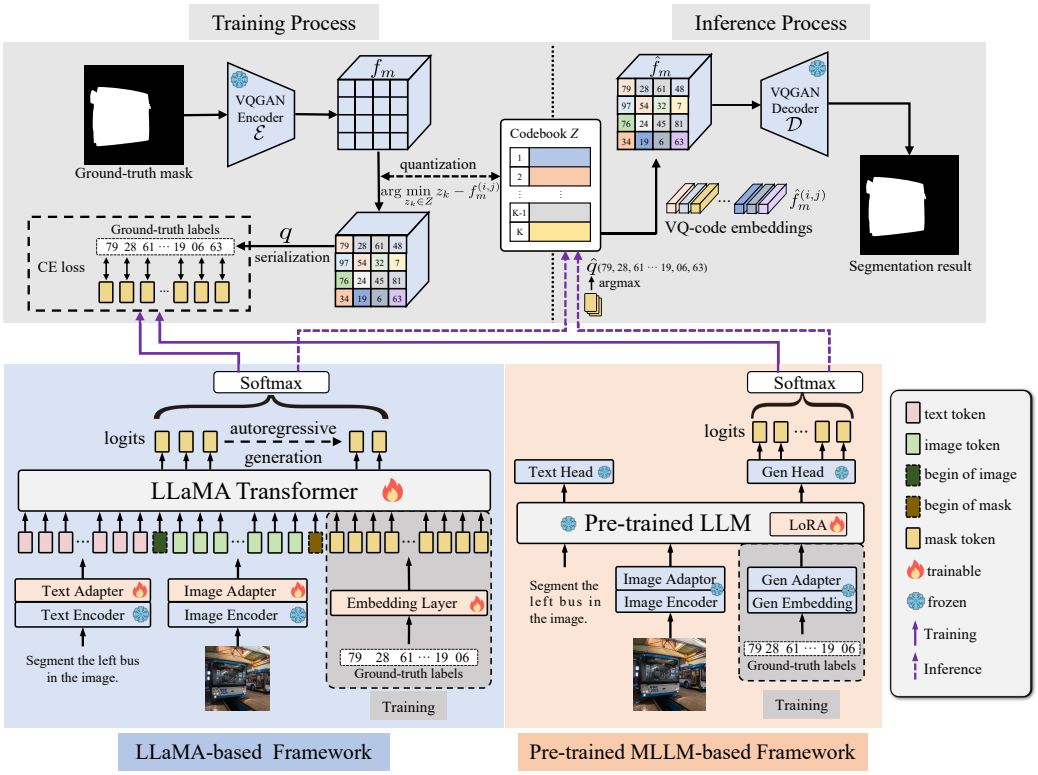

Figure 2: Overall framework of **LlamaSeg**. The model consists of a VQGAN-based mask tokenizer and a LLaMA-style autoregressive generator. During training, ground-truth masks are quantized into discrete codes to supervise token prediction; during inference, predicted tokens are decoded to produce segmentation masks. We support training either from scratch or by adapting a pre-trained MLLM with lightweight modules.

the closest vector in Z in terms of Euclidean distance and finds the corresponding code index:

$$q^{(i,j)} = \arg\min_{z_k \in Z} ||z_k - f_m^{(i,j)}|| \in [K], \tag{1}$$

where z_k is the vector within Z. Flattening the code indices to form a 1D sequence can serve as supervised training data for an autoregressive model. At inference time, for a 1D token output by an autoregressive model, the token IDs are equivalent to the code indices in the mask tokenizer. By "looking up in table", vectors $\hat{z}_k \in \mathbb{R}^{d_{vq}}$ can be found in the codebook according to $\hat{q} \in [K]^{h \times w}$, which are rearranged into a 2D shape from token IDs and form the feature map $\hat{f}_m \in \mathbb{R}^{h \times w \times d_{vq}}$. The normalized mask image $\hat{M}_I \in [-1, 1] \subset \mathbb{R}^{H \times W \times 3}$ is reconstructed by the decoder $\mathcal{D}$ using $\hat{f}_m$. Average $\hat{M}_I$ across the channel dimension and then apply a binarization step to produce the final segmentation result $\hat{M}_{bin}^{(i,j)}$:

$$\hat{M} = \frac{1}{3}\sum_{i=0}^{2}(\hat{M}_I[:,:,i]), \qquad \hat{M}_{bin}^{(i,j)} = \begin{cases} 1, & \hat{M}^{(i,j)} \geq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

