

001 **SALMA: Mask-Guided Structure-Aware Alignment for** 001
002 **Referring Segmentation** 002

003 Anonymous ECCV 2026 Submission 003

004 Paper ID #***** 004

005 **Abstract.** Unified multimodal large language models (MLLMs) equipped with 005
006 segmentation decoders have enabled referring segmentation, yet they often exhibit 006
007 attention drift under relational or occluded queries, attending to salient distrac- 007
008 tors instead of the instructed target. We propose SALMA, a lightweight structure- 008
009 aware attention framework that injects class-agnostic structural priors into cross- 009
010 modal interaction without adding extra tokens or auxiliary heads. SALMA runs a 010
011 cheap pre-pass with a frozen SAM-2 decoder to obtain a coarse mask prior, con- 011
012 verts it into a soft spatial gate, and modulates cross-attention outputs through a 012
013 learnable residual scale for robust degradation under imperfect priors. We further 013
014 enforce semantic and structural alignment using a text-mask contrastive objec- 014
015 tive and a boundary consistency loss. On Ref-DAVIS17 and Ref-Youtube-VOS, 015
016 SALMA improves J&F by +3.4 and +1.7 over Sa2VA-1B with only ~0.7% la- 016
017 tency overhead, demonstrating that stronger structural grounding can be achieved 017
018 without sacrificing efficiency. Code and models will be made publicly available. 018

019 **Keywords:** Referring Segmentation · Multimodal LLMs · Spatial Alignment 019

020 **1 Introduction** 020

021 Unified Multimodal Large Language Models (MLLMs) have recently emerged as a pow- 021
022 erful paradigm for visual understanding. Frameworks like Sa2VA [35] and LIRA [12] 022
023 integrate LLM reasoning (e.g., Qwen [1], InternVL [3]) with the pixel-level precision of 023
024 SAM-2 [22], aiming to enable general-purpose grounding within a single architecture. 024

025 However, a critical semantic-structural gap persists. Existing MLLMs often treat 025
026 visual features as abstract semantic tokens, stripping away the low-level spatial priors 026
027 essential for boundary delineation. This leads to attention hallucination: as shown in 027
028 Figure 1, complex spatial queries (e.g., “*the cup behind the laptop*”) cause attention 028
029 to drift toward salient distractors rather than the target defined by spatial constraints. 029
030 Consequently, performance on structural benchmarks like MeVis [5] often lags behind 030
031 human capability. 031

032 To bridge this gap, we propose SALMA, a unified framework that re-introduces 032
033 structural fidelity into the attention mechanism. We find that improving per-frame struc- 033
034 tural grounding yields substantial gains even without introducing additional temporal 034
035 modules. Unlike methods that rely on explicit user prompts (SegGPT [24]) or implicit 035
036 alignment (LIRA), our approach injects Mask-Biased Attention priors directly into the 036
037 transformer layers. By modulating attention weights with low-level cues from the visual 037
038 backbone, we suppress noise and lock focus onto the target topology. 038

