THINK, COMPRESS, ANSWER: REINFORCEMENT LEARNING FOR MULTIMODAL GROUNDING VIA REASONING COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

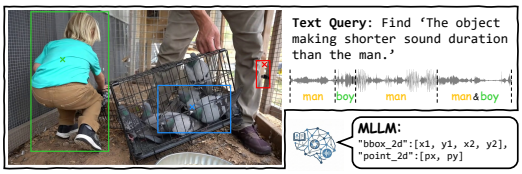
Extended reasoning has substantially improved multimodal large language models (MLLMs) on tasks requiring compositional and relational inference. However, multimodal grounding imposes a distinct requirement: reasoning must ultimately converge on a target-specific compact representation that supports precise localization. As reasoning trajectories grow longer, they often expand into broad scene narratives and peripheral relational details, which can dilute grounded perceptual evidence and lead to ambiguous or incorrect predictions. Motivated by the information bottleneck principle, we propose **Think-Compress-Answer (TCA)**, a reinforcement learning framework that decomposes multimodal grounding into three phases: free-form reasoning, target-aware context compression, and answer generation conditioned on the compressed context. TCA is optimized through a multi-stage Group Relative Policy Optimization (GRPO) with progressive rewards that establish the structured output protocol, minimize reasoning-specific information while preserving target-relevant content, and directly improve localization accuracy. This staged design enables MLLMs to retain the benefits of extended reasoning while distilling it into the minimal context required for grounding. Experiments across six multimodal grounding benchmarks show that TCA attains substantial zero-shot performance gains, demonstrating the effectiveness of reasoning compression.

1 INTRODUCTION

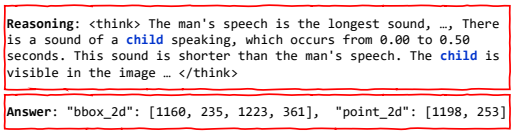
Multimodal grounding requires a model to identify the entity, event, or region in multimodal observations that is referred to by a text query. Although the final output is typically compact (*e.g.*, bounding box, segmentation mask, center point, or temporal interval Yu et al. (2016); Seo et al. (2020); Khoreva et al. (2018); Ding et al. (2023); Wang et al. (2024)), the model must interpret the query, distinguish among perceptually similar candidates, resolve spatial or temporal relations, and integrate evidence distributed across entities, observations, and modalities. Recent multimodal large language models (MLLMs) have supported these capabilities through explicit reasoning procedures, including chain-of-thought (CoT) prompting Wei et al. (2022); Zhang et al. (2024b), instruction-following Liu et al. (2023); Dai et al. (2023), process supervision and intermediate verification Zheng et al. (2023a); Shao et al. (2024a), and test-time scaling DeepSeek-AI (2025); OpenAI (2024); Snell et al. (2024) that allocates additional computation to explore and refine reasoning paths. These advances suggest that reasoning is important for grounding targets that cannot be identified through direct query-perception matching alone. However, multimodal grounding also imposes a requirement that differs from open-ended reasoning: the reasoning process must ultimately isolate the compact, target-specific evidence necessary for precise localization. This raises a fundamental question: *How can MLLMs retain the benefits of extended reasoning while preserving only the information required for grounding?*

Recent works have demonstrated that MLLMs can drift away from perceptual evidence, increasingly rely on language priors, and introduce hallucinated or weakly relevant details, as reasoning paths grow longer Zhou et al. (2024); Li et al. (2023); Liu et al. (2025a). This problem is particularly pronounced in multimodal grounding, where CoT often expands a target-oriented analysis into broad scene descriptions and relational narratives that exceed the information required to localize the

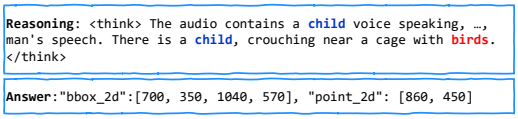
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



