SALMA: Mask-Guided Structure-Aware Alignment for Referring Segmentation

Anonymous ECCV 2026 Submission

Paper ID #*****

Abstract. Unified multimodal large language models (MLLMs) equipped with segmentation decoders have enabled referring segmentation, yet they often exhibit attention drift under relational or occluded queries, attending to salient distractors instead of the instructed target. We propose SALMA, a lightweight structure-aware attention framework that injects class-agnostic structural priors into cross-modal interaction without adding extra tokens or auxiliary heads. SALMA runs a cheap pre-pass with a frozen SAM-2 decoder to obtain a coarse mask prior, converts it into a soft spatial gate, and modulates cross-attention outputs through a learnable residual scale for robust degradation under imperfect priors. We further enforce semantic and structural alignment using a text–mask contrastive objective and a boundary consistency loss. On Ref-DAVIS17 and Ref-Youtube-VOS, SALMA improves J&F by +3.4 and +1.7 over Sa2VA-1B with only ∼0.7% latency overhead, demonstrating that stronger structural grounding can be achieved without sacrificing efficiency. Code and models will be made publicly available.

Keywords: Referring Segmentation · Multimodal LLMs · Spatial Alignment

1 Introduction

Unified Multimodal Large Language Models (MLLMs) have recently emerged as a powerful paradigm for visual understanding. Frameworks like Sa2VA [35] and LIRA [12] integrate LLM reasoning (e.g., Qwen [1], InternVL [3]) with the pixel-level precision of SAM-2 [22], aiming to enable general-purpose grounding within a single architecture.

However, a critical semantic-structural gap persists. Existing MLLMs often treat visual features as abstract semantic tokens, stripping away the low-level spatial priors essential for boundary delineation. This leads to attention hallucination: as shown in Figure 1, complex spatial queries (e.g., "*the cup behind the laptop*") cause attention to drift toward salient distractors rather than the target defined by spatial constraints. Consequently, performance on structural benchmarks like MeVis [5] often lags behind human capability.

To bridge this gap, we propose SALMA, a unified framework that re-introduces structural fidelity into the attention mechanism. We find that improving per-frame structural grounding yields substantial gains even without introducing additional temporal modules. Unlike methods that rely on explicit user prompts (SegGPT [24]) or implicit alignment (LIRA), our approach injects Mask-Biased Attention priors directly into the transformer layers. By modulating attention weights with low-level cues from the visual backbone, we suppress noise and lock focus onto the target topology.

Our contributions are three-fold:

Fig. 1: Addressing Attention Hallucination. Unlike OMG-LLaVA [37] (Center), SALMA (Right) uses structural priors to precisely ground objects with complex spatial constraints (e.g., depth, ordering), avoiding attention drift.

– **Implicit Structure-Aware Attention.** We introduce Mask-Biased Attention (MBA), which injects structural priors through *implicit attention gating*. Unlike token-based approaches (e.g., GLaMM [21], LISA [10]) that compress spatial information into discrete, hard bottleneck tokens, our MBA mechanism acts as a *dense, soft spatial constraint*. This design preserves the continuous nature of visual features, avoiding the information loss associated with discrete tokenization and allowing for fine-grained modulation directly within the reasoning layers.

– **Dual-Mode Decoder Strategy.** We propose a *shared decoder mechanism* where the same physical SAM-2 decoder serves two roles: frozen prior extraction during the pre-pass and fine-tuned segmentation during decoding. A learnable gating factor γ dynamically weights this structural guidance, ensuring robustness against misleading priors.

– **High Efficiency and Performance.** SALMA achieves 71.9% J&F on Ref-DAVIS 17 [20] (+3.4%) and 78.4 cIoU on RefCOCOg [18], with a negligible 0.7% latency increase. This verifies that precise structural grounding can be achieved with minimal computational overhead.

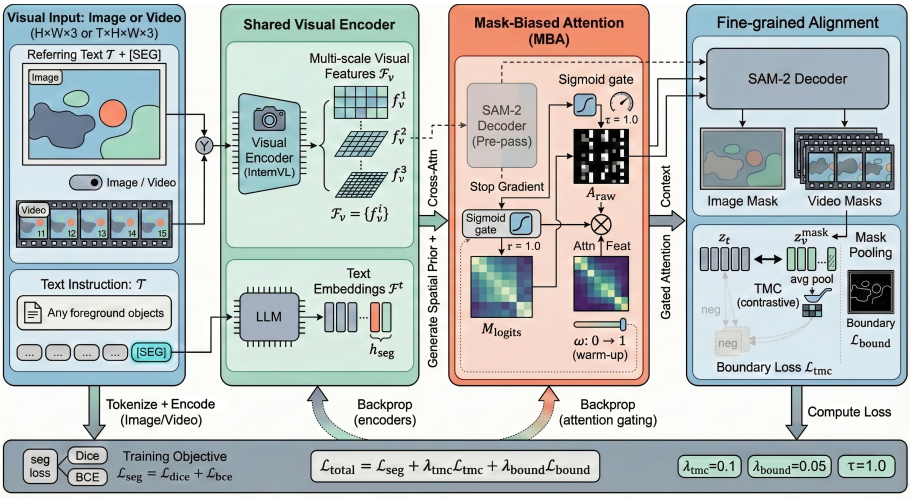

Fig. 2: Framework Overview. (a) Shared Visual Encoder. Multi-scale visual features $\mathcal{F}_v$ and text embeddings are extracted. **(b) Mask-Biased Attention.** Instead of a separate auxiliary head, we perform a pre-pass inference using the shared SAM-2 decoder to generate a prior $\mathcal{M}_{prior}$. This prior is transformed via a Sigmoid into a spatial Gate G, which is multiplied ($\otimes$) with the Cross-Attention output features to suppress noise. **(c) Fine-grained Alignment.** We enforce semantic alignment via TMC Loss ($\leftrightarrow$) and structural precision via Boundary Loss (highlighted contours).