039 Our contributions are three-fold:

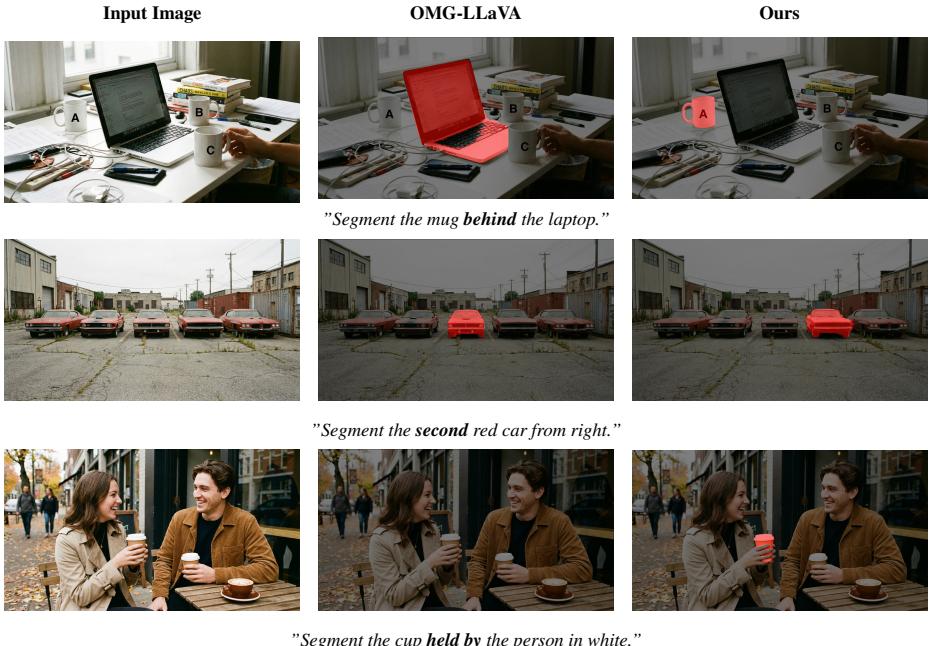


Fig. 1: Addressing Attention Hallucination. Unlike OMG-LLaVA [37] (Center), SALMA (Right) uses structural priors to precisely ground objects with complex spatial constraints (e.g., depth, ordering), avoiding attention drift.

- 040 – **Implicit Structure-Aware Attention.** We introduce Mask-Biased Attention (MBA),
 041 which injects structural priors through *implicit attention gating*. Unlike token-based
 042 approaches (e.g., GLaMM [21], LISA [10]) that compress spatial information into
 043 discrete, hard bottleneck tokens, our MBA mechanism acts as a *dense, soft spatial
 044 constraint*. This design preserves the continuous nature of visual features, avoiding
 045 the information loss associated with discrete tokenization and allowing for fine-
 046 grained modulation directly within the reasoning layers.
- 047 – **Dual-Mode Decoder Strategy.** We propose a *shared decoder mechanism* where
 048 the same physical SAM-2 decoder serves two roles: frozen prior extraction during
 049 the pre-pass and fine-tuned segmentation during decoding. A learnable gating
 050 factor γ dynamically weights this structural guidance, ensuring robustness against
 051 misleading priors.
- 052 – **High Efficiency and Performance.** SALMA achieves 71.9% J&F on Ref-DAVIS
 053 17 [20] (+3.4%) and 78.4 cIoU on RefCOCOg [18], with a negligible 0.7% latency
 054 increase. This verifies that precise structural grounding can be achieved with mini-
 055 mal computational overhead.

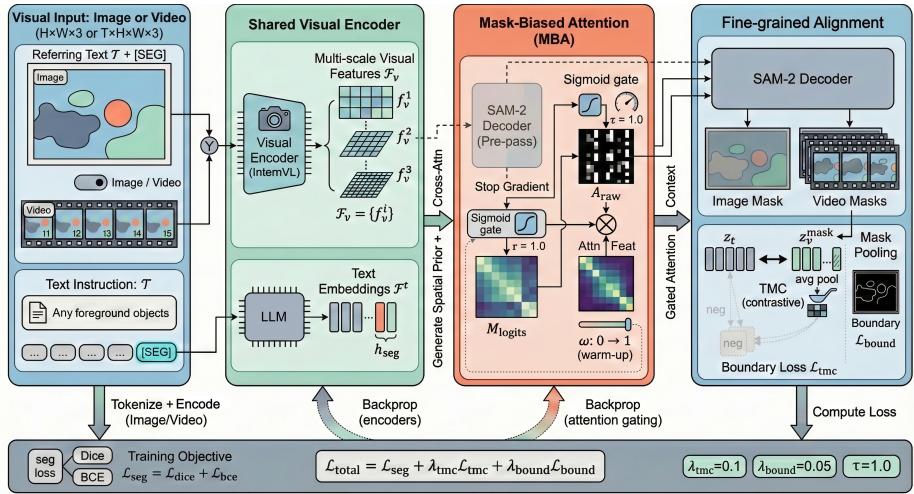


Fig. 2: Framework Overview. (a) **Shared Visual Encoder.** Multi-scale visual features \mathcal{F}_v and text embeddings are extracted. (b) **Mask-Biased Attention.** Instead of a separate auxiliary head, we perform a pre-pass inference using the shared SAM-2 decoder to generate a prior M_{prior} . This 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

