STEERGROUNDING: INSTANCE-ADAPTIVE REPRESENTATION STEERING FOR REASONING SEGMENTATION WITH FROZEN LMMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning segmentation requires models to infer an implicitly described target from an image–query pair and delineate it at the pixel level. Existing frozen-LMM grounding methods seek to extract spatial cues from pretrained multimodal representations and decode them into pixel-level predictions while largely preserving the general multimodal capabilities of the pretrained LMMs. However, representations produced by a frozen LMM may not always provide sufficiently reliable target-specific spatial evidence for each image–query pair, leaving downstream refinement to operate on incomplete or ambiguous cues without directly adapting the target-related hidden states. To address this problem, we propose SteerGrounding, which introduces instance-adaptive representation steering into a frozen LMM. Specifically, we introduce Instance-Conditioned Latent Steering (ICLS), which trains a lightweight steering predictor using layer-wise representation differences between the original image and its target-erased counterpart as target-presence supervision. At inference, the predictor directly estimates instance-specific steering from the original image–query pair and injects it into the frozen LMM hidden states for representation adaptation. Building on this, we introduce a Steering-Conditioned Grounding Decoder (SCGD), which reuses the predicted steering representation as a target condition to guide progressive multi-scale spatial decoding, followed by SAM refinement for the final mask prediction. In this way, the same instance-specific steering signal couples internal representation adaptation with downstream spatial grounding. Experiments on ReasonSeg and the RefCOCO series show that SteerGrounding improves ReasonSeg Test-All gIoU by 11.6 points with only 7.3M additional trainable parameters and achieves consistent gains on RefCOCO, RefCOCO+, and RefCOCOg.

1 INTRODUCTION

Reasoning segmentation requires a model to infer an implicitly described target from an image–query pair and predict its pixel-level mask (Lai et al., 2024). Unlike conventional referring segmentation, where the target is typically specified through an explicit referring expression, reasoning segmentation often describes the target indirectly through functional, contextual, or commonsense cues. Large multimodal models (LMMs), with their strong semantic understanding and reasoning capabilities, provide a natural foundation for resolving such implicit references. However, successful reasoning segmentation requires more than identifying the semantic answer: the inferred target must also remain accurately aligned with its corresponding visual instance so that it can be converted into a precise dense mask. The key challenge is therefore to bridge high-level multimodal reasoning and fine-grained spatial grounding.

As illustrated in Fig. 1, existing LMM-based approaches mainly differ in where task adaptation occurs. Fine-tuning-based methods adapt task-related representations by updating part of the multimodal model under segmentation supervision (Lai et al., 2024; Wang et al., 2025; Lu et al., 2026). In contrast, frozen-LMM grounding methods preserve the pretrained backbone and extract spatial cues from representations produced by the original multimodal computation (Wu et al., 2025; Kang et al., 2025). Although these pretrained representations contain useful localization information, task-specific adaptation in such methods mainly occurs after the hidden states and attention

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

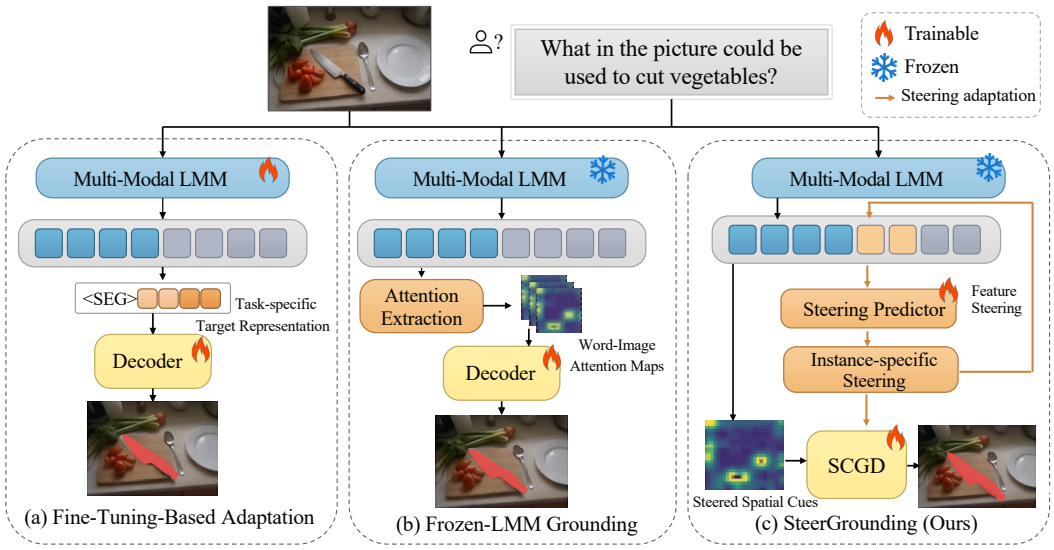


Figure 1: **Illustration of different adaptation paradigms for reasoning segmentation.** (a) Fine-tuning-based methods adapt task-related representations through LMM parameter updates, whereas (b) frozen-LMM grounding methods decode spatial cues from already-formed multimodal representations. (c) SteerGrounding keeps the LMM frozen while adaptively steering its internal representations and reuses the same instance-specific steering signal to guide downstream spatial grounding.

responses have already been formed. When these representations provide incomplete or ambiguous target-specific spatial evidence, downstream feature aggregation and mask refinement can improve the decoding of existing cues, but cannot directly adapt the LMM representations from which those cues originate. This creates a representation-adaptation bottleneck for frozen-LMM grounding.

