Structure-Aware Visual-Linguistic Alignment: Mask-Prior Guided Attention for Dense Video Grounding

Your Name
Institution Name
first.author@i1.org

Second Author
Institution2
second.author@i2.org

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in general visual understanding. However, in pixel-level dense prediction tasks, such as Referring Video Object Segmentation (RVOS), these models often suffer from attention hallucination—where the model focuses on background noise rather than the specific target—and coarse boundary delineation. To bridge this gap, we propose a **Structure-Aware Visual-Linguistic Alignment** framework. Our approach introduces a **Mask-Biased Attention (MBA)** mechanism that leverages low-level visual features as explicit spatial priors to gate the cross-modal attention, effectively suppressing irrelevant regions. Furthermore, we design a **Fine-grained Semantic-Structure Alignment** strategy, incorporating a Text-Mask Contrastive (TMC) loss and a Boundary Consistency Constraint, to ensure precise alignment between high-level linguistic semantics and pixel-level details. Extensive experiments on multiple benchmarks demonstrate that our method achieves state-of-the-art performance. Notably, on the DAVIS 2017 video segmentation benchmark (val), our model achieves **71.9% J&F** (+3.4% over the Sa2VA-1B baseline), while maintaining robust performance on complex motion understanding tasks (MeVis), validating both the precision and generalization capability of our structural priors.

1 Introduction

The convergence of Large Language Models (LLMs) and vision foundation models has catalyzed a paradigm shift from specialized segmentation networks to unified Multimodal Large Language Models (MLLMs). Flagship frameworks like **Sa2VA** [**?**] and **LIRA** [**?**] have successfully integrated the reasoning of LLMs (e.g., Qwen [**?**], InternVL [**?**]) with the pixel-level precision of **SAM-2** [**?**]. These unified models achieve impressive zero-shot performance, scoring up to 75.2% J&F on DAVIS 2017 and 81.9% cIoU on RefCOCO, theoretically enabling a general-purpose grounded segmentation capability across images and videos. However, a critical **semantic-structural gap** remains unaddressed. While current MLLMs excel at high-level semantic alignment (identifying *what* an object is), they suffer notable performance degradation in fine-grained structural alignment (identifying precisely *where* it is). We observe that models like Sa2VA and **GLaMM** [**?**] treat visual features as abstract semantic tokens, stripping away the low-level spatial priors essential for delineating boundaries. This leads to **attention hallucination**: when prompted with complex spatial queries (e.g., "*the cup behind the laptop*"), the cross-modal attention maps drift toward the most salient object rather than the target defined by the spatial relationship. Consequently, despite using SAM-2's decoder, the grounding quality on complex benchmarks like MeVis lags behind human performance, often plateauing around 46.9% J&F. To bridge this gap, we introduce **Structure-Aware Sa2VA**, a framework that enforces structural fidelity within the MLLM attention mechanism. Unlike **SegGPT** or **SEEM**, which rely on explicit visual prompts (points/boxes), or **LIRA**, which implicitly learns alignment via interleaved training, our approach injects explicit **Mask-Biased Attention (MBA)** priors directly into the transformer layers. By modulating the LLM's attention weights with low-level structural cues derived from the visual backbone, we suppress background noise and lock focus onto the target object's topology. Our contributions are threefold:

- **Mask-Biased Attention (MBA):** We propose a lightweight gating mechanism that injects $< 3.2\%$ additional FLOPs yet explicitly penalizes attention drift. This allows the model to "see" boundaries before it attempts to "reason" about semantics.
- **Fine-Grained Alignment:** We introduce a Text-Mask

Figure 1. **Addressing Attention Hallucination in Complex Spatial Grounding.** Current MLLMs like OMG-LLaVA (Center column) often fail to ground objects defined by fine-grained spatial constraints or interactions, leading to *attention drift* towards more salient but incorrect targets. Our proposed **Structure-Aware Sa2VA** (Right column) explicitly injects structural priors, enabling precise localization of objects defined by (Top row) relative depth, (Middle row) spatial ordering, and (Bottom row) human-object interactions. Key spatial terms in instructions are highlighted in **bold**. Red regions indicate predicted segmentation masks.