3.3 Image and Text Feature Extraction

We construct textual inputs by combining object labels or descriptions with 10 predefined templates, such as "Produce a segmentation mask for the {*object name*}." For the autoregressive model trained from scratch, we employ a frozen SigLIP2 (Tschannen et al., 2025) with a patch size of 16 to encode both images and text. The vision encoder's patch size matches the mask tokenizer's downsampling rate, ensuring that input tokens correspond to the same pixel regions as the segmentation mask for precise token-level alignment. Furthermore, trainable adaptors are used to project the image and text features into the embedding dimension of the autoregressive model. Each adaptor has two linear layers and a GELU (Hendrycks & Gimpel, 2016) activation function. We consider visual encoders with input resolutions of 256 and 384, corresponding to 256 and 576 visual tokens, respectively.

3.4 AUTOREGRESSIVE MODEL

LLaMA Structure. The autoregressive model is based on LLaMA (Touvron et al., 2023a;b) model Each Transformer layer employs 1D RoPE, treating both images and masks as 1D sequences. We utilize two model variants of different sizes. The base model comprises 770 million parameters, 16 layers, a hidden size of 1920, and 20 attention heads. The large model contains 1.5 billion parameters, 22 layers, a hidden size of 2304, and 32 attention heads. During training, input image and text features are concatenated along the channel dimension. To distinguish between text, image, and mask tokens, we introduce learnable separators: `<BOI>` (*Begin-Of-Image*) and `<BOM>` (*Begin-Of-Mask*). The input sequence during training is structured as follows:

`[text tokens]<BOI>[image tokens]<BOM>[mask tokens].`

The model is trained using a cross-entropy loss, computed solely on the mask tokens. At inference time, only `<BOM>` and the preceding tokens are fed into the model, allowing it to generate mask tokens autoregressively.

Utilization of pre-trained MLLM. We adopt Janus Pro (Chen et al., 2025), an MLLM for multi-modal understanding and generation with separate language and image pathways. For multimodal understanding, images are encoded by the SigLIP (Zhai et al., 2023) vision encoder. For image generation, however, visual inputs are synthesized via an image tokenizer (Esser et al., 2021). The model operates at an image resolution of 384. To enable visual generation, we utilize components responsible for producing mask tokens, denoted as "Gen Embedding", "Gen Adapter", and "Gen Head" in Fig. 2. These components are kept frozen to retain the generative capabilities acquired through large-scale data training. During training, the sequence input to the model is:

`<|User|>:` `[text tokens]<image>` **`<|Assistant|>:`** `<begin_of_mask>[mask tokens]`

where `<|User|>` and `<|Assistant|>` denote the dialogue roles, `<image>` represents the input image, and `<begin_of_mask>` indicates the beginning of the mask token sequence. During inference, only `<begin_of_mask>` and its preceding context are provided as input.

4 SA-OVRS DATASET CONSTRUCTION

Existing methods (Lai et al., 2024; Xia et al., 2024) leverage semantic and referring segmentation datasets for language-guided segmentation. Semantic datasets such as COCO-Stuff (Caesar et al., 2018) treat category names as text, but hierarchical labels (e.g., "vegetables" vs. "broccoli") are split into independent classes, leading to semantic inconsistencies and limiting open-vocabulary learning. In addition, existing datasets are too small to support the dense cross-modal alignment required by autoregressive models. Figure 3 illustrates the two-stage annotation pipeline used to construct SA-OVRS. In the first stage, we generate candidate open-vocabulary labels for SA-1B (Kirillov et al., 2023) images and match them with their corresponding masks through quality filtering, while the second stage generates natural-language descriptions for each object instance, encompassing both referring and reasoning expressions. More details are presented in the following subsection.