This bottleneck points to an alternative in which the pretrained parameters remain fixed while the internal representations are adapted to each image–query pair. Activation steering (Rimsky et al., 2024; Xing et al., 2026; Cheraghian et al., 2026) provides a natural mechanism for such representation-space adaptation by intervening directly on hidden states without updating model weights. Most existing multimodal steering methods have focused on generation-side behaviors, such as reducing hallucination, improving visual faithfulness, or controlling model responses, while recent input-dependent approaches further predict sample-specific interventions rather than applying a fixed steering direction (Li et al., 2025; Parekh et al., 2025). For reasoning segmentation, this leaves two questions: how can a frozen LMM learn target-related, instance-specific representation interventions from the current image–query pair without updating its pretrained parameters, and how can the steering information consistently guide subsequent spatial grounding?

To address these questions, we propose SteerGrounding, which introduces instance-adaptive representation steering into frozen-LMM grounding. Specifically, Instance-Conditioned Latent Steering (ICLS) derives target-presence supervision from layer-wise representation differences between the original image and its target-erased counterpart under the same textual context. These differences supervise a lightweight predictor to estimate instance-specific steering directly from the original image–query pair. At inference, the predicted steering is injected into the frozen LMM hidden states without requiring target annotations or target-erased images. We further introduce a Steering-Conditioned Grounding Decoder (SCGD), which reuses the same steering representation to condition progressive multi-scale spatial decoding. Together, ICLS and SCGD make the steering representation a shared interface between internal representation adaptation and downstream grounding. Experiments on ReasonSeg and the RefCOCO series demonstrate substantial improvements on reasoning segmentation and consistent gains across referring segmentation benchmarks.

Our main contributions are summarized as follows:

- We formulate frozen-LMM reasoning segmentation from the perspective of **instance-adaptive representation steering**, allowing internal representations to adapt to individual

image–query pairs while largely preserving the general multimodal capabilities inherited from the pretrained LMM.

- We develop **Instance-Conditioned Latent Steering (ICLS)**, which derives target-presence supervision from layer-wise representation differences between the original image and its target-erased counterpart, and learns to predict instance-specific steering directly from the original image–query pair.
- We introduce the **Steering-Conditioned Grounding Decoder (SCGD)**, which reuses the same predicted steering representation to condition progressive multi-scale spatial decoding, making steering a shared interface between internal representation adaptation and downstream grounding.
- Experiments on ReasonSeg and the RefCOCO series demonstrate consistent improvements in both reasoning and referring expression segmentation.

2 Related Work

Grounding with Fine-Tuned Large Multimodal Models. Reasoning segmentation extends referring segmentation to implicit queries, requiring semantic reasoning to be translated into spatial predictions. LISA established the embedding-as-mask paradigm (Lai et al., 2024), extended by PixelLM, GLaMM, SegLLM, and MMR to richer pixel grounding and multi-target or interactive reasoning (Ren et al., 2024b; Rasheed et al., 2024; Wang et al., 2025; Jang et al., 2025). CoPRS (Lu et al., 2026) makes this interface explicit through a chain-of-thought-derived positional prior, while VGR and RegionReasoner incorporate region evidence into reasoning (Wang et al., 2026; Sun et al., 2026). These advances strengthen the reasoning-to-spatial interface or trajectory, but reasoning must still distinguish the intended instance from plausible alternatives. We address this challenge by adapting target-associated hidden states inside a frozen LMM before grounding.

Grounding with Frozen Large Multimodal Models. Frozen-LMM grounding exploits pretrained spatial cues without updating the backbone. F-LMM and localization-head analysis demonstrate the grounding potential of word–image attention (Wu et al., 2025; Kang et al., 2025); HG-LMM and LENS further organize pretrained cues into instance-aware features or keypoint prompts (Li et al., 2026; Liu & Chen, 2025). These approaches primarily improve how existing cues are selected and decoded. SteerSeg instead conditions attention through soft prompts and chain-of-thought-derived attributes for reasoning video segmentation (Cheraghian et al., 2026). Related region-understanding methods also guide frozen models: CRG contrasts output distributions with and without region information (Wan et al., 2024), while ControlMLLM optimizes latent variables at inference to direct attention toward visually prompted regions (Wu et al., 2024a). SteerGrounding instead learns target-associated hidden-state interventions from target-presence supervision and reuses the predicted signal for mask decoding, without requiring target-region prompts at inference.

Activation Steering for Multimodal Models. Activation steering adapts frozen-model computation through hidden-state interventions. Contrastive Activation Addition derives directions from activation differences (Rimsky et al., 2024); VTI, VISTA, and SteerVLM extend steering to visual representation enhancement and multimodal behavior control (Liu et al., 2025; Li et al., 2025; Sivakumar et al., 2025). Text-derived steering vectors can also improve multimodal visual understanding (Gan et al., 2026). ReFT learns interventions on frozen language-model representations (Wu et al., 2024b), and MoReS learns visual-subspace transformations for multimodal instruction tuning (Bi et al., 2025). Learn-to-Steer further predicts input-specific shifts (Parekh et al., 2025). Closely related, CSteer applies contextual latent interventions to improve multi-region understanding from visual prompts (Xing et al., 2026). Our task instead requires predicting a target mask from an image–query pair, with an intervention signal that also supports spatial decoding. ICLS learns this signal from a target-presence contrast and uses the same representation for both internal intervention and decoder conditioning, connecting representation adaptation to target localization.