2 Related Work

2.1 Large Multimodal Foundation Models

The advancement of Large Language Models (LLMs) [19] has transformed AI. This progress extended to Multimodal LLMs (MLLMs), aligning visual encoders with LLMs. Integrating visual instruction tuning, pioneering models like LLaVA [15] focused on static image understanding. To handle fine-grained details, subsequent works scaled up resolution [1, 3] and reasoning capabilities [16]. Others adapted these architectures for video understanding [8, 14]. Despite comprehensive benchmarks [6, 11, 17], standard MLLMs often struggle with structure, treating images as abstract tokens.

2.2 Visual Segmentation and Grounding

Prior to the unified MLLM era, referring segmentation was dominated by specialist models. Early approaches relied on modular designs to decouple language and vision. With the advent of Transformers, attention mechanisms became central to multi-modal fusion. Methods like LAVT [32] and CRIS [25] focus on early linguistic injection and contrastive alignment. In the video domain, ReferFormer [27] formulates Referring Video Object Segmentation (RVOS) as a sequence prediction problem. Concurrently, universal segmentation has seen significant progress. Interactive models like SEEM [39] and

SegGPT [24] unify various segmentation tasks. SAM-2 [22] further introduces a promptable streaming memory mechanism. However, these specialist models generally lack the massive reasoning capabilities of LLMs.

2.3 Unified Multimodal Perception

To bridge the gap between semantic reasoning and spatial grounding, recent trends aim to integrate pixel-level decoding directly with MLLMs [10, 12, 21, 23, 29, 35]. To harmonize model optimization, some approaches leverage additional decoders to decode segmentation tokens. LISA [10] utilizes special `[SEG]` tokens to trigger segmentation pixel decoders like SAM. PixelLM [23] replaces SAM with a lightweight pixel decoder and a segmentation codebook. GLaMM [21] incorporates a region-level adapter for simultaneous grounding. PSALM [38] uses mask tokens to unify diverse tasks. Other works focus on fine-grained alignment; for instance, Mask Grounding [4] introduces auxiliary tasks to teach correspondence between masked text and visual objects. Unlike these approaches, our SALMA explicitly injects low-level structural priors via Mask-Biased Attention and enforces alignment through Text-Mask Contrastive constraints, strictly confining the VLM's reasoning to the object of interest.

3 Methodology

The overall pipeline of our proposed SALMA is illustrated in Figure 2 and summarized in Algorithm 1. The framework consists of three integrated stages: (1) **Shared Visual Encoder:** A visual encoder and an LLM process the input video and text to extract multi-scale visual features and text embeddings. (2) **Mask-Biased Attention:** The proposed MBA mechanism injects pixel-level structural priors from the SAM-2 decoder into the cross-modal interaction to prevent attention drift. (3) **Fine-grained Alignment:** Finally, the model is supervised by a dual-constraint strategy involving TMC loss and Boundary Consistency loss to ensure high-fidelity segmentation.

3.1 Problem Formulation

We design SALMA as a structure-first evolution of the unified MLLM paradigm. We adopt the **Sa2VA** [35] architecture as our foundational baseline, as it efficiently integrates the semantic reasoning of MLLMs with the robust segmentation decoding of SAM-2 [22]. This architectural choice provides a strong starting point by combining open-ended generalization with foundational mask quality.

However, we identify a critical structural bottleneck inherent to this architecture. While Sa2VA excels at high-level semantic recognition (e.g., identifying "a dog"), it suffers from varying degrees of degradation when handling fine-grained spatial instructions:

- **Semantic Alignment:** The model effectively aligns text with visual regions for distinct objects. However, the implicit injection of the `[SEG]` token often leads to coarse localization when objects share similar semantic attributes.

– **Structural Hallucination:** As highlighted in Section 1, a severe limitation arises in complex reasoning scenarios (e.g., relative positioning or heavy occlusion). Since Sa2VA treats video frames as abstract 1D semantic tokens, it strips away the low-level 2D/3D spatial geometry required for precise boundary delineation. Without explicit spatial constraints, the cross-modal attention mechanism exhibits attention drift, where the model hallucinates focus on background noise or neighboring distractors rather than the target's topology.

This diagnosis indicates that simply scaling data or parameters cannot resolve the structural deficit. Instead, it necessitates an explicit mechanism to re-introduce spatial priors into the MLLM's reasoning process, motivating our proposed Mask-Biased Attention within the SALMA framework.

3.2 Mask-Biased Attention

Standard cross-attention often introduces noise by allowing global interaction. We propose Mask-Biased Attention (MBA) to strictly gate semantic injection using low-level spatial priors.