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

073
074
075

076 2.3 Unified Multimodal Perception

076

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

077
078
079
080
081
082
083
084
085
086
087
088

089 3 Methodology

089

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

090
091
092
093
094
095
096
097

098 3.1 Problem Formulation

098

099 We design SALMA as a structure-first evolution of the unified MLLM paradigm. We
 100 adopt the **Sa2VA** [35] architecture as our foundational baseline, as it efficiently integrates
 101 the semantic reasoning of MLLMs with the robust segmentation decoding of SAM-
 102 2 [22]. This architectural choice provides a strong starting point by combining open-
 103 ended generalization with foundational mask quality.

099
100
101
102
103

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

104
105
106
107

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

108
109
110

– **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 (Section 3.3), 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*” → [car, red]). Empirically, $K = 2$ provides the best trade-off between preserving essential semantics and filtering noise tokens (e.g., articles, prepositions):

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

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

Mask-Biased Attention Mechanism

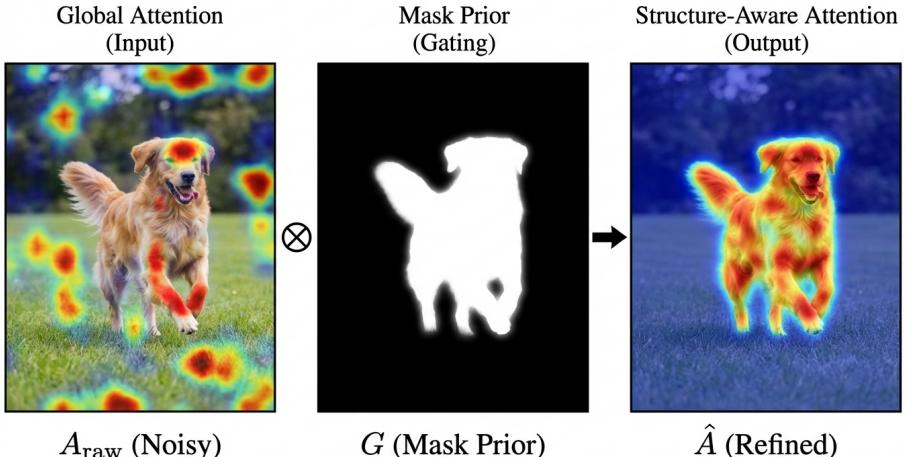


Fig. 3: Mask-Biased Attention Mechanism. The spatial prior G gates the cross-attention output. The injected semantics are strictly confined to the target structure via element-wise multiplication \otimes .

Figure 3. Crucially, this gating is applied to the *output* of the cross-attention block (after 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 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:

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

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

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

174
175

176 3.3 Fine-grained Alignment

176

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

177
178
179
180
181
182
183
184
185

186 **Text-Mask Contrastive Learning.** To prevent the visual features from drifting away
 187 from the textual instructions during the decoding process, we enforce a semantic con-
 188 sistency constraint. Specifically, we extract the region-level visual feature z_v^{mask} by
 189 average-pooling the feature map \mathcal{F}_v over the ground-truth foreground region \mathcal{M}^{gt} . We
 190 use the ground-truth mask during training to ensure stable semantic alignment, decou-
 191pling it from early-stage segmentation errors. Let z_t be the embedding of the special
 192 [SEG] token corresponding to the referring expression. We employ a Symmetric In-
 193 foNCE loss where **negatives are constructed from other text-image pairs within the**
 194 **same batch**, maximizing the mutual information between the matched text-visual pair
 195 (z_t, z_v^{mask}):

$$196 \quad \mathcal{L}_{tmc} = \mathcal{L}_{v \rightarrow t} + \mathcal{L}_{t \rightarrow v} \quad (3) \quad 196$$

196

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

197
198
199
200

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

$$206 \quad \mathcal{L}_{bound} = \|\nabla \hat{\mathcal{M}} - \nabla \mathcal{M}^{gt}\|_1 \quad (4) \quad 206$$