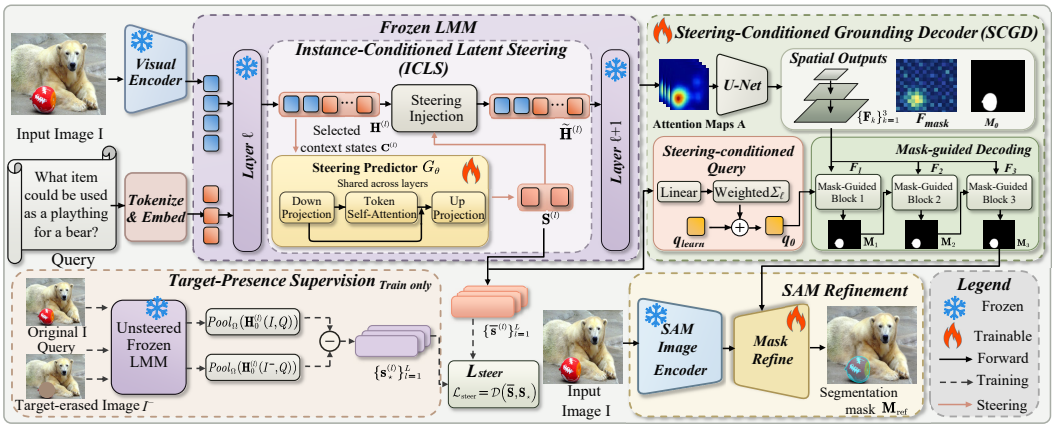


Figure 2: **Overview of SteerGrounding.** ICLS predicts token-specific steering vectors to adapt target-related hidden representations within a frozen LMM. The instance-specific steering is reused by SCGD to condition progressive multi-scale spatial decoding, followed by SAM refinement. During training, layer-wise representation differences between the original and target-erased images provide target-presence supervision.

3 METHOD

3.1 OVERVIEW

Given an image I and a textual query Q, SteerGrounding predicts a pixel-level target mask while keeping the pretrained LMM frozen. As illustrated in Fig. 2, the framework comprises Instance-Conditioned Latent Steering (ICLS), which adapts hidden representations within the frozen LMM, and a Steering-Conditioned Grounding Decoder (SCGD), which reuses the predicted steering to guide spatial mask decoding.

For reasoning queries, the LMM first generates a target description R from the image–query pair (I, Q), and a subsequent grounding pass uses both Q and R as textual context. For referring expressions, the grounding pass directly uses Q without generating an additional description. During grounding, ICLS predicts token-specific steering vectors from textual hidden states and injects them into the frozen LMM across layers. The resulting text-to-image attention is aggregated and transformed by a lightweight U-Net into multi-scale spatial features. The predicted steering vectors are also aggregated to initialize the SCGD decoding query, which guides progressive mask decoding over these spatial features. The same steering signal thus connects internal representation adaptation with downstream spatial grounding. Finally, the decoded mask is converted into prompts for SAM (Kirillov et al., 2023) refinement to obtain the final segmentation prediction.

3.2 INSTANCE-CONDITIONED LATENT STEERING

ICLS is designed to make the frozen LMM's internal representations more responsive to the target specified by each image–query pair. It predicts instance-specific steering vectors to adjust target-related hidden states across layers, enabling task-specific representation adaptation.

Steering prediction. We predict token-specific steering vectors by modeling dependencies among the selected textual hidden states in a low-dimensional bottleneck space. At layer l, these states form $\mathbf{C}^{(l)} = [\mathbf{H}_t^{(l)}]_{t\in\mathcal{T}} \in \mathbb{R}^{T\times d}$, where $\mathcal{T}$ is the set of selected textual token positions and $\mathbf{H}_t^{(l)} \in \mathbb{R}^d$ is the hidden state of token t produced by the l-th LMM block before steering injection. A lightweight self-attention predictor contextualizes these states while retaining their token-wise representations through a residual connection, then maps the result back to the LMM hidden dimension:

$$\begin{aligned} \mathbf{Z}^{(l)} &= \phi_{\downarrow}\left(\mathbf{C}^{(l)}\right), \\ \hat{\mathbf{S}}^{(l)} &= \phi_{\uparrow}\left(\mathbf{Z}^{(l)} + \mathrm{MHA}\left(\mathrm{LN}\left(\mathbf{Z}^{(l)}\right)\right)\right). \end{aligned} \tag{1}$$

Here, $\mathbf{Z}^{(l)} \in \mathbb{R}^{T\times r}$ denotes the bottleneck representation with dimension r, and $\hat{\mathbf{S}}^{(l)} \in \mathbb{R}^{T\times d}$ contains the predicted steering vectors. The token-wise projections $\phi_\downarrow$ and $\phi_\uparrow$ map between the LMM hidden space and the bottleneck space. MHA and LN denote multi-head self-attention and layer normalization, respectively. The predictor parameters are shared across LMM layers, and each row $\hat{\mathbf{s}}_t^{(l)}$ of $\hat{\mathbf{S}}^{(l)}$ is the steering vector for token t at layer l.