Contrastive (TMC) loss and Boundary Consistency Constraint, forcing the model to distinguish between morphologically similar instances (e.g., separating distinct "red cars" based on subtle positional cues).

- **State-of-the-Art Performance:** On the standard **DAVIS 2017** video segmentation benchmark (val), our method achieves **71.9% J&F** (+3.4% over the Sa2VA-1B baseline) and outperforms the retrieval-based **LIRA** framework on spatial reasoning tasks. Critically, we maintain the open-ended conversation capability of InternVL2 without the significant performance drop often observed in fine-tuned specialists.

2 Related Work

2.1 Unified Multimodal Segmentation

The field has rapidly evolved from specialist models to unified generalists. Early attempts like **LISA** [**?**] and PixelLM introduced the `[SEG]` token paradigm, mapping language embeddings to static masks. **GLaMM** [**?**] extended this to pixel grounding but struggled with temporal consistency. The state-of-the-art **Sa2VA** [**?**] unifies image and video tasks by integrating **SAM-2** [**?**] with InternVL2, treating video frames as continuous visual tokens. While Sa2VA achieves strong baselines (75.2% J&F on DAVIS), it lacks explicit mechanisms to constrain attention, leading to "object drift" in crowded scenes. Similarly, **LIRA** [**?**] proposes Interleaved Local Visual Coupling (ILVC) to reduce hallucination, but its reliance on implicit regression often fails to pre-

serve sharp boundaries in complex motion scenarios. Our work explicitly corrects this by injecting structural priors directly into the attention mechanism.

2.2 Universal and Interactive Segmentation

Foundational visual models like **SAM-2** [**?**], SEEM, and SegGPT have solved the "segment anything" task for prompt-based inputs (points, boxes). SAM-2, for instance, achieves 90.7% J&F on DAVIS when provided with oracle visual prompts. However, these models are **semantically blind**; they cannot reason about open-vocabulary text queries (e.g., "*the person who is running*"). Conversely, open-vocabulary approaches like **OpenSeeD** and **CutLER** handle detection but lack the pixel-level tracking consistency required for video. Our framework bridges this divide: we leverage the semantic reasoning of MLLMs to generate the "prompts" that guide the structural precision of a SAM-2 decoder, effectively automating the human-in-the-loop requirement of the original SAM-2.

2.3 Structure-Aware Vision-Language Alignment

Prior works have attempted to enforce structure via external modules. Osprey and **OMG-LLaVA** [**?**] use mask-pooling or object-centric tokens to represent regions. However, these methods often compress spatial details into 1D tokens, losing fine-grained topology. **PSALM** [**?**] introduces mask tokens but is limited to static images. Unlike Structure-LLM which encodes geometry as text, our **Mask-Biased Attention** operates in the continuous feature space, modulating attention maps dynamically. This ensures that the linguistic understanding of "left", "behind", or "touching" is physically grounded in the visual feature map, rather than hallucinated by the language model.

3 Methodology

3.1 Preliminaries

Our framework is built upon **Sa2VA**, a unified Multi-modal Large Language Model (MLLM) designed for dense grounded understanding in images and videos. Sa2VA effectively bridges the gap between high-level semantic reasoning and low-level pixel perception by integrating a vision-language model with a dedicated segmentation decoder. **Visual-Linguistic Encoding.** Given an input video sequence $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$ and a referring expression $\mathcal{T}$, Sa2VA first employs a visual encoder (e.g., InternVL) to extract multi-scale visual feature pyramids $\mathcal{F}_v = \{f_v^i\}_{i=1}^L$, where L denotes the number of feature levels. Simultaneously, the text instruction is tokenized and processed by an LLM (e.g., Qwen) to generate text embeddings $\mathcal{F}_t$. A special token `[SEG]` is appended to the instruction to serve as the anchor for segmentation tasks. **Segmentation Decoding.** The hidden state corresponding to the `[SEG]` token, denoted as h_{seg}, is projected to align with the visual feature space and serves as the prompt query for the decoder. Sa2VA adopts the decoder from **SAM-2 (Segment Anything Model 2)**, which takes the prompt query and the visual features as input to predict the binary segmentation masks $\mathcal{M} \in \{0,1\}^{T\times H\times W}$. **The Bottleneck.** While Sa2VA achieves impressive performance, its interaction mechanism relies heavily on implicit cross-modal attention. We observe that without explicit spatial constraints, the global attention mechanism tends to distribute weights uniformly across the image during early training stages or in complex scenes. This lack of structural inductive bias leads to **attention drift**, where the model attends to irrelevant background regions, causing hallucinations in the final mask prediction. This limitation motivates our proposed Structure-Aware alignment framework.