206

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

207
208
209
210
211
212
213

Algorithm 1 Overall Training Pipeline

-
- 1: **Input:** Image I , Language Instruction T
 2: **Output:** Segmentation Mask \mathcal{M}
 3: **Parameters:** Gate Temp $\tau = 1.0$, Routing $K = 2$, loss weights $\lambda_{tmc}, \lambda_{bound}$
- 4: **Stage 1: Multimodal Feature Extraction**
- 5: $F_v \leftarrow \text{VisualBackbone}(I)$ ▷ Extract multi-scale visual features
 6: $F_t \leftarrow \text{LLM}(T)$ ▷ Extract text embeddings
- 7: **Stage 2: Structural Prior Generation (Pre-pass)**
- 8: $P_{bg} \leftarrow \{(0, 0), \text{label} = -1\}$ ▷ Background point prompt
 9: **Freeze** SAM-2 Decoder parameters ▷ Stop gradient flow
- 10: $\mathcal{M}_{logits} \leftarrow \text{SAM2Decoder}(F_v, P_{bg})$
 11: $G \leftarrow \sigma(\mathcal{M}_{logits}/\tau)$ ▷ Generate soft spatial gate
- 12: **Unfreeze** SAM-2 Decoder
- 13: **Stage 3: Mask-Biased Attention (MBA)**
- 14: **for** each layer l in VLM **do**
- 15: $S \leftarrow \text{MLP}(F_l)$ ▷ Predict token saliency
 16: $\hat{F}_l \leftarrow \text{TopK}(F_l, S, K)$ ▷ Select top-K relevant tokens
- 17: $O_{attn} \leftarrow \text{CrossAttn}(Q = F_v, K = \hat{F}_l, V = \hat{F}_l)$
 18: # Gated Injection: Semantic \times Structural Prior
 19: $F_v \leftarrow F_v + \gamma_l \cdot (O_{attn} \odot G)$
- 20: **end for**
- 21: **Stage 4: Final Decoding & Optimization**
- 22: $\hat{\mathcal{M}} \leftarrow \text{SAM2Decoder}(F_v, T)$ ▷ Text-conditioned decoding
- 23: $\mathcal{L}_{seg} \leftarrow \mathcal{L}_{dice}(\hat{\mathcal{M}}, \mathcal{M}^{gt}) + \mathcal{L}_{bce}(\hat{\mathcal{M}}, \mathcal{M}^{gt})$
 24: $z_v^{mask} \leftarrow \text{Pool}(F_v, \mathcal{M}^{gt})$ ▷ Target region features
 25: $\mathcal{L}_{tmc} \leftarrow \text{Contrastive}(z_v^{mask}, F_t)$ ▷ Semantic Alignment
 26: $\mathcal{L}_{bound} \leftarrow \text{BoundaryLoss}(\hat{\mathcal{M}}, \mathcal{M}^{gt})$ ▷ Structural Precision
- 27: $\mathcal{L}_{total} \leftarrow \mathcal{L}_{seg} + \lambda_{tmc} \mathcal{L}_{tmc} + \lambda_{bound} \mathcal{L}_{bound}$
 28: **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:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{tmc} \mathcal{L}_{tmc} + \lambda_{bound} \mathcal{L}_{bound} \quad (5)$$

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 $\tau_{gate} =$

226 1.0, which corresponds to applying the standard Sigmoid function without any scaling.
227 This choice directly uses the SAM-2 decoder’s output logits as the spatial prior, preserv-
228 ing the decoder’s learned confidence distribution.

229 4 Experiments

230 4.1 Implementation Details

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

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

242 **Training Data and Splitting Policy.** We curate a comprehensive multi-modal in-
243 struction tuning dataset consisting of: (1) **Image Referring Segmentation:** RefCOCO/+g [9]
244 [18] (upsampled 4×) and the GCG collection (grounded samples from GranDf [21] and
245 RefCOCOg). (2) **Video Segmentation:** To ensure temporal structural awareness, we in-
246 clude MeVis (4×), ReViOS [31] (10× to emphasize tracking), Ref-Youtube-VOS [30]/Ref-
247 DAVIS [20] (4×), and the SA-V dataset (4×) from SAM-2. (3) **General Instructions:**
248 LLaVA-v1.5 Mix665k [15], Video-ChatUniVi [7], and Osprey-724k [33] are included to
249 maintain broad reasoning capabilities. Importantly, for all benchmarks evaluated (Ref-
250 DAVIS17, Ref-Youtube-VOS, MeVis), we strictly utilize their official **training splits**
251 during the training phase, ensuring that validation and test splits remain completely un-
252 seen to prevent data leakage.