4.1 LABEL GENERATION AND MASK MATCHING

For each SA-1B image, we first obtain candidate open-vocabulary labels using Qwen2-VL-72B with a tailored prompt, limiting outputs to at most ten labels per image. GroundingDINO (Liu et al., 2024b) then detects bounding boxes for each label. To ensure high-quality matches, we apply a three-step filtering procedure: **(i)** Remove labels associated with excessive boxes (more than four). **(ii)** Eliminate nested detections where a smaller box is almost entirely enclosed in a larger one (IoU > 0.97). **(iii)** Keep only matches with sufficient confidence (score > 0.3) and strong mask overlap (IoU > 0.85 for single-box labels, 0.9 for multi-box labels). Masks linked to the same label are merged to produce semantic segmentation samples.

4.2 TEXTUAL DATA GENERATION AND VERIFICATION

To produce unique natural-language descriptions for each instance, we highlight every matched mask with a green contour and prompt Qwen2-VL-72B to generate a referring expression that distinguishes it from all other objects in the image. We further perform cross-verification by recoloring the mask

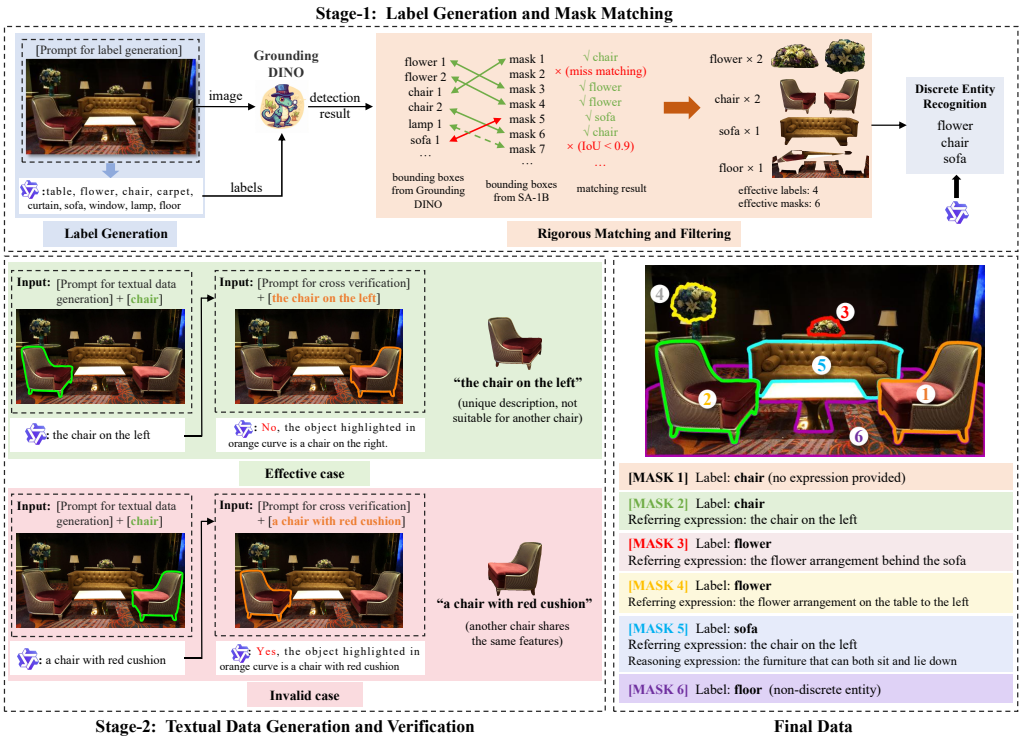

Figure 3: A two-stage annotation pipeline for the construction of the **SA-OVRS** dataset. **Stage 1: Label Generation and Mask Matching** (top) leverages Qwen2-VL to generate candidate open-vocabulary labels, which are then aligned with SA-1B masks using GroundingDINO through strict matching and filtering procedures. **Stage 2: Textual Data Generation and Verification** (bottom left) produces high-quality referring expressions and reasoning descriptions.

contour and prompting the model to confirm that the expression is unique and unambiguous. This ensures high-quality referring segmentation annotations.

To further enhance data diversity, we introduce reasoning segmentation. Unlike referring expressions that directly identify objects, reasoning expressions omit object names and instead describe them by function or commonsense attributes (e.g., "the furniture that can both sit and lie down" $\rightarrow$ sofa). This design encourages models to leverage semantic reasoning beyond surface lexical cues. This process differs in prompt design and omits the verification step.

In total, SA-OVRS contains **1.93M validated instance masks, 1.15M semantic samples, and 850K textual expressions (800K referring, 50K reasoning)** spanning over 5,800 open-vocabulary labels.