Steering injection. We inject the predicted steering vectors into the selected hidden states to adapt the frozen LMM's internal representations to the current image–query pair. The injection adjusts representation directions while preserving the original hidden-state norms, with a learnable layer-wise strength controlling the steering contribution. The steering injection is formulated as

$$\begin{aligned} \mathbf{z}_t^{(l)} &= \mathcal{N}\left(\mathbf{H}_t^{(l)}\right) + \lambda_l \mathcal{N}\left(\hat{\mathbf{s}}_t^{(l)}\right), \\ \tilde{\mathbf{H}}_t^{(l)} &= \|\mathbf{H}_t^{(l)}\|_2 \mathcal{N}\left(\mathbf{z}_t^{(l)}\right), \qquad t \in \mathcal{T}. \end{aligned} \tag{2}$$

Here, $\mathcal{N}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$ denotes L_2 normalization,λ_l controls the learnable steering strength at layer l, and $\tilde{\mathbf{H}}_t^{(l)}$ is the updated hidden state passed to subsequent transformer blocks of LMM. Hidden states outside $\mathcal{T}$ are not directly modified by the injection operation. Since subsequent blocks process the modified representations, steering predictions at later layers depend on hidden states already modified by earlier interventions.

3.3 Steering-Conditioned Grounding Decoder

SCGD employs the instance-specific steering learned by ICLS to condition spatial mask decoding. By reusing the steering representation as a decoding condition, it connects internal representation adaptation with target localization over the steered LMM's spatial features.

Spatial outputs. We construct spatial outputs from the steered LMM's attention to support both query updating and mask prediction. Following F-LMM, text-to-image attention from selected layers and heads is aggregated into a spatial representation $\mathbf{A}$. A lightweight U-Net transforms $\mathbf{A}$ into multi-scale features for query updating, a shared pixel embedding for mask prediction, and initial mask logits for spatial guidance:

$$\left(\{\mathbf{F}_k\}_{k=1}^K, \mathbf{F}_{\text{mask}}, \mathbf{M}_0\right) = \mathcal{U}(\mathbf{A}), \tag{3}$$

where $\mathbf{F}_k$ denotes the spatial feature used for query updating at decoding stage k. The shared pixel embedding $\mathbf{F}_{\text{mask}}$ is obtained by applying a separate 1×1 convolutional projection to the U-Net's final feature map. It provides a D-dimensional representation at each spatial location and is shared across decoding stages to predict masks from the updated queries. The initial mask logits $\mathbf{M}_0$ provide spatial guidance for the first decoding stage. We use $K = 3$ stages.

Steering-conditioned query. To incorporate instance-specific steering into mask decoding, we first aggregate the token-wise steering vectors within each layer. The layer-wise representations are then projected and combined to initialize the decoding query:

$$\begin{aligned} \bar{\mathbf{s}}^{(l)} &= \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \hat{\mathbf{s}}_t^{(l)}, \\ \mathbf{q}_0 &= \text{LN}\left(\mathbf{q}_{\text{learn}} + \beta \sum_{l=1}^{L} \pi_l \phi_s\left(\bar{\mathbf{s}}^{(l)}\right)\right). \end{aligned} \tag{4}$$

Here, $\bar{\mathbf{s}}^{(l)}$ is the mean steering vector at layer l, and $\phi_s : \mathbb{R}^d \to \mathbb{R}^D$ is a shared projection into the decoder space. The weights π_l are softmax-normalized learnable layer weights over the L LMM layers. The learnable query $\mathbf{q}_{\text{learn}}$ provides a shared task-level component, while β controls the contribution of the instance-specific steering condition.

Mask-guided decoding. SCGD uses the preceding mask prediction as spatial guidance for progressively updating the steering-conditioned query. At each stage, soft mask-based weights modulate the spatial features used in cross-attention without excluding alternative locations. The updated query predicts a new mask through the shared pixel embedding, providing guidance for the next stage.

Starting from $\mathbf{q}_0$ and $\mathbf{M}_0$, decoding proceeds from coarse to fine as

$$\begin{aligned} \mathbf{W}_k &= 1 + \alpha\sigma\left(\mathrm{Resize}_k\left(\mathbf{M}_{k-1}\right)\right), \\ \mathbf{q}_k &= \mathrm{FFN}_k\left(\mathrm{CA}_k\left(\mathbf{q}_{k-1}, \mathbf{F}_k \odot \mathbf{W}_k, \mathbf{F}_k \odot \mathbf{W}_k\right)\right), \\ \mathbf{M}_k &= \phi_m(\mathbf{q}_k)^\top \mathbf{F}_{\mathrm{mask}}, \qquad k = 1, \ldots, K. \end{aligned} \tag{5}$$

Here, σ denotes the sigmoid function, and Resize_k aligns the preceding mask logits with the spatial resolution of $\mathbf{F}_k$. The coefficient α controls the modulation strength, and $\odot$ applies the spatial weights across feature channels. CA_k and FFN_k denote cross-attention and the feed-forward network at stage k, respectively. The projection ϕ_m maps the updated query to mask weights, whose pointwise inner products with $\mathbf{F}_{\mathrm{mask}}$ yield $\mathbf{M}_k$. After the final decoding stage, $\mathbf{M}_K$ is converted into prompts for SAM refinement, yielding the final mask logits $\mathbf{M}_{\mathrm{ref}}$.