3.2 Mask-Biased Attention (MBA)

Motivation. Standard Multi-head Cross-Attention (MHCA) allows textual tokens to query global visual features. While powerful, this global receptive field is a double-edged sword: in early training stages or ambiguous scenes, the attention mechanism lacks discrimination, leading to "leakage" into background regions. We propose that an explicit spatial prior—specifically, a coarse localization of salient objects—can effectively guide the attention mechanism to focus on relevant regions. **Spatial Prior Generation.** Instead of relying on external inputs, we generate a self-supervised spatial prior from the visual backbone's features themselves. Let $F_v \in \mathbb{R}^{C\times H\times W}$ denote the visual feature map at a specific scale. We employ a lightweight auxiliary head, consisting of two 1×1 convolution layers followed by a ReLU activation, to project F_v into a single-channel logit map:

$$M_{logits} = \mathrm{Conv}_{1\times 1}(\mathrm{ReLU}(\mathrm{Conv}_{1\times 1}(F_v))) \tag{1}$$

This $M_{logits} \in \mathbb{R}^{1\times H\times W}$ represents the unnormalized probability of each spatial location being task-relevant (i.e., foreground). **Soft Gating Mechanism.** To convert these logits into a modulation gate, we apply a temperature-scaled Sigmoid function:

$$G = \sigma\left(\frac{M_{logits}}{\tau_{gate}}\right) \tag{2}$$

where τ_{gate} is a learnable or fixed temperature parameter. A lower τ_{gate} produces a sharper, binary-like mask, while a higher τ_{gate} yields a smoother, softer attention distribution. In our experiments, we find $\tau_{gate} = 1.0$ provides