253 **Hyper-parameters & Stabilization.** We set the loss weights $\lambda_{tmc} = 0.1$ and $\lambda_{bound} =$
254 0.05 to balance semantic alignment with boundary precision. The gating temperature
255 τ_{gate} is fixed at 1.0. Notably, to ensure training stability, we apply a **5% linear warm-**
256 **up** to both the auxiliary loss weights and the learnable gating factor γ . This progressive
257 alignment strategy prevents the strong structural bias from potentially disrupting the
258 pre-trained semantic feature space during early training iterations.

259 4.2 Evaluation Benchmarks

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

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

Table 1: Comparison with State-of-the-Art Methods. We report cloU 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]		
	Val	TestA	TestB	Val	TestA	TestB	Val	Test		MME [6]	MMB [17]	SEED [11]
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	76.5	68.2	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
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 cloU, 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 cloU scores partitioned by object area ratio. Our method shows consistent improvements on both salient and non-salient objects.

Method	Salient (> 5%)	Non-Salient ($\leq 5\%$)
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.



“Segment a small carrot that has a larger carrot to its left and onions to its right.”



“Segment the fry on top of the pile of fries.”

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.



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.

347
348
349
350
351
352
353

347
348
349
350
351
352
353

354 5 Discussion

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

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

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

382 6 Conclusion

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

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

397

References

397

- 398 1. Bai, J., et al.: Qwen-VL: A versatile vision-language model for understanding, localization,
399 text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 399
- 400 2. Chen, Z., et al.: Expanding performance boundaries of open-source multimodal models with
401 model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 400
- 402 3. Chen, Z., et al.: InternVL: Scaling up vision foundation models and aligning for generic
403 visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision
404 and Pattern Recognition (CVPR) (2024) 402
- 405 4. Chng, Y.X., Zheng, H., Han, Y., Liu, X., Kankanhalli, M.: Mask grounding for referring
406 image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and
407 Pattern Recognition (CVPR) (2024) 405
- 408 5. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: MeViS: A large-scale benchmark for
409 video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International
410 Conference on Computer Vision (ICCV) (2023) 409
- 411 6. Fu, C., et al.: MME: A comprehensive evaluation benchmark for multimodal large language
412 models. arXiv preprint arXiv:2306.13394 (2023) 411
- 413 7. Jin, P., Takanobu, R., Zhang, C., Cao, X., Yuan, L.: Chat-UniVi: Unified visual representation
414 empowers large language models with image and video understanding. In: Proceedings of the
415 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 413
- 416 8. Jin, S., et al.: One token to seg them all: Language instructed reasoning segmentation in
417 videos. arXiv preprint arXiv:2409.19603 (2024) 416
- 418 9. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referitgame: Referring to objects in
419 photographs of natural scenes. In: EMNLP (2014) 418
- 420 10. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: LISA: Reasoning segmentation
421 via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision
422 and Pattern Recognition (CVPR) (2024) 420
- 423 11. Li, B., et al.: SEED-Bench: Benchmarking multimodal large language models. In: Proceed-
424 ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
425 (2024) 423
- 426 12. Li, X., et al.: LIRA: Inferring segmentation in large multi-modal models with local interleaved
427 region assistance. arXiv preprint arXiv:2501.00000 (2025) 426
- 428 13. Li, Y., Wang, C., Wu, J.: Llama-vid: An image-to-video token for video understanding. arXiv
429 preprint arXiv:2311.17043 (2023) 428
- 430 14. Lin, B., et al.: Video-LLaVA: Learning united visual representation by alignment before pro-
431 jection. arXiv preprint arXiv:2311.10122 (2023) 430
- 432 15. Liu, H., et al.: Improved baselines with visual instruction tuning. In: Proceedings of the
433 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 432
- 434 16. Liu, H., et al.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. LLaVA Blog
435 (2024) 434
- 436 17. Liu, Y., et al.: MMBench: Is your multi-modal model an all-around player? In: Proceedings
437 of the European Conference on Computer Vision (ECCV) (2024) 436
- 438 18. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and com-
439 prehension of unambiguous object descriptions. In: CVPR (2016) 438
- 440 19. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 440
- 441 20. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The
442 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017) 441
- 443 21. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Fahad, S., Khan, F.S.: GLaMM:
444 Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on
445 Computer Vision and Pattern Recognition (CVPR) (2024) 443