3.4 Training Objectives

Joint training objective. We combine target-presence supervision for steering prediction with segmentation supervision on the initial, stage-wise, and SAM-refined masks:

$$\mathcal{L} = \lambda_{\mathrm{steer}}\mathcal{L}_{\mathrm{steer}} + \lambda_{\mathrm{coarse}} \sum_{k=0}^{K} \omega_k \ell_{\mathrm{seg}}(\mathbf{M}_k, Y) + \lambda_{\mathrm{refine}}\ell_{\mathrm{seg}}(\mathbf{M}_{\mathrm{ref}}, Y). \tag{6}$$

Here, Y is the ground-truth mask, and ℓ_{seg} combines binary cross-entropy on logits with Dice loss on sigmoid probabilities. The λ coefficients balance the three supervision terms, while $\omega_k \geq 0$ and $\sum_{k=0}^{K} \omega_k = 1$ weight the initial ($k = 0$) and stage-wise mask losses.

Target-presence supervision. We derive steering targets by contrasting representations with and without the annotated target under the same textual context. The target-erased image I^- is constructed by replacing pixels inside Y with the channel-wise image mean. The frozen LMM processes both images without steering, using the grounding context Q described in Section 3.1:

$$\mathbf{s}_\star^{(l)} = \frac{1}{|\Omega|} \sum_{t \in \Omega} \left[\mathbf{H}_{0,t}^{(l)}(I, P) - \mathbf{H}_{0,t}^{(l)}(I^-, P) \right]. \tag{7}$$

Here, $\mathbf{H}_{0,t}^{(l)}$ is the unsteered hidden state of token t after block l, and Ω includes the textual positions used for teacher extraction and target-region visual tokens. The teacher is detached and used only during training. We match the layer-wise mean prediction $\bar{\mathbf{s}}^{(l)}$ in Eq. 4 to $\mathbf{s}_\star^{(l)}$ through complementary element-wise, directional, and magnitude losses:

$$\mathcal{L}_{\mathrm{steer}} = \mathcal{L}_{\mathrm{mse}} + \eta_1\mathcal{L}_1 + \eta_{\cos}\mathcal{L}_{\cos} + \eta_{\mathrm{norm}}\mathcal{L}_{\mathrm{norm}}. \tag{8}$$

All terms are averaged across LMM layers. $\mathcal{L}_{\mathrm{mse}}$ and $\mathcal{L}_1$ measure element-wise discrepancies, while $\mathcal{L}_{\cos}$ and $\mathcal{L}_{\mathrm{norm}}$ match direction and magnitude. The η coefficients balance these terms.

4 Experiments

4.1 Experimental Details

SteerGrounding is implemented in PyTorch 2.1.2 and trained on four NVIDIA A800 GPUs with a per-GPU batch size of 1. We build on F-LMM with DeepSeek-VL-1.3B and DeepSeek-VL-7B (Lu et al., 2024), while keeping the LMM backbone and the SAM image encoder frozen. Training proceeds in two stages. We first train the model for four epochs on the combined training splits of RefCOCO, RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016) using AdamW with a weight decay of 1×10^{-4}. The learning rates for ICLS and SCGD are set to 1×10^{-5} and 1×10^{-4}, respectively. We then fine-tune our model on the ReasonSeg training split (Lai et al., 2024) for 30 epochs, initializing from their respective checkpoints trained on the RefCOCO series in the first stage. The learning rates for ICLS and SCGD are increased to 5×10^{-5} and 5×10^{-4}, respectively. Additionally, we set the ICLS bottleneck dimension of $r = 256$ and a three-stage SCGD with decoder dimension $D = 256$. Text-to-image attention maps are aggregated from selected Transformer layers and attention heads. We set the spatial feature modulation strength to $\alpha = 1.0$ and the steering-conditioning coefficient to $\beta = 0.1$. The loss weights λ_{steer}, $\lambda_{\mathrm{coarse}}$, and $\lambda_{\mathrm{refine}}$ are set to 0.1, 0.45, and 0.45, respectively.

Table 1: Comparison with existing methods on reasoning segmentation and referring expression segmentation. * denotes results reproduced by us following the F-LMM experimental setup. All results are reported in percentage (%).