5 Experiments

5.1 Experiment Settings

Training Data. For semantic segmentation, we select ADE20K (Zhou et al., 2019) and COCO-Stuff (Caesar et al., 2018) dataset. For referring segmentation, we choose the refCOCO (Yu et al., 2016) series and refCLEF (Kazemzadeh et al., 2014) datasets. Additionally, we adopt the semantic and referring segmentation data from SA-OVRS.

Tasks and Evaluation Metrics. For closed-set and open-vocabulary semantic segmentation, we employed mean IoU (mIoU). For referring segmentation, we follow prior works and use the cumulative intersection over cumulative union (cIoU). To evaluate the contour accuracy of masks produced by visual generative models, we propose the *average Hausdorff Distance* (d_{AHD}) metric under multi-level IoU thresholds. Given the boundary point sets of a predicted mask and GT mask, d_{AHD} is defined as:

$$d_{AHD}(X,Y) = \frac{1}{2}\left(\frac{1}{X}\sum_{x\in X}\min_{y\in Y} dist(x,y) + \frac{1}{Y}\sum_{y\in Y}\min_{x\in X} dist(x,y)\right), \tag{3}$$

Table 1: Comparisons with visual generative models and MLLMs on closed-set and open-vocabulary semantic segmentation datasets. Unified-IO and Unified-IO2 are evaluated using their open-source weights. "Params" denotes model size and "384 pix." indicates a 384-pixel input resolution. **Bold** marks the best result.

Type	Model	Params	ADE20K	COCO-Stuff	PC-459	PC-59	PAS-20
MLLM	LaSagnA (Wei et al., 2024)	7B	42.0	43.9	9.8	39.6	61.8
Visual generative model	Unified-IO-Large	0.77B	39.9	50.7	-	62.8	62.7
	Unified-IO-XL	2.9B	45.2	55.2	-	**64.0**	74.9
	Unified-IO2-Large	1.1B	35.6	48.8	-	55.8	71.5
	Unified-IO2-XL	3.2B	39.4	49.9	-	56.9	71.5
	Unified-IO2-XXL	6.6B	51.9	55.1	-	-	-
Visual generative model (ours)	**LlamaSeg-B**	0.77B	50.5	54.5	28.5	61.7	74.5
	LlamaSeg-1B$_{(MLLM)}$	1B	45.1	50.1	**35.6**	58.2	71.1
	LlamaSeg-L	1.5B	**52.0**	**55.7**	35.5	62.5	**75.1**
	LlamaSeg-L$_{(384\ pix.)}$	1.5B	55.9 (+3.9)	58.1 (+2.4)	36.7 (+1.2)	63.8 (+1.3)	77.1 (+2.0)

Table 2: Comparison results on referring segmentation datasets. We evaluated the performance of Unified-IO and Unified-IO2 on a subset of datasets.Green indicates the best result among discriminative segmentation models, while orange indicates the best result among visual generative models.

Type	Model	Params	refCOCO			refCOCO+			refCOCOg	
			val	testA	testB	val	testA	testB	val	test
Discriminative segmentation model	LAVT (Yang et al., 2022)	-	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
	ReLA (Liu et al., 2023a)	-	**73.8**	**76.5**	**70.2**	**66.0**	**71.0**	**57.7**	**65.0**	**66.0**
Visual generative model	Unified-IO-Large	0.77B	35.7	39.2	-	34.7	39.4	-	42.2	42.3
	Unified-IO-XL	2.9B	42.4	49.5	-	39.4	42.7	-	50.3	50.3
	Unified-IO2-Large	1.1B	40.3	47.2	-	33.0	39.1	-	43.2	44.1
	Unified-IO2-XL	3.2B	50.7	55.7	-	39.4	**47.4**	-	**52.6**	**54.5**
	Unified-IO2-XXL	6.6B	54.8	-	-	**44.8**	-	-	-	-
Visual generative model (ours)	**LlamaSeg-B**	0.77B	50.9	56.1	38.9	38.9	44.5	32.9	51.0	50.5
	LlamaSeg-1B$_{(MLLM)}$	1B	52.3	57.8	47.1	44.6	51.6	36.5	49.0	49.7
	LlamaSeg-L	1.5B	**56.5**	**60.5**	**52.6**	41.6	46.2	**36.2**	51.0	50.4

where $dist(x, y)$ denotes the Euclidean distance between x and y. This metric captures bidirectional distances between the two point sets, providing an aggregate measure of global boundary alignment, where a **lower score** indicates better performance. For each threshold in [0.5, 0.6, 0.7, 0.8, 0.9], we collect predictions with IoU above it, compute d_{AHD}, and report the mean average Hausdorff distance ($mAHD$)