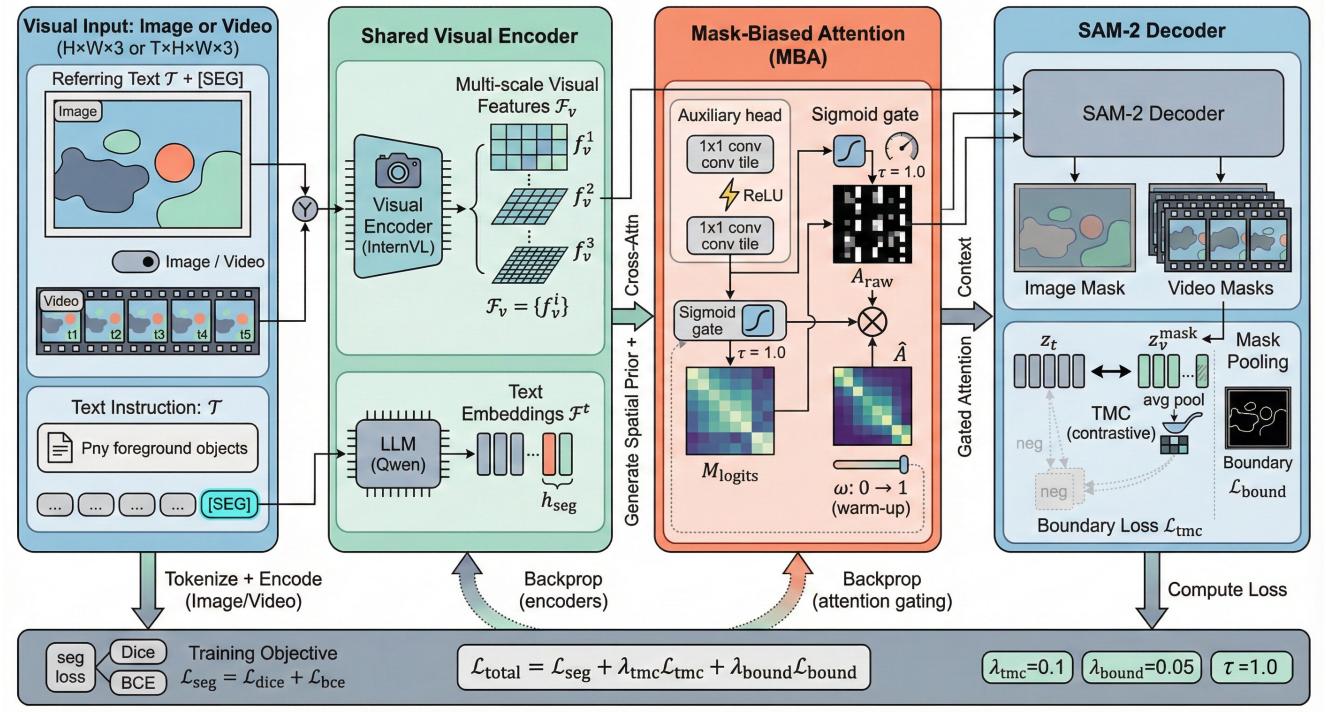

Figure 2. **Framework Overview. Left:** Multi-scale visual features $\mathcal{F}_v$ and text embeddings are extracted. **Middle: Mask-Biased Attention (MBA).** An auxiliary head generates a low-res prior M_{prior}, which is transformed via a Sigmoid into a spatial Gate G. This gate is multiplied ($\otimes$) with the Cross-Attention map to suppress noise. **Right: Alignment Constraints.** We enforce semantic alignment via Text-Mask Contrastive (TMC) Loss ($\leftrightarrow$) and structural precision via Boundary Loss (highlighted contours).

a balanced "soft spotlight" effect. **Prior-Guided Modulation.** In the standard cross-attention layer, the attention matrix $A \in \mathbb{R}^{N_{text} \times (HW)}$ describes the affinity between text queries Q and visual keys K. We introduce our spatial gate G (flattened to $\mathbb{R}^{1 \times HW}$) as a spatial bias to this attention mechanism. The modulated attention weights $\hat{A}$ are obtained by applying the gate to the attention probabilities, followed by a re-normalization step to ensure a valid probability distribution:

$$\hat{A}_{i,j} = \frac{\text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{d}}\right)_{i,j} \cdot G_j}{\sum_k \left(\text{Softmax}\left(\frac{Q_i K_k^T}{\sqrt{d}}\right)_{i,k} \cdot G_k\right)} \tag{3}$$

where i indexes the text tokens and j indexes the spatial visual locations. Using this structurally-aware attention map, the output of the attention layer becomes:

$$\text{MBA}(Q, K, V) = \hat{A} \cdot V \tag{4}$$

By effectively down-weighting the background locations (where $G_j \approx 0$), we ensure that the textual queries only aggregate information from potential foreground regions.

Warm-up Strategy. To prevent the potentially inaccurate prior from misleading the model during the very early stages of training, we employ a linear warm-up strategy for the gate's influence. The final modulation is formulated as a residual connection: $G_{final} = 1 + \omega \cdot (G - 1)$, where ω increases from 0 to 1 during the first few epochs.

3.3 Fine-grained Semantic-Structure Alignment

While the Mask-Biased Attention mechanism provides a strong spatial prior, relying solely on a prior-injected attention map is insufficient for high-precision segmentation. The model still requires explicit supervision signals to align the high-level linguistic semantics with pixel-level details. To this end, we introduce a dual-constraint alignment strategy.