Spatial Prior via Auxiliary Execution. To obtain structural guidance without training a separate head, we leverage the *shared* SAM-2 decoder for a lightweight pre-pass. Algorithmically, we supply a **constant** dummy negative point (label -1) at $(0, 0)$ during **both training and inference**. This fixed prompt satisfies the decoder's interface while ensuring the mask generation is driven purely by SAM-2's learned internal saliency prior, independent of specific user guidance. During pre-pass, output logits are detached from the computation graph; during final decoding, decoder weights are fine-tuned end-to-end. This dual-mode usage extracts robust priors without memory overhead.

Temporal Processing. For video, priors are generated independently per frame to prevent error accumulation; SAM-2's internal memory maintains temporal consistency during final decoding.

Visual-Query Attention with Top-K Routing. We invert the standard interaction direction by treating visual features $\mathcal{F}_v \in \mathbb{R}^{HW \times C}$ as Queries and text tokens $\mathcal{F}_t$ as Keys/Values. To enhance computational efficiency and focus on relevant semantics, we employ a Top-K Token Routing strategy. We first compute a scalar importance score s_i for each token via an MLP router, and then select the top-K subset $\hat{\mathcal{F}}_t$ for attention. We set $K = 2$ by default, as referring expressions typically contain at most two key semantic concepts (e.g., "*the red car*" $\rightarrow$ [car, red]). Empirically, $K = 2$ provides the best trade-off between preserving essential semantics and filtering noise tokens (e.g., articles, prepositions):

$$s_i = \mathrm{MLP}(t_i), \quad \hat{\mathcal{F}}_t = \{t_i \mid i \in \mathrm{TopK}(s, K)\} \tag{1}$$

The retrieved semantic features O_{attn} are then computed via standard Multi-Head Attention (MHA) between the visual queries and the filtered text tokens.

Mask-Gated Modulation. Different from implicit bias mechanisms, we explicitly modulate the attention output features using the generated spatial prior, as illustrated in Figure 3. Crucially, this gating is applied to the *output* of the cross-attention block (after

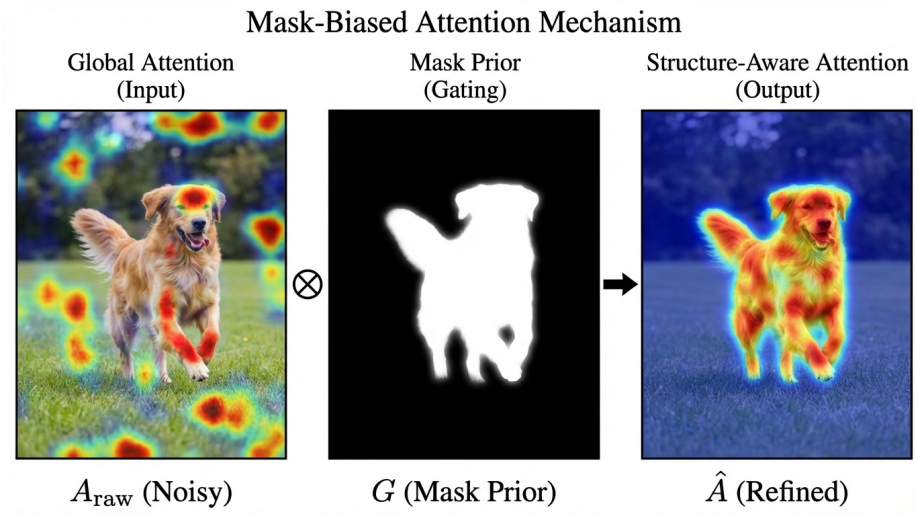

Fig. 3: Mask-Biased Attention Mechanism. The spatial prior G gates the cross-attention output. The injected semantics are strictly confined to the target structure via element-wise multiplication $\otimes$.

the final linear projection), not to the internal attention probability map. This design choice ensures that we modulate the semantic information flow without distorting the probabilistic alignment between tokens.

We first normalize the logits $\mathcal{M}_{logits}$ into a soft spatial gate G using a temperature-scaled Sigmoid. The structure-aware features are then injected into the visual backbone via a gated residual connection:

$$G = \sigma\left(\frac{\mathcal{M}_{logits}}{\tau_{gate}}\right), \quad \mathcal{F}_v' = \mathcal{F}_v + \gamma \cdot (O_{attn} \odot G) \tag{2}$$

where G is resized via bilinear interpolation to match the spatial resolution $H \times W$ of the feature map $\mathcal{F}_v$. O_{attn} is the projected output of the Multi-Head Attention, $\odot$ denotes element-wise multiplication, and γ is a learnable scaling factor. This strategy prevents training instability and enables graceful degradation when priors are unreliable—the model learns to down-weight γ via TMC supervision when the pre-pass mask is erroneous.

Soft Constraint vs. Hard Tokens. A critical distinction of our approach is the nature of the spatial injection. Existing methods like LISA [10] or GLaMM [21] rely on "hard" tokens (e.g., `[SEG]`) to trigger segmentation. This creates a representational bottleneck: the complex spatial intent must be compressed into a single vector, often leading to "all-or-nothing" failures if the token embedding is misaligned. In contrast, MBA imposes a *soft, dense constraint* over the entire visual feature map. The residual connection $\mathcal{F}_v' = \mathcal{F}_v + \dots$ ensures that the original semantic information is never discarded, only refined. If the prior is incorrect, the learnable γ can converge to zero, allowing the model to

revert to standard attention, a robustness property structurally impossible in hard-token architectures.