Type	Method	Reasoning Segmentation								Referring Expression Segmentation		
		Val All		Test Short		Test Long		Test All		RefCOCO	RefCOCO+	RefCOCOg
		gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	cIoU	cIoU	cIoU
Without LMM	X-Decoder (Zou et al., 2023a)	22.6	–	20.4	–	22.2	–	21.7	–	–	–	64.6
	SEEM (Zou et al., 2023b)	25.5	21.2	20.1	11.5	25.6	20.8	24.3	18.7	–	–	65.7
	Grounded SAM (Ren et al., 2024a)	26.0	14.5	17.8	10.8	22.4	18.6	21.3	16.4	–	–	–
	OVSeg (Liang et al., 2023)	28.5	18.6	18.0	15.5	28.7	22.5	26.1	20.8	–	–	–
Fine-tuned LMM	SegAgent-7B (Zhu et al., 2025)	33.0	25.4	–	–	–	–	33.5	31.3	–	–	–
	VISA-7B (Yan et al., 2024)	52.7	57.8	–	–	–	–	–	–	72.4	59.8	65.5
	LISA-7B (Lai et al., 2024)	52.9	54.0	40.6	40.6	49.4	51.0	47.3	48.4	74.9	65.1	67.9
	Seg-R1-7B (You & Wu, 2025)	58.6	41.2	–	–	–	–	56.7	53.7	74.3	62.6	71.0
Frozen LMM	LENS-7B (Liu & Chen, 2025)	**61.4**	**65.3**	47.8	41.7	59.3	**61.6**	56.5	**57.3**	76.5	66.1	69.4
	Qwen2.5-SAM-7B (Hao et al., 2026)	57.6	48.3	–	–	–	–	50.1	41.2	72.4	59.8	65.5
	RSVP-7B (Lu et al., 2025)	59.2	56.7	**47.9**	**42.0**	58.4	53.0	55.9	50.7	–	–	–
	F-LMM-1.3B* (Wu et al., 2025)	39.0	23.2	29.8	23.6	44.3	29.9	40.7	28.4	75.0	62.8	68.2
	F-LMM-7B* (Wu et al., 2025)	44.9	29.2	35.9	25.6	52.4	40.3	48.4	36.5	76.1	66.4	70.1
	SteerGrounding-1.3B (Ours)	55.5	44.7	40.4	29.3	59.5	53.4	54.8	46.7	76.8	64.7	70.0
	SteerGrounding-7B (Ours)	59.3	55.3	46.1	36.5	**63.0**	57.6	**60.0**	51.9	**77.9**	**69.1**	**73.6**

4.2 Results on Segmentation and Multimodal Question Answering

Tables 1 and 2 compare segmentation and multimodal question-answering performance. SteerGrounding consistently improves over the corresponding F-LMM baselines on segmentation while maintaining comparable performance on the evaluated question-answering benchmarks.

Reasoning Segmentation. As shown in Table 1, SteerGrounding-7B achieves the highest Test-All and Test-Long gIoU among the methods compared in Table 1, reaching 60.0 and 63.0, respectively. Compared with our reproduced F-LMM baseline using the same DeepSeek-VL-7B backbone, it improves Test All gIoU and cIoU by 11.6 and 15.4 points. The gIoU gains extend to both short and long queries, with improvements of 10.2 and 10.6 points, respectively. The 1.3B model improves Test All gIoU/cIoU from 40.7/28.4 to 54.8/46.7, showing consistent gains at both model scales.

Referring Expression Segmentation. SteerGrounding-7B achieves validation cIoU scores of 77.9, 69.1, and 73.6 on RefCOCO, RefCOCO+, and RefCOCOg, respectively, outperforming the corresponding F-LMM baseline by 1.8, 2.7, and 3.5 points. These scores are the highest among the methods reported in Table 1. The 1.3B model also improves over F-LMM on all three datasets, supporting the effectiveness of the proposed representation adaptation for referring expression segmentation at both scales.

Multimodal Question Answering. Table 2 evaluates multimodal question answering on MME (Fu et al., 2023), MMBench (Liu et al., 2024), MM-Vet (Yu et al., 2024), and LLaVA-W (Liu et al., 2023). SteerGrounding remains close to F-LMM using the same DeepSeek-VL backbone. For the 7B model, MMBench decreases by 1.1 points, while MM-Vet and LLaVA-W increase by 0.3 and 0.8 points. The 1.3B model shows similarly limited changes on these three benchmarks. Together with the segmentation results, these findings show that improved grounding is accompanied by largely maintained performance on the evaluated question-answering benchmarks.

Table 2: Comparison on multimodal question answering benchmarks.

Model	MME	MMBench	MM-Vet	LLaVA-W
PixelLM-7B (Ren et al., 2024b)	309/135	17.4	15.9	46.4
LISA-7B (Lai et al., 2024)	1/1	0.4	19.1	47.5
LENS (Liu & Chen, 2025)	1801/–	64.0	–	–
LaSagnA-7B (Wei et al., 2024)	0/0	0.0	16.7	34.5
GSVA-7B (Xia et al., 2024)	446/18	17.8	19.4	38.3
F-LMM-1.3B (Wu et al., 2025)	1286/215	64.3	27.2	44.1
F-LMM-7B (Wu et al., 2025)	1466/303	**74.2**	40.1	74.0
SteerGrounding-1.3B	1258/227	64.1	26.7	43.8
SteerGrounding-7B	1460/309	73.1	**40.4**	**74.8**

Table 3: Component ablation on ReasonSeg (gIoU). **Gray row** denotes the default configuration.

ICLS	SCGD	Params.	Val	Short	Long	All
×	×	12.8 M	44.9	35.9	52.4	48.4
✓	×	15.1 M	57.7	46.0	62.8	58.8
×	✓	17.8 M	53.2	39.0	56.2	52.0
✓	✓	20.1 M	**59.3**	**46.1**	**63.0**	**60.0**

Table 4: SCGD query component ablation on ReasonSeg (gIoU). **Gray row** denotes the default configuration.

Learned	Steering	Val	Short	Long	All
×	×	57.7	46.0	62.8	58.8
✓	×	58.6	45.7	62.9	58.8
×	✓	**60.1**	43.8	62.8	58.2
✓	✓	59.3	**46.1**	**63.0**	**60.0**

Table 5: Steering injection ablation on ReasonSeg (gIoU). **Gray row** denotes the default configuration.