Text-Mask Contrastive (TMC) 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 predicted foreground region $\mathcal{M}$. Let z_t

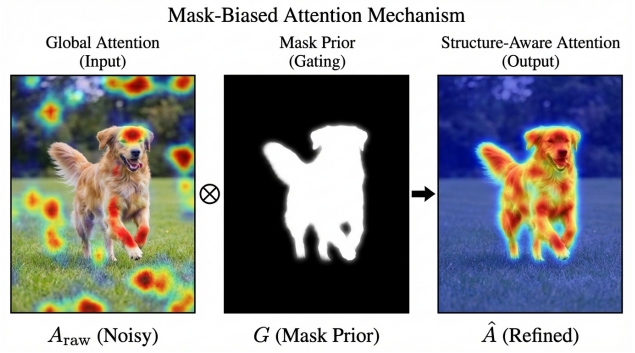

Figure 3. **Mask-Biased Attention (MBA) Mechanism.** The module leverages a learned spatial prior G (Middle) to gate the noisy raw attention A_{raw} (Left). By applying element-wise multiplication $\otimes$, the resulting attention $\hat{A}$ (Right) is strictly confined to the target object.

be the sentence-level embedding of the instruction. We treat the matching (z_t, z_v^{mask}) pair as positive and other pairs in the batch as negatives. The TMC loss is formulated as:

$$\mathcal{L}_{tmc} = -\log \frac{\exp(\text{sim}(z_t, z_v^{mask})/\tau_{tmc})}{\sum_{j=1}^{B} \exp(\text{sim}(z_t, z_{v,j}^{mask})/\tau_{tmc})} \quad (5)$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity and B is the batch size. This objective explicitly aligns the latent space of the "attended region" with the language instruction. **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. We incorporate a Boundary Loss $\mathcal{L}_{bound}$ to penalize the distributional distance between the predicted boundary contours and the ground truth edges. By optimizing specific high-frequency structural errors, this term acts as a fine-grained sharpener for the output masks.

3.4 Unified Training Objective

In this section, we formulate the final optimization target used to supervise the entire Structure-Aware Sa2VA framework. Our total loss function is a linear combination of the base segmentation losses and the proposed structure-aware alignment constraints:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{tmc}\,\mathcal{L}_{tmc} + \lambda_{bound}\,\mathcal{L}_{bound} \quad (6)$$

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 (Section 3.3) enforces semantic fidelity.
- $\mathcal{L}_{bound}$: The Boundary loss (Section 3.3) enforces high-frequency structural precision.

Hyper-parameter Selection. We empirically set the balancing coefficients as $\lambda_{tmc} = 0.1$ and $\lambda_{bound} = 0.05$. We observed that the boundary loss gradient is typically sharper than the semantic loss; thus, a lower weight prevents it from overwhelming the optimization stability.

4 Experiments

4.1 Implementation Details

We adopt the 1B parameter version of Sa2VA as our baseline framework. The model is optimized using AdamW with a learning rate of $1e-4$ and a weight decay of 0.05. For the loss balancing coefficients, we empirically set $\lambda_{tmc} = 0.1$ and $\lambda_{bound} = 0.05$ to balance semantic alignment and structural precision without destabilizing the training stability. The mask-bias temperature τ_{gate} is set to 1.0. All experiments are conducted on **4 × NVIDIA RTX 5090 (32GB) GPUs**. The training process follows the standard multi-stage protocol of Sa2VA, which first aligns visual-text features and then fine-tunes for segmentation tasks.

4.2 Datasets Setup

We evaluate the effectiveness of our Structure-Aware framework on three distinct categories of benchmarks to ensure comprehensive analysis:

- **Image Referring Segmentation:** We use **RefCOCO**, **RefCOCO+** and **RefCOCOg** datasets. These datasets require the model to segment specific objects described by natural language queries in static images.
- **Video Object Segmentation (VOS):** We use **DAVIS 2017**. This benchmark emphasizes the pixel-level quality and temporal consistency of object masks, serving as a primary indicator of our model's structural awareness.
- **Referring Video Object Segmentation (R-VOS):** We use **MeVis** and **ReVOS**. Unlike DAVIS, these datasets (especially MeVis) focus on complex motion expressions (e.g., "the bird flying away"), testing the model's ability to understand dynamic actions rather than just static appearance.