3.3 Fine-grained Alignment

While the Mask-Biased Attention mechanism successfully injects spatial priors at the feature level, relying solely on this interaction is insufficient for high-precision segmentation. We observe that without explicit supervision signals targeting fine-grained details, the model may still converge to suboptimal solutions where visual features drift towards semantically similar but spatially distinct instances (e.g., merging two adjacent "red cars"). This necessitates a dual-constraint strategy to explicitly enforce alignment between high-level linguistic semantics and pixel-level structural boundaries during the optimization phase. To this end, we introduce a dual-constraint alignment strategy, as depicted in the **Right** panel of Figure 2.

Text-Mask Contrastive Learning. To prevent the visual features from drifting away from the textual instructions during the decoding process, we enforce a semantic consistency constraint. Specifically, we extract the region-level visual feature z_v^{mask} by average-pooling the feature map $\mathcal{F}_v$ over the ground-truth foreground region $\mathcal{M}^{gt}$. We use the ground-truth mask during training to ensure stable semantic alignment, decoupling it from early-stage segmentation errors. Let z_t be the embedding of the special `[SEG]` token corresponding to the referring expression. We employ a Symmetric InfoNCE loss where **negatives are constructed from other text-image pairs within the same batch**, maximizing the mutual information between the matched text-visual pair (z_t, z_v^{mask}):

$$\mathcal{L}_{tmc} = \mathcal{L}_{v\to t} + \mathcal{L}_{t\to v} \tag{3}$$

where both directions (vision-to-text and text-to-vision) are normalized by temperature τ_{tmc}. This symmetric objective explicitly aligns the latent space of the "attended region" with the language instruction, ensuring the model not only looks at the right place but also understands the correct semantics.

Boundary Consistency Constraint. Standard binary cross-entropy (BCE) and Dice losses primarily focus on the overall mask area but are insensitive to boundary errors. In video segmentation, however, fuzzy boundaries are a major source of qualitative degradation (e.g., temporal jitter). We incorporate a Boundary Loss $\mathcal{L}_{bound}$ that measures the discrepancy between predicted and ground-truth edge maps:

$$\mathcal{L}_{bound} = \|\nabla\hat{\mathcal{M}} - \nabla\mathcal{M}^{gt}\|_1 \tag{4}$$

where $\nabla\hat{\mathcal{M}}$ and $\nabla\mathcal{M}^{gt}$ denote the gradient magnitudes of the predicted and ground-truth masks, respectively, computed via Sobel edge detection with 3×3 kernels. We use the standard 3×3 kernel size following common practice in boundary-aware losses; larger kernels (e.g., 5×5) were tested but showed no significant improvement while increasing computational cost. By optimizing this boundary-specific objective, the model is explicitly penalized for edge inaccuracies, acting as a fine-grained sharpener for the output masks.

Algorithm 1 Overall Training Pipeline

Input: Image I, Language Instruction T
Output: Segmentation Mask $\mathcal{M}$
Parameters: Gate Temp $\tau = 1.0$, Routing $K = 2$, loss weights $\lambda_{tmc}, \lambda_{bound}$

Stage 1: Multimodal Feature Extraction
$F_v \leftarrow \text{VisualBackbone}(I)$ ▷ *Extract multi-scale visual features*
$F_t \leftarrow \text{LLM}(T)$ ▷ *Extract text embeddings*

Stage 2: Structural Prior Generation (Pre-pass)
$P_{bg} \leftarrow \{(0,0), \text{label} = -1\}$ ▷ *Background point prompt*
Freeze SAM-2 Decoder parameters ▷ *Stop gradient flow*
$\mathcal{M}_{logits} \leftarrow \text{SAM2Decoder}(F_v, P_{bg})$
$G \leftarrow \sigma(\mathcal{M}_{logits}/\tau)$ ▷ *Generate soft spatial gate*
Unfreeze SAM-2 Decoder

Stage 3: Mask-Biased Attention (MBA)
for each layer l in VLM **do**
  $S \leftarrow \text{MLP}(F_t)$ ▷ *Predict token saliency*
  $\hat{F}_t \leftarrow \text{TopK}(F_t, S, K)$ ▷ *Select top-K relevant tokens*
  $O_{attn} \leftarrow \text{CrossAttn}(Q = F_v, K = \hat{F}_t, V = \hat{F}_t)$
  # Gated Injection: Semantic × Structural Prior
  $F_v \leftarrow F_v + \gamma_l \cdot (O_{attn} \odot G)$
end for

Stage 4: Final Decoding & Optimization
$\hat{\mathcal{M}} \leftarrow \text{SAM2Decoder}(F_v, T)$ ▷ *Text-conditioned decoding*
$\mathcal{L}_{seg} \leftarrow \mathcal{L}_{dice}(\hat{\mathcal{M}}, \mathcal{M}^{gt}) + \mathcal{L}_{bce}(\hat{\mathcal{M}}, \mathcal{M}^{gt})$
$z_v^{mask} \leftarrow \text{Pool}(F_v, \mathcal{M}^{gt})$ ▷ *Target region features*
$\mathcal{L}_{tmc} \leftarrow \text{Contrastive}(z_v^{mask}, F_t)$ ▷ *Semantic Alignment*
$\mathcal{L}_{bound} \leftarrow \text{BoundaryLoss}(\hat{\mathcal{M}}, \mathcal{M}^{gt})$ ▷ *Structural Precision*
$\mathcal{L}_{total} \leftarrow \mathcal{L}_{seg} + \lambda_{tmc}\mathcal{L}_{tmc} + \lambda_{bound}\mathcal{L}_{bound}$
Update model parameters using $\nabla\mathcal{L}_{total}$