Pre-training. We adopt a two-stage training protocol consisting of pre-training followed by fine-tuning. During pre-training, the model is trained on the SA-OVRS dataset and the referring segmentation datasets. We use the AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.05, and run for 4 epochs.

Fine-tuning. The semantic and referring segmentation tasks are trained separately on their respective datasets. We use AdamW with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.99$, zero weight decay, and a WarmupCosineDecay scheduler with 1% linear warmup. We fine-tune for 10 epochs on the semantic segmentation datasets and 20 epochs on the referring segmentation datasets. All training is carried out on 8 NVIDIA GPUs (A800 or H20) under PyTorch's distributed data parallel framework.

Model Variants. For the LLaMA-based framework, mask tokens are generated via greedy search. When leveraging an MLLM-based framework, we apply LoRA (Hu et al., 2022) for efficient fine-tuning and remove the unconditional generation component while keeping other configurations unchanged. We develop two models within the LLaMA-based framework (from scratch), namely **LlamaSeg-B (base)** and **LlamaSeg-L (large)**, and further construct **LlamaSeg-1B$_{(MLLM)}$** within the Pre-trained MLLM-based Segmentation Framework (with LoRA/Adapter).

5.2 Main Results

Semantic Segmentation Result Analysis. Table 1 reports closed-set and open-vocabulary semantic segmentation results. LlamaSeg consistently outperforms existing visual generative models across all benchmarks. LlamaSeg-L attains 52.0/55.7 mIoU on ADE20K/COCO-Stuff, surpassing Unified-IO-XL and Unified-IO2-XXL by large margins with fewer parameters. Using higher-resolution mask tokens, LlamaSeg-L (384 pix.) further improves to 55.9 (+3.9) and 58.1 (+2.4), with consistent gains on PC-459/PC-59/PAS-20 (e.g., +2.0 on PAS-20). These results show that our visual-token

Table 3: Comparison of the $mAHD$ metric after normalization to a 256-pixel resolution. Lower values correspond to a closer match between the predicted mask contours and the ground truth. Bold entries highlight the best results.

Model	ADE20K					refCOCO				
	IoU-0.5	IoU-0.6	IoU-0.7	IoU-0.8	IoU-0.9	IoU-0.5	IoU-0.6	IoU-0.7	IoU-0.8	IoU-0.9
Unified-IO-Base	76.41	74.84	72.36	68.72	62.39	85.14	83.22	81.26	80.02	79.06
Unified-IO-Large	76.35	75.31	72.21	67.73	62.62	83.08	81.02	79.81	78.38	75.61
Unified-IO-XL	75.88	75.16	72.35	69.35	65.13	72.58	71.69	71.16	69.99	85.65
Unified-IO2-Large	78.45	76.96	75.02	72.62	67.37	83.32	82.57	81.68	80.59	79.14
Unified-IO2-XL	80.15	78.51	76.48	74.21	67.82	82.05	81.42	80.91	79.95	78.18
Unified-IO2-XXL	79.72	78.02	75.76	71.81	65.96	80.98	80.10	79.45	78.24	75.33
LlamaSeg-B	26.61	25.91	25.61	25.98	29.28	14.40	13.63	12.69	11.36	9.94
LlamaSeg-1B$_{(MLLM)}$	59.24	58.91	59.20	60.77	69.38	45.47	42.41	39.29	37.80	37.28
LlamaSeg-L	**25.54**	**24.73**	**24.28**	**24.40**	**26.51**	**10.45**	**9.68**	**8.73**	**7.42**	**5.27**

formulation scales well with model size and resolution, enabling fine-grained masks and better boundary preservation. Compared with the MLLM-based baseline (LaSagnA) and the Unified-IO series, LlamaSeg achieves higher accuracy while remaining fully autoregressive and end-to-end.