Injection	Val	Short	Long	All
Final token	57.0	42.9	62.1	57.4
Text	**59.3**	**46.1**	**63.0**	**60.0**
Image	55.1	45.1	60.4	56.7
Text+Image	59.2	44.5	61.3	57.2

Table 6: Steering source ablation on ReasonSeg (gIoU). **Gray row** denotes the default setting.

Source	Val	Short	Long	All
Zero	51.1	38.1	57.6	52.9
Static	52.9	41.1	59.6	55.1
Random	50.8	38.6	56.3	52.0
GT-derived	55.0	44.6	61.5	57.4
Ours	**59.3**	**46.1**	**63.0**	**60.0**

4.3 Ablation Studies

We conduct all ablation studies on ReasonSeg with DeepSeek-VL-7B. Unless otherwise stated, the training and evaluation protocols follow those used in the main experiments. We first isolate the contributions of ICLS and SCGD and then examine SCGD query composition, the steering injection location, and steering source.

Component contributions. Table 3 examines the contributions of ICLS and SCGD. ICLS alone improves Test All gIoU from 48.4 to 58.8, while SCGD alone reaches 52.0. Combining both modules achieves the highest gIoU across all splits, reaching 60.0 on Test All, an improvement of 11.6 points over the baseline and 1.2 points over ICLS alone. These results identify ICLS as the main contributor to the improvement, with SCGD providing additional gains. The complete model uses 20.1M trainable parameters: 12.8M for the F-LMM adapter, 2.3M for ICLS, and 5.0M for SCGD.

Query components. Table 4 examines the query components, with the ICLS-only model serving as the reference without SCGD. The learned query alone reaches 58.8 Test All gIoU, while steering alone achieves the highest Val gIoU of 60.1 but obtains 58.2 on Test All. Combining the learned query with the steering condition achieves the highest scores on all three test splits (46.1/63.0/60.0), improving Test All gIoU by 1.2 and 1.8 points over the learned-only and steering-only queries.

Steering injection. Table 5 compares four injection locations. Text-only injection achieves the highest gIoU across all splits, reaching 60.0 on Test All. It outperforms final-token, image-only, and joint text-image injection by 2.6, 3.3, and 2.8 points, respectively. These results support steering the selected textual states, with no further gain from direct visual-token injection.

Steering sources. Table 6 replaces the steering source while keeping the remaining model fixed. Predicted steering achieves the highest gIoU across all splits, reaching 60.0 on Test All compared with 52.9, 55.1, and 52.0 for zero, static, and norm-matched random steering, respectively. The ground-truth-derived source obtains 57.4. This advantage may reflect better alignment of the learned steering with the injection and decoding mechanisms through joint training.

4.4 Visualization Analysis

Effect of steering on LMM attention. Figure 3 illustrates how steering alters the attention responses of the frozen LMM. Without steering, target-related responses are relatively weak or spatially incomplete and are mixed with activations in unrelated regions. After steering, the responses become more prominent over the door in panel (a) and cover a larger portion of the target refrigerator in panels (b) and (c). These observations suggest that steering makes target-related spatial evidence

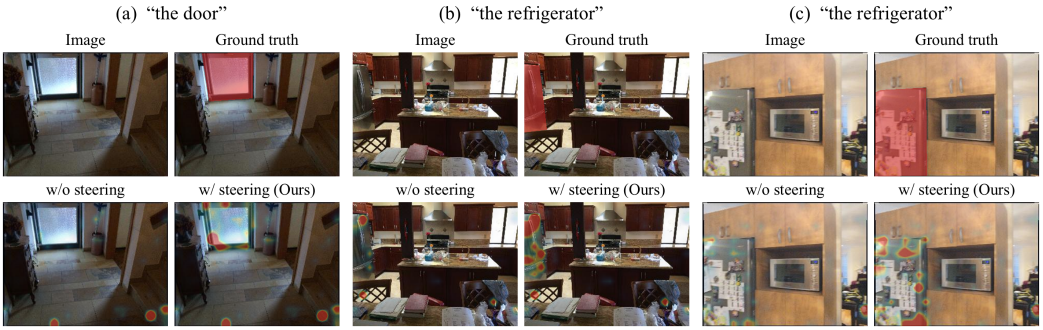


Figure 3: Qualitative comparison of LMM attention maps without and with steering.

Query	Input image	Reasoning excerpts	Prediction
What is the area that displays the time in this image?		The illuminated digits are located on the digital display at the top of the clock.	
What is something that helps in the pollination of plants in this image?		The image shows a butterfly resting on a purple flower, so the butterfly is the object serving this function.	
Graduation is a significant milestone in one's education. What in the picture could be worn by someone to signify their achievement during the graduation ceremony?		The person is wearing a square graduation cap, which visibly signals the educational achievement described in the query.	
In many countries, it is illegal to park a car in certain areas without proper permission. What part of the picture can help identify the ownership and registration of the car to ensure it is legally parked?		A vehicle's ownership and registration are identified by its unique number plate.	
In cold weather, dogs may need extra protection to keep them warm. What object in the picture can a dog wear to provide warmth during snowy walks?		The dog is wearing a bright red coat, making this coat the item that keeps it warm during snowy walks.	