3.4 Unified Training Objective

We formulate a unified optimization target that balances global semantic fidelity with local structural precision:

Our total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{tmc}\,\mathcal{L}_{tmc} + \lambda_{bound}\,\mathcal{L}_{bound} \tag{5}$$

where:

- $\mathcal{L}_{seg} = \mathcal{L}_{dice} + \mathcal{L}_{bce}$: This is the standard linear combination of Dice loss and Binary Cross-Entropy loss used in SAM-2, responsible for the general mask quality.
- $\mathcal{L}_{tmc}$: The Text-Mask Contrastive loss enforces semantic fidelity.
- $\mathcal{L}_{bound}$: The Boundary loss enforces high-frequency structural precision.

Hyper-parameter Selection. We set $\lambda_{tmc} = 0.1$ and $\lambda_{bound} = 0.05$ to balance semantic alignment and structural precision. For the gating temperature, we set $\tau_{gate} =$

1.0, which corresponds to applying the standard Sigmoid function without any scaling. This choice directly uses the SAM-2 decoder's output logits as the spatial prior, preserving the decoder's learned confidence distribution.

4 Experiments

4.1 Implementation Details

Architecture Configuration. We adopt the efficient 1B parameter variant of Sa2VA as our architectural backbone, incorporating InternVL2.5-1B [2] as the large language model and the Hiera-Large encoder from SAM-2 as the vision tower. The SAM-2 mask decoder is uniquely repurposed to serve a dual role: it operates in a frozen state for heuristic prior generation during the pre-pass and is fine-tuned for text-conditioned decoding during the final stage.

Training Protocols. The model is trained end-to-end on 4× NVIDIA RTX 5090 GPUs for 1 epoch. We employ the AdamW optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and a weight decay of 0.05. The learning rate is initialized at 4×10^{-5} and follows a cosine decay schedule. We utilize DeepSpeed and BF16 mixed-precision training to maximize efficiency.

Training Data and Splitting Policy. We curate a comprehensive multi-modal instruction tuning dataset consisting of: (1) **Image Referring Segmentation**: RefCOCO/+/g [9, 18] (upsampled 4×) and the GCG collection (grounded samples from GranDf [21] and RefCOCOg). (2) **Video Segmentation**: To ensure temporal structural awareness, we include MeVis (4×), ReVOS [31] (10× to emphasize tracking), Ref-Youtube-VOS [30]/Ref-DAVIS [20] (4×), and the SA-V dataset (4×) from SAM-2. (3) **General Instructions**: LLaVA-v1.5 Mix665k [15], Video-ChatUniVi [7], and Osprey-724k [33] are included to maintain broad reasoning capabilities. Importantly, for all benchmarks evaluated (Ref-DAVIS17, Ref-Youtube-VOS, MeVis), we strictly utilize their official **training splits** during the training phase, ensuring that validation and test splits remain completely unseen to prevent data leakage.

Hyper-parameters & Stabilization. We set the loss weights $\lambda_{tmc} = 0.1$ and $\lambda_{bound} = 0.05$ to balance semantic alignment with boundary precision. The gating temperature τ_{gate} is fixed at 1.0. Notably, to ensure training stability, we apply a **5% linear warm-up** to both the auxiliary loss weights and the learnable gating factor γ. This progressive alignment strategy prevents the strong structural bias from potentially disrupting the pre-trained semantic feature space during early training iterations.

4.2 Evaluation Benchmarks

To rigorously verify SALMA's structure-aware capabilities, we conduct comprehensive evaluations across three complementary domains:

– **Referring Image Segmentation.** We evaluate on the classic **RefCOCO**, **RefCOCO+**, and **RefCOCOg** benchmarks. These datasets test the model's ability to ground static objects from natural language descriptions. RefCOCOg is particularly challenging due to its longer, more complex expressions. We report performance using the standard Cumulative Intersection-over-Union (cIoU) metric.

Table 1: Comparison with State-of-the-Art Methods. We report cIoU for image referring segmentation tasks, $\mathcal{J}\&\mathcal{F}$ for video segmentation tasks, and standard metrics for multimodal understanding. The best results are highlighted in **bold**. SALMA achieves leading performance among unified MLLMs on video segmentation tasks and complex referring scenarios (e.g., RefCOCOg), while maintaining competitive general understanding capabilities.