Referring Segmentation Result Analysis. As shown in Table 2, our model outperforms existing visual generative models across multiple referring segmentation datasets under the same parameter constraints. On the RefCOCO dataset, our LlamaSeg-Large achieves a score 1.7 points higher than the best result of Unified-IO2, demonstrating superior language understanding capabilities. In contrast, a noticeable gap remains compared to specialized discriminative segmentation models.

Boundary Accuracy Analysis. Although our model achieves slightly lower overall scores on referring segmentation tasks than other strong baselines, it delivers notably higher boundary precision. We evaluate $mAHD$ under multiple IoU thresholds on ADE20K and RefCOCO (Table 3). While semantic segmentation data—aimed at pixel-level discrimination—often exhibits complex contours, referring segmentation masks are comparatively simpler. Our approach consistently yields lower $mAHD$ than previous visual generative models, indicating more accurate alignment with ground-truth contours and the ability to capture fine object boundaries with high fidelity. Figure 4 further illustrates these improvements in mask quality and segmentation accuracy.

2D Spatial Relationship. To examine whether the model can learn 2D spatial relationships from 1D tokens, we select an example in resolution of 256 and export the attention map in the last self-attention layer of the Transformer model, as shown in Fig. 5. Under the causal attention mechanism, the attention map exhibits a distinct band parallel to the main diagonal, revealing that mask tokens preferentially attend to tokens located in preceding rows but within the same column in the original 2D layout. Moreover, the presence of regular dark vertical stripes indicates that, despite

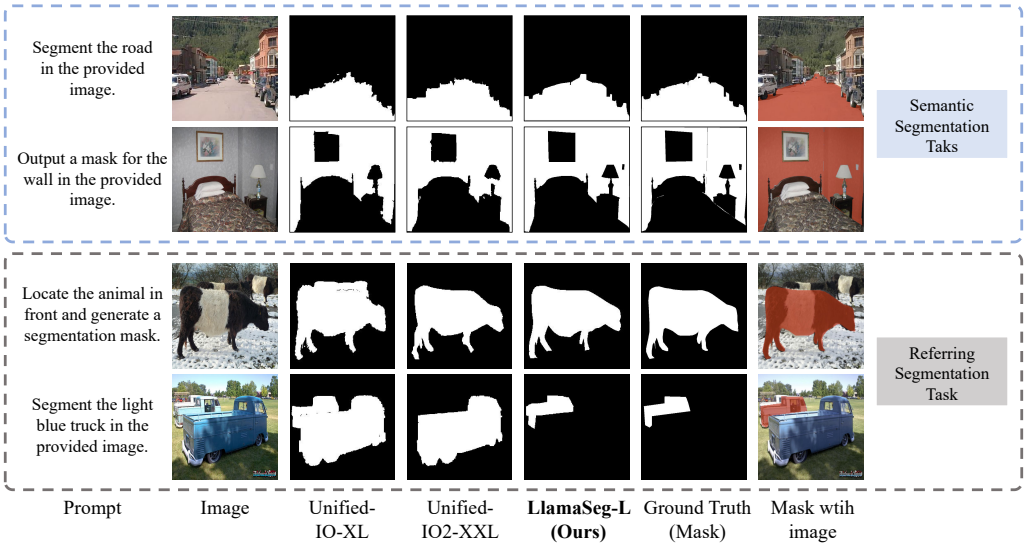

Figure 4: Visual comparison between our model and other visual generative models.

the linear adjacency of row-end and next-row tokens in the 1D sequence, the model consistently suppresses attention from the leading tokens of each row to the trailing tokens of the previous row.

These observations demonstrate that our method does not merely exploit 1D sequential proximity but internally reconstructs and leverages authentic 2D spatial topology.

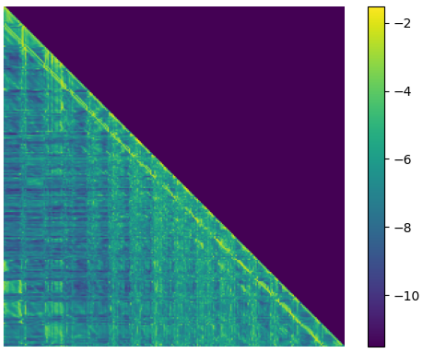

Figure 5: Attention heatmap of mask tokens in the last self-attention layer of the autoregressive model. Scores have been log-transformed to enhance visual contrast.