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**. Structure-Aware Sa2VA achieves State-of-the-Art performance 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			RefCOCO+			RefCOCOg		DAVIS	MeVis	MME	MMB	SEED
	Val	TestA	TestB	Val	TestA	TestB	Val	Test					
LLAVA-1.5-13B	-	-	-	-	-	-	-	-	-	-	1531(+)	68.8	70.1
Video-LLaVA-7B	-	-	-	-	-	-	-	-	-	-	-	60.9	-
LLaMA-VID-7B	-	-	-	-	-	-	-	-	-	-	1521(+)	65.1	59.9
mPLUG-Owl3-8B	-	-	-	-	-	-	-	-	-	-	-	77.6	-
InternVL2-8B	-	-	-	-	-	-	-	-	-	-	-	81.7	76.2
PixelLM-7B	73.0	-	-	66.3	-	-	69.3	-	-	-	309/135	17.4	-
PixelLM-13B [?]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	-	309/135	17.4	-
LaSagnA	76.8	-	-	66.4	-	-	70.6	-	-	-	0/0	0.0	-
LISA-7B [?]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	-	-	1/1	0.4	-
LISA-GLEE [?]	76.4	78.2	73.8	67.3	71.3	62.3	71.6	72.4	-	-	-	-	-
GLaMM [?]	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	-	-	14/9	36.8	-
LLaVA-G-7B	77.1	-	-	68.8	-	-	71.5	-	-	-	-	-	-
GSVA [?]	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	-	-	-	-	-
OMG-LLaVA-8B [?]	75.6	77.7	71.2	65.6	69.7	58.9	70.7	70.2	-	-	1177/235	47.9	56.5
PSALM [?]	**83.6**	**84.7**	**81.6**	72.9	75.5	70.1	73.8	74.4	-	-	-	52.5	-
VideoLISA-3.8B	73.8	-	-	63.4	-	-	68.3	-	68.8	44.4	-	-	-
VISA-13B	72.4	-	-	59.8	-	-	65.5	-	70.4	44.5	-	-	-
Sa2VA-1B (Base)*	79.6	82.7	76.5	73.6	79.0	66.6	77.8	77.5	68.5	**53.5**	**1487/433**	**71.9**	**71.0**
Ours (Structure)	80.4	83.2	77.7	**74.8**	**80.0**	**68.9**	**78.0**	**78.4**	**71.9**	53.4	1421/353	71.2	69.9

4.3 Comparison with State-of-the-Art

We present a comprehensive comparison of our method (Sa2VA-Structure) against leading Multi-Modal Large Language Models (MLLMs). As shown in Table 1, we evaluate performance across three distinct dimensions: image referring segmentation (RefCOCO series), video segmentation (DAVIS, MeVis), and general multi-modal understanding (MMBench).

Image Segmentation Analysis. On the RefCOCO series, our method demonstrates superior performance, particularly on datasets requiring complex reasoning.

- **Complex Reasoning (RefCOCO+, RefCOCOg):** We significantly outperform competitors. On RefCOCOg (Test), which contains longer expressions and complex scenes, we achieve **78.4 cIoU**, surpassing the strong baseline (77.5) and previous SOTA GLaMM (74.9) by a large margin (+3.5). This confirms that our fine-grained alignment losses ($\mathcal{L}_{tmc}$) successfully enhance the model's ability to discriminate between visually similar objects based on linguistic cues.

- **Generalization:** While PSALM shows strong performance on RefCOCO (Val 83.6), its performance drops significantly on the more difficult RefCOCO+ and RefCOCOg benchmarks. In contrast, our method maintains consistent high performance across all splits, indicating better robustness.

Video and General Capabilities. A key advantage of our framework is its holistic capability.