Method	Image Segmentation								Video Segmentation			Image Chat		
	RefCOCO [9]			RefCOCO+ [9]			RefCOCOg [18]		Ref-DAVIS 17 [20]	Ref-YTVOS [30]	MeVis [5]	MME [6]	MMB [17]	SEED [11]
	Val	TestA	TestB	Val	TestA	TestB	Val	Test						
LLAVA-1.5-13B [15]	-	-	-	-	-	-	-	-	-	-	-	1531(+)	68.8	70.1
Video-LLaVA-7B [14]	-	-	-	-	-	-	-	-	-	-	-	-	60.9	-
LLaMA-VID-7B [13]	-	-	-	-	-	-	-	-	-	-	-	1521(+)	65.1	59.9
mPLUG-Owl3-8B [34]	-	-	-	-	-	-	-	-	-	-	-	-	77.6	-
InternVL2-8B [3]	-	-	-	-	-	-	-	-	-	-	-	-	81.7	76.2
PixelLM-7B [23]	73.0	-	-	66.3	-	-	69.3	-	-	-	-	309/135	17.4	-
PixelLM-13B [23]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	-	-	309/135	17.4	-
LaSagnA [26]	76.8	-	-	66.4	-	-	70.6	-	-	-	-	0/0	0.0	-
LISA-7B [10]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	-	-	-	1/1	0.4	-
LISA-GLEE [28]	76.4	78.2	73.8	67.3	71.3	62.3	71.6	72.4	-	-	-	-	-	-
GLaMM [21]	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	-	-	-	14/9	36.8	-
LLaVA-G-7B [36]	77.1	-	-	68.8	-	-	71.5	-	-	-	-	-	-	-
GSVA [29]	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	-	-	-	-	-	-
OMG-LLaVA-8B [37]	75.6	77.7	71.2	65.6	69.7	58.9	70.7	70.2	-	-	-	1177/235	47.9	56.5
PSALM [38]	**83.6**	**84.7**	**81.6**	72.9	75.5	70.1	73.8	74.4	-	-	-	-	52.5	-
VideoLISA-3.8B [8]	73.8	-	-	63.4	-	-	68.3	-	68.8	63.7	44.4	-	-	-
VISA-13B [31]	72.4	-	-	59.8	-	-	65.5	-	70.4	63.0	44.5	-	-	-
Sa2VA-1B (Base)*	79.6	82.7	76.5	73.6	79.0	66.6	77.8	77.5	68.5	65.3	41.7	**1487/433**	**71.9**	**71.0**
SALMA	80.4	83.2	77.7	**74.8**	**80.0**	**68.9**	**78.0**	**78.4**	**71.9**	**67.0**	**46.0**	1421/353	71.2	69.9

– **Referring Video Segmentation.** To assess structural stability in dynamic scenes, we test on **Ref-DAVIS 2017** and **Ref-Youtube-VOS**. These benchmarks require the model to maintain consistent tracking of a referred object across video frames. We report the $\mathcal{J}\&\mathcal{F}$ mean score, which aggregates region similarity ($\mathcal{J}$) and boundary contour accuracy ($\mathcal{F}$).
– **Motion-Centric Understanding.** We utilize the **MeVis** benchmark, which features queries explicitly dependent on motion cues (e.g., "the fish swimming to the left"). This tests whether our structural priors remain robust under complex temporal dynamics.

4.3 Comparison with Leading MLLMs

We present a comprehensive comparison of our method (SALMA) against leading Multi-Modal Large Language Models (MLLMs). As shown in Table 1, we evaluate performance across three distinct dimensions: image referring segmentation, video segmentation, and general multi-modal understanding.

Quantitative Analysis. As shown in Table 1, our method outperforms baselines across segmentation tasks. On RefCOCOg, we achieve 78.4 cIoU, confirming that $\mathcal{L}_{tmc}$ enhances discrimination. For video, we achieve 71.9 J&F on Ref-DAVIS 17 (+3.4 over baseline) and 67.0 J&F on Ref-Youtube-VOS (+1.7 over baseline), validating our structural priors on standard benchmarks.

Generalization-Specialization Trade-off. We observe modest decreases on general MLLM benchmarks. However, our core task improvements (+3.4% on Ref-DAVIS, +1.7% on Ref-Youtube-VOS) substantially outweigh this degradation. Notably, PSALM [38]

Table 2: Ablation Study of Components. We progressively add Mask-Biased Attention, Text-Mask Contrastive and Boundary Loss.

Configuration	DAVIS (J&F)	RefCOCOg (Val)
Baseline (Sa2VA)	68.52	77.76
+ MBA	69.83	78.38
+ MBA + TMC	70.57	78.15
+ MBA + TMC + Bound	**71.87**	**78.42**

Table 3: Analysis of Feature Modulation. Comparing our simplified spatial gating vs. complex FiLM modulation. Red text indicates a performance drop relative to MBA-only.

Method	Ref-DAVIS17	MeVis (Val_u)
Ours (Full)	**71.87**	**53.37**
Ours (w/ FiLM)	68.13 (-3.74)	47.46 (-5.91)

reports only 52.5 on MMBench versus our 71.2, suggesting our implicit gating preserves more reasoning capacity than explicit token-based approaches. We argue that this is a justifiable **Specialization for Fine-grained Grounding**: in practical agentic applications, precise physical localization is often more critical than generic QA capabilities, making this trade-off highly favorable for grounding-centric tasks.

4.4 Ablation Study

In this section, we analyze the contribution of each component to the final performance. Table 2 shows the progressive improvement of our framework.

- **Baseline:** The vanilla Sa2VA model suffers from global attention noise.
- **+ MBA:** Injecting the mask-biased spatial prior provides the most significant gain on DAVIS, proving that attention gating is specific to structural details.
- **+ TMC & Boundary:** Adding the fine-grained alignment losses further refines the performance, particularly on RefCOCOg, where semantic alignment is crucial.