Figure 4: Qualitative analysis of reasoning segmentation.

more explicit before downstream decoding. Background activations remain visible, indicating that the attention maps provide intermediate spatial cues rather than complete segmentation masks.

Qualitative analysis of reasoning segmentation. Figure 4 presents qualitative examples involving functional, contextual, and commonsense reasoning. In the pollination, graduation, and warmth-related cases, SteerGrounding correctly associates the implicit descriptions with the butterfly, graduation cap, and dog coat, respectively. The predictions also preserve fine-grained spatial specificity, localizing the digital display rather than the entire clock and the number plate rather than the whole car. These examples illustrate that SteerGrounding can translate high-level reasoning into spatially selective masks for both object-level and part-level targets.

5 Conclusion

We introduce SteerGrounding, which adapts a frozen LMM's internal representations to each image–query pair and reuses the predicted steering for spatial decoding. ICLS learns instance-specific steering from representation differences between original and target-erased images and injects it into the frozen LMM, while SCGD reuses the same steering representation to condition progressive multi-scale mask decoding. Experiments on ReasonSeg and the RefCOCO series show consistent gains over the corresponding baselines, alongside comparable performance on the evaluated multi-modal question-answering benchmarks. These results support instance-specific steering as a shared representation for connecting internal representation adaptation with pixel-level grounding.

AI USE STATEMENT

We used OpenAI ChatGPT/Codex to assist with drafting and revising manuscript sections, improving clarity and readability, and identifying relevant literature. The authors take full responsibility for the final content of the paper, including its claims, results, and references.

REPRODUCIBILITY STATEMENT

Sections 3.2 and 3.3 describe the steering predictor, representation intervention, and grounding decoder. Section 3.4 presents the training objectives, and Section 4.1 reports the model backbones, training schedule, optimization settings, and evaluation protocol. Our experiments use the publicly available ReasonSeg and RefCOCO-series datasets.

REFERENCES

Jinhe Bi, Yujun Wang, Haokun Chen, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. LLaVA steering: Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15230–15250, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.739.

Ali Cheraghian, Hamidreza Dastmalchi, Abdelwahed Khamis, Morteza Saberi, Aijun An, and Lars Petersson. SteerSeg: Attention steering for reasoning video segmentation. *arXiv preprint arXiv:2605.14908*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2023.

Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40056–40087, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1861.

Chao Hao, Jun Xu, Ji Du, Shuo Ye, Ziyue Qiao, Xiaodong Cun, Guangcong Wang, Xubin Zheng, and Zitong Yu. Seg-agent: Test-time multimodal reasoning for training-free language-guided segmentation. *arXiv preprint arXiv:2605.12953*, 2026.

Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Dae-Shik Kim. MMR: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. In *International Conference on Learning Representations*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9339–9350, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

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

Wenjie Li, Jiale Cao, Jin Xie, Aiping Yang, and Yanwei Pang. HG-LMM: Unleashing high-quality pixel grounding capabilities in frozen large multimodal models. *IEEE Transactions on Image Processing*, 35:9099–9112, 2026. doi: 10.1109/TIP.2026.3718446.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large

vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819, 2025.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516.

Jiazhen Liu and Long Chen. Segmentation as a plug-and-play capability for frozen multimodal LLMs. *arXiv preprint arXiv:2510.16785*, 2025.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? In *European Conference on Computer Vision*, pp. 216–233, 2024.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, and Wenbo Zhu. RSVP: Reasoning segmentation via visual prompting and multi-modal chain-of-thought. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Yan Feng, Bin Chen, Ke Chen, and Yaowei Wang. CoPRS: Learning positional prior from chain-of-thought for reasoning segmentation. In *International Conference on Learning Representations*, 2026.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Arnaud Dapogny, Alasdair Newson, and Matthieu Cord. Learning to steer: Input-dependent steering for multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5341.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024a.

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

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Anushka Sivakumar, Andrew Zhang, Zaber Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665, 2025.

Wenfang Sun, Hao Chen, Yingjun Du, Yefeng Zheng, and Cees G. M. Snoek. RegionReasoner: Region-grounded multi-round visual reasoning. In *International Conference on Learning Representations*, 2026.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive Region Guidance: Improving Grounding in Vision-Language Models Without Training. In *European Conference on Computer Vision*, pp. 198–215, 2024. doi: 10.1007/978-3-031-72986-7_12.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. VGR: Visual grounded reasoning. In *International Conference on Learning Representations*, 2026.

Xudong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. SegLLM: Multi-round reasoning segmentation with large language models. In *International Conference on Learning Representations*, 2025.

Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. LaSagnA: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2024.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. ControlMLLM: Training-free visual prompt learning for multimodal large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 45206–45234. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1436.

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-LMM: Grounding frozen large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24710–24721, 2025.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 63908–63962. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2041.

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

Yun Xing, Hanyuan Liu, Jiahao Nie, and Shijian Lu. Referring Multiple Regions with Large Multimodal Models via Contextual Latent Steering. *arXiv preprint arXiv:2605.01827*, 2026.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, 2024.

Zuyao You and Zuxuan Wu. Seg-R1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *European Conference on Computer Vision*, pp. 69–85, 2016. doi: 10.1007/978-3-319-46475-6_5.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754, 2024.

Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. SegAgent: Exploring pixel understanding capabilities in MLLMs by imitating human annotator trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3686–3696, 2025.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15116–15127, 2023a.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*, 2023b.