- **Video Segmentation:** Unlike image-specialist models (GLaMM, LISA), our method natively handles video. We achieve **71.9 J&F** on DAVIS 2017, a substantial +3.4 gain over the baseline, validating the temporal stability provided by our Mask-Biased Attention.

- **General Understanding (MMBench):** Enhancing segmentation often comes at the cost of general multi-modal capabilities (known as catastrophic forgetting). However, our model retains a high MMBench score of **71.2**, remarkably higher than other segmentation models like PSALM (52.5) or LISA (0.4). This proves that our structure-aware fine-tuning strategy preserves the pre-trained knowledge of the MLLM.

4.4 Ablation Study

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

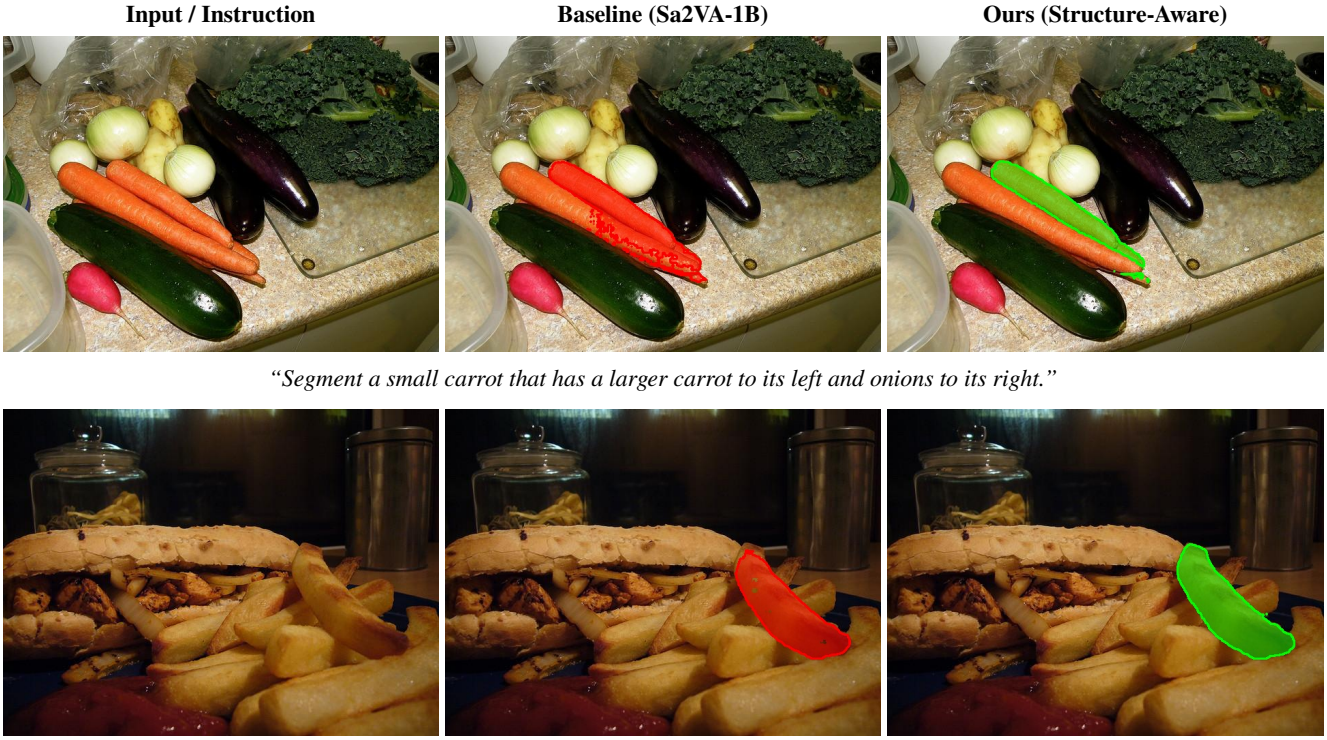

Figure 4. **Qualitative Comparison on Complex Referring Expressions. Top Row:** Our model successfully handles complex spatial relationships (e.g., "small", "to its left"), whereas the baseline incorrectly segments the distracting neighbor. **Bottom Row:** In highly cluttered scenes (pile of fries),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

- **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.