Impact of Residual FiLM. We investigated a more complex modulation mechanism called Residual FiLM, where feature-wise affine transformations are applied in addition to our spatial gating. As shown in Table 3, FiLM causes significant performance degradation on both benchmarks: DAVIS drops from 71.87 to 68.13 and MeVis drops from 53.37 to 47.46.

We attribute this consistent degradation to FiLM's *over-aggressive feature modulation*. Unlike our lightweight gating mechanism that preserves the original feature distribution while selectively injecting structural priors, FiLM applies a full affine transformation ($\gamma \cdot \mathcal{F} + \beta$) that fundamentally distorts the learned visual representations. This distortion is particularly harmful because: (1) the SAM-2 decoder was pre-trained on features with specific statistical properties, and aggressive modulation breaks this

Table 4: Saliency Bias Analysis on RefCOCOg (Val). We report cIoU scores partitioned by object area ratio. Our method shows consistent improvements on both salient and non-salient objects.

Method	Salient (> 5%)	Non-Salient (≤ 5%)
Baseline (Sa2VA)	82.1	61.6
Ours	**83.4** (+1.3)	**62.4** (+0.8)

assumption; (2) the semantic richness encoded in the MLLM features is corrupted, impairing both structural (DAVIS) and motion-centric (MeVis) reasoning.

This finding validates our design choice of minimal intervention: the gating mechanism should enhance attention focus without distorting the underlying feature space.

Quantitative Saliency Bias Analysis. To address the concern that our prompt-free Mask-Bias might overfit to salient objects, we evaluated performance on the RefCOCOg validation set partitioned by object scale. We define "Salient" (Easy) objects as those occupying $> 5\%$ of the image area, and "Non-Salient" (Hard) as $\leq 5\%$. As shown in Table 4, our method outperforms the baseline on both subsets. Significantly, on the Non-Salient subset, we achieve a consistent gain (+0.8%), demonstrating that the learned residual gating factor γ effectively suppresses the mask prior when it mistakenly highlights the wrong salient object, allowing the semantic branch to dominate.

We acknowledge that our 5% threshold is relatively generous compared to detection standards (e.g., COCO small objects are $< 32^2$ pixels, roughly 0.25%). For true "micro-objects" ($< 1\%$ area), the pre-pass mask effectiveness drops, effectively reverting to the baseline semantic branch. Adoption of stricter thresholds for evaluation is a pertinent future direction.

Inference Efficiency. A potential concern with multi-stage inference approaches is the computational cost. To evaluate this, we benchmarked the inference speed on a standard NVIDIA RTX 5090 GPU environment using the DAVIS validation set (batch size=1). Our method achieves 17.84 FPS, exhibiting a negligible latency increase compared to the Sa2VA-1B baseline (17.97 FPS). This minimal drop (~0.7%) is expected given the architectural asymmetry: the auxiliary SAM-2 decoder (~10M params) is computationally insignificant compared to the massive InternVL backbone (~1B params) and vision encoder (~300M params). Furthermore, because the pre-pass decoder operates on detached features without storing gradients, the additional memory footprint is negligible, validating our design for practical real-time applications.

4.5 Visualization

To intuitively understand the effectiveness of our Structure-Aware framework, we provide qualitative comparisons and attention visualizations.

Qualitative Results. As shown in Figure 4, in the first row, the instruction "*a small carrot...to its left*" requires spatial reasoning; the baseline segments the wrong object while ours succeeds. In the second row, our method produces coherent masks with sharp boundaries compared to the fragmented baseline outputs.

Fig. 4: Qualitative Comparison on Complex Referring Expressions. Top Row: Our model successfully handles complex spatial relationships (e.g., "small", "to its left"), whereas the baseline incorrectly segments the distracting neighbor. **Bottom Row:** In highly cluttered scenes (pile of fries), our method produces masks with superior boundary crispness and completeness compared to the fragmented outputs of the baseline.

(a) Input **(b)** Baseline **(c)** Ours

Fig. 5: Visualization of Cross-Modal Attention. We visualize the attention for the instruction *"Segment the fry on top of the pile"*. (b) Baseline exhibits diffused attention. (c) MBA concentrates focus on the target.

Effect of Mask-Biased Attention. To verify the hypothesis that MBA reduces hallucinations, we visualize the cross-modal attention maps as shown in Figure 5. In the baseline model, the attention weights are often diffuse, activating on irrelevant background areas that share low-level texture similarities with the target. In contrast, our MBA-guided attention exhibits a clear "spotlight" effect. By injecting the spatial mask prior, the attention is strictly confined to the structural extent of the object, verifying the effectiveness of our gating mechanism.

5 Discussion

Mechanistic Interpretation of MBA. Why does Mask-Biased Attention work? We hypothesize that MBA acts as a *soft structural constraint* on the multimodal attention mechanism. In standard MLLMs, the cross-attention layers must simultaneously learn *where* to look (localization) and *what* features to extract (recognition). This dual burden often leads to "attention drift" when semantic cues are ambiguous. By injecting a class-agnostic saliency prior, MBA effectively reduces the spatial search space, allowing the attention heads to focus their capacity on fine-grained semantic matching within the proposed regions. The detailed visualization in Figure 5 supports this: the attention distribution shifts from diffuse "bag-of-words" matching to precise, edges-aligned grounding.