5.3 Ablation Study

Mask Tokenizer Analysis. We further fine-tune the pre-trained VQGAN on segmentation masks with a batch size of 128. The overall IoU and $mAHD$ results (without IoU threshold constraints) for mask reconstruction are reported in Table 4. While incorporating additional training data slightly improves reconstruction quality, the overall performance remains below that of the original pre-trained model, suggesting that extended fine-tuning leads to a degradation of the learned feature representations.

Decoding Strategy Analysis. We conduct experiments on the RefCOCO validation set using the LlamaSeg-B model to systematically compare several decoding strategies, as reported in Table 5. Given that image segmentation tasks are paired with deterministic ground-truth masks, the model is required to produce a single, optimal prediction rather than diverse outputs. These results demonstrate that greedy decoding is the most effective and reliable strategy for autoregressive image segmentation in this setting.

Table 4: Ablation study on mask tokenizer. **Bold** indicates the best result.

Train steps	Total IoU↑	mAHD↓
0 (frozen)	**96.8**	**6.1**
5000	93.0	12.5
10000	93.5	10.6

Table 5: Ablation study on decoding strategy. The IoU threshold of mAHD is set to 0.5.

Decoding strategy	RefCOCO	
	cIoU↑	mAHD↓
Greedy search	**50.9**	**14.4**
Beam search (B=3)	47.3	15.1
Top-K (K=3)	49.4	15.3
Top-P (P=0.9)	50.2	15.3
Random sample	49.2	15.7

Table 6: Ablation study on training data. "Sem." denotes semantic segmentation data and "Ref." denotes referring segmentation data. The IoU threshold of mAHD is set to 0.5.

Pre-training data	Fine-tuning data	ADE20K		refCOCO	
		mIoU↑	mAHD↓	cIoU↑	mAHD↓
SA-OVRS+Ref.	Sem.	**50.5**	**26.6**	-	-
-	Sem.	48.9	28.7	-	-
SA-OVRS+Ref.	Ref.	-	-	**50.9**	**14.4**
-	Ref.	-	-	46.0	16.2

SA-OVRS Contribution Analysis. To assess the contribution of the SA-OVRS dataset to segmentation performance, we train the LlamaSeg-B model with varying combinations of training data and report the results in Table 6. The ablation shows a clear performance drop when pre-training data are omitted, highlighting the importance of large-scale pre-training for capturing rich visual representations. Nevertheless, even without pre-training data, our model still surpasses existing visual generative models of comparable parameter size.

6 Conclusion

In this paper, we propose **LlamaSeg**, a novel visual autoregressive image segmentation method that integrates segmentation into a standard autoregressive framework. We further introduce a two-stage data annotation pipeline to construct **SA-OVRS**, a large-scale open-vocabulary segmentation dataset containing 2M high-quality samples, which provides valuable resources for training and future research. Extensive experiments show that LlamaSeg not only surpasses existing visual generative models of comparable size across multiple benchmarks but also provides a scalable and unified paradigm that can inspire future multimodal and autoregressive segmentation research.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Jiaqi Chen, Jiachen Lu, Xiatian Zhu, and Li Zhang. Generative semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7111–7120, 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16321–16330, 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23592–23601, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024.

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14076–14088, 2024.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26374–26383, 2024.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. *URL https://arxiv.org/abs/2405.09818*, 9.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Haiyang Wang, Hao Tang, Li Jiang, Shaoshuai Shi, Muhammad Ferjad Naeem, Hongsheng Li, Bernt Schiele, and Liwei Wang. Git: Towards generalist vision transformer through universal language interface. In *European Conference on Computer Vision*, pp. 55–73. Springer, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11686–11695, 2022.

Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma Lasagna. Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2(3):6, 2024.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3858–3869, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Jinjin Xu, Liwu Xu, Yuzhe Yang, Xiang Li, Fanyi Wang, Yanchun Xie, Yi-Jie Huang, and Yaqian Li. u-llava: Unifying multi-modal tasks via large language model. In *ECAI 2024*, pp. 618–625. IOS Press, 2024.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18155–18165, 2022.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in Neural Information Processing Systems*, 37:71737–71767, 2024.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.