- 446 22. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland,
447 C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint
448 arXiv:2408.00714 (2024) 449
- 449 23. Ren, Z., Ji, Z., Lan, G., Wang, Z., Cui, Y., Zhai, W., Feng, J.: PixelLM: Pixel reasoning with
450 large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision
451 and Pattern Recognition (CVPR) (2024) 452
- 452 24. Wang, X., et al.: SegGPT: Segmenting everything in context. In: Proceedings of the
453 IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 454
- 454 25. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: CRIS: CLIP-driven referring
455 image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and
456 Pattern Recognition (CVPR) (2022) 457
- 457 26. Wei, C., Tan, H., Zhong, Y., Yang, Y., Ma, L.: LaSagnA: Language-based segmentation as-
458 sistant for complex queries. arXiv preprint arXiv:2404.02646 (2024) 459
- 459 27. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Tan, P.: ReferFormer: A simple baseline for referring
460 image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and
461 Pattern Recognition (CVPR) (2022) 462
- 462 28. Wu, J., Jiang, Y., Liu, Q., Yuan, Z., Bai, X., Bai, S.: GLEE: General object foundation model
463 for images and videos at scale. In: Proceedings of the IEEE/CVF Conference on Computer
464 Vision and Pattern Recognition (CVPR) (2024) 465
- 465 29. Xia, Z., Han, X., Xue, Y., Zhang, W.: GSVA: Generalized segmentation via multimodal large
466 language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and
467 Pattern Recognition (CVPR) (2024) 468
- 468 30. Xu, N., Yang, L., Fan, Y., Yang, D., Yue, Y., Liang, Y., PRICE, F., Cohen, S., Huang, T.:
469 Youtube-vos: A large-scale video object segmentation benchmark. In: ECCV (2018) 470
- 470 31. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: VISA: Reason-
471 ing video object segmentation via large language models. arXiv preprint arXiv:2407.11325
472 (2024) 473
- 473 32. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: LAVT: Language-aware vision
474 transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference
475 on Computer Vision and Pattern Recognition (CVPR) (2022) 476
- 476 33. Yao, Y., Gisiger, T., Peng, Y., et al.: Osprey: Pixel understanding with visual instruction tun-
477 ing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-
478 nition (CVPR) (2024) 479
- 479 34. Ye, J., et al.: mplug-owl3: Towards long image-sequence understanding in multi-modal large
480 language models. arXiv preprint arXiv:2408.04840 (2024) 481
- 481 35. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang,
482 M.H.: Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images
483 and videos. arXiv preprint arXiv:2501.04001 (2025) 484
- 484 36. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Zhang, L., Li, C., Yang,
485 J.: LLaVA-Grounding: Grounded visual chat with large multimodal models. arXiv preprint
486 arXiv:2312.02949 (2023) 487
- 487 37. Zhang, T., Li, X., Yuan, H., Wan, S., Yang, M.H.: OMG-LLaVA: Bridging image-level,
488 object-level, pixel-level reasoning and understanding. arXiv preprint arXiv:2406.19389
489 (2024) 490
- 490 38. Zhang, T., Li, X., Yuan, H., Wan, S., Yang, M.H.: PSALM: Pixelwise segmentation with large
491 multi-modal model. arXiv preprint arXiv:2403.14598 (2024) 492
- 492 39. Zou, X., et al.: SEEM: Segment everything everywhere all at once. In: Advances in Neural
493 Information Processing Systems (NeurIPS) (2023) 494