Scalability and Efficiency. A key advantage of our "pre-pass" design is its minimal footprint. Since the SAM-2 decoder is lightweight (~10M parameters) compared to the vision encoder (~300M+) and LLM (~7B+), the computational cost of generating priors is negligible (< 1%). For long-form video, our frame-independent prior generation naturally parallelizes, avoiding the recurrent bottlenecks of temporal modules. While we demonstrated this on 1B models, the gating mechanism is architecture-agnostic and should theoretically scale to larger backbones (e.g., LLaVA-Next-34B) where structural hallucination remains a persistent challenge.

Limitations and Trade-offs. Despite these gains, we acknowledge two limitations. First, *Saliency Dependency*: Our method relies on the initial saliency hypothesis. In scenarios where the target is highly non-salient (e.g., < 1% area) and visually merged with the background, the prior may be misleading. Our learnable γ helps mitigate this but cannot fully eliminate bias in extreme cases. Second, *Generalization Capability*: The slight dip in general benchmarks suggests that strong structural biasing might suppress serendipitous background context useful for open-ended VQA. Future work could explore *instruction-aware dynamic gating* to toggle MBA only when identifying grounding-specific intents.

6 Conclusion

In this paper, we identify and address the "semantic-structural gap" in unified Multimodal Large Language Models, where attention mechanisms often hallucinate in the absence of explicit spatial constraints. To bridge this gap, we propose SALMA, a novel framework that integrates a Mask-Biased Attention mechanism and a fine-grained dual-alignment strategy. By effectively leveraging the low-level mask priors from the shared SAM-2 decoder, we transform the abstract semantic interaction into a structure-guided process, significantly reducing attention drift.

Extensive evaluations demonstrate that our method achieves significant improvements over existing unified baselines on video segmentation benchmarks and complex referring segmentation, while maintaining robust performance on general multimodal tasks. We believe that explicitly modeling the interplay between low-level pixel structure and high-level semantic reasoning is a crucial step towards building truly grounded foundation models. Future work will explore extending this paradigm to larger-scale backbones and addressing the limitation of saliency dependency in highly cluttered scenes.

References

1. Bai, J., et al.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
2. Chen, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
3. Chen, Z., et al.: 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 (CVPR) (2024)
4. Chng, Y.X., Zheng, H., Han, Y., Liu, X., Kankanhalli, M.: Mask grounding for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
5. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: MeViS: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
6. Fu, C., et al.: MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
7. Jin, P., Takanobu, R., Zhang, C., Cao, X., Yuan, L.: Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
8. Jin, S., et al.: One token to seg them all: Language instructed reasoning segmentation in videos. arXiv preprint arXiv:2409.19603 (2024)
9. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referitgame: Referring to objects in photographs of natural scenes. In: EMNLP (2014)
10. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: LISA: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
11. Li, B., et al.: SEED-Bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
12. Li, X., et al.: LIRA: Inferring segmentation in large multi-modal models with local interleaved region assistance. arXiv preprint arXiv:2501.00000 (2025)
13. Li, Y., Wang, C., Wu, J.: Llama-vid: An image-to-video token for video understanding. arXiv preprint arXiv:2311.17043 (2023)
14. Lin, B., et al.: Video-LLaVA: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
15. Liu, H., et al.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
16. Liu, H., et al.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. LLaVA Blog (2024)
17. Liu, Y., et al.: MMBench: Is your multi-modal model an all-around player? In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
18. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR (2016)
19. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
20. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
21. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Fahad, S., Khan, F.S.: GLaMM: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

22. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
23. Ren, Z., Ji, Z., Lan, G., Wang, Z., Cui, Y., Zhai, W., Feng, J.: PixelLM: Pixel reasoning with large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
24. Wang, X., et al.: SegGPT: Segmenting everything in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
25. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: CRIS: CLIP-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
26. Wei, C., Tan, H., Zhong, Y., Yang, Y., Ma, L.: LaSagnA: Language-based segmentation assistant for complex queries. arXiv preprint arXiv:2404.02646 (2024)
27. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Tan, P.: ReferFormer: A simple baseline for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
28. Wu, J., Jiang, Y., Liu, Q., Yuan, Z., Bai, X., Bai, S.: GLEE: General object foundation model for images and videos at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
29. Xia, Z., Han, X., Xue, Y., Zhang, W.: GSVA: Generalized segmentation via multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
30. Xu, N., Yang, L., Fan, Y., Yang, D., Yue, Y., Liang, Y., PRICE, F., Cohen, S., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. In: ECCV (2018)
31. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: VISA: Reasoning video object segmentation via large language models. arXiv preprint arXiv:2407.11325 (2024)
32. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: LAVT: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
33. Yao, Y., Gisiger, T., Peng, Y., et al.: Osprey: Pixel understanding with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
34. Ye, J., et al.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840 (2024)
35. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
36. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., Yang, J.: LLaVA-Grounding: Grounded visual chat with large multimodal models. arXiv preprint arXiv:2312.02949 (2023)
37. Zhang, T., Li, X., Yuan, H., Wan, S., Yang, M.H.: OMG-LLaVA: Bridging image-level, object-level, pixel-level reasoning and understanding. arXiv preprint arXiv:2406.19389 (2024)
38. Zhang, T., Li, X., Yuan, H., Wan, S., Yang, M.H.: PSALM: Pixelwise segmentation with large multi-modal model. arXiv preprint arXiv:2403.14598 (2024)
39. Zou, X., et al.: SEEM: Segment everything everywhere all at once. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)