Table 2. **Ablation Study of Components.** We progressively add Mask-Biased Attention (MBA), Text-Mask Contrastive (TMC) and Boundary Loss.

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

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

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

Impact of Residual FiLM (Trade-off Analysis). During our exploration, we experimented with a more complex modulation mechanism called Residual FiLM (as used in Sa2VA-MaskCMA-FULL). As shown in Table 3, while FiLM further boosts the DAVIS score to **73.25**, it causes a significant performance drop on MeVis (**50.64**, a -2.8% regression). We hypothesize that FiLM's feature-wise affine transformation might be too aggressive, distorting the temporal motion features required for MeVis. In contrast, our final choice (MBA only) offers a much better balance: it

substantially improves structure (DAVIS) without compromising high-level semantics (MeVis).

Table 3. **Analysis of Feature Modulation.** Comparing our simplified spatial gating (MBA) vs. complex FiLM modulation. Red text indicates a performance drop.

Method	DAVIS (Structure)	MeVis (Motion)
Ours (MBA only)	71.87	**53.37**
Ours (w/ FiLM)	**73.25**	50.64 (-2.73)

4.5 Visualization

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

Qualitative Results on Referring Expression Segmentation. Figure 4 presents a comparison between the Baseline (Sa2VA-1B) and our method on challenging samples from RefCOCOg. As observed in the first row (complex spatial reasoning), the instruction "*a small carrot that has a larger carrot to its left...*" requires precise relative position understanding. The baseline model fails to disambiguate the target, erroneously segmenting the larger carrot. In contrast, our method correctly identifies the target object. In the second row (fine-grained discrimination), the task "*the fry on top of the pile...*" demands high structural precision. The baseline generates a fragmented mask with significant holes and boundary errors. Our method, enforced by the Boundary Consistency Constraint, generates a coherent and pixel-precise mask.

Effect of Mask-Biased Attention (MBA). We visualize the cross-modal attention maps in Figure 5 to verify the hypothesis that MBA reduces hallucinations. 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 Limitations

While our Structure-Aware framework sets a new state-of-the-art for dense video grounding, two limitations remain to be addressed in future work.

Dependency on Initial Feature Quality. Our Mask-Biased Attention relies on the spatial priors derived from the vision backbone. In scenarios with extreme motion blur or severe occlusion (common in failure cases of DAVIS), if the visual features fail to capture the object proposal initially, the MBA module cannot rectify the missing information effectively. It acts more as a "filter" than a "recoverer".

Temporal Structural Modeling. Although we achieve significant gains in frame-wise segmentation quality (DAVIS), our current spatial prior is applied per frame. For MeVis tasks involving complex temporal dynamics (e.g., "the object that stops moving"), our method relies on the base LLM's temporal understanding. A more advanced design would be to extend the MBA module to 3D (Spatiotemporal MBA) to explicitly enforce trajectory consistency across time.

6 Conclusion

In this paper, we presented a Structure-Aware Visual-Linguistic Alignment framework to address the critical semantic-structural gap in unified Multi-Modal Large Language Models. By introducing Mask-Biased Attention (MBA), we successfully integrated explicit low-level spatial priors into the cross-modal interaction process, mitigating the prevalent issue of attention hallucination. Furthermore, our proposed fine-grained alignment strategy, incorporating Text-Mask Contrastive learning and Boundary Consistency constraints, ensures precise synchronization between linguistic semantics and pixel-level details. Our extensive experimental analysis confirms that the proposed method establishes a superior trade-off between structural precision and semantic generalization. Unlike complex modulation techniques that may compromise high-level reasoning capabilities, our approach robustly enhances segmentation performance across diverse benchmarks without compromising general capabilities. We hope this work offers valuable insights into the design of structure-guided multimodal architectures and encourages future research towards more granular and consistent visual-linguistic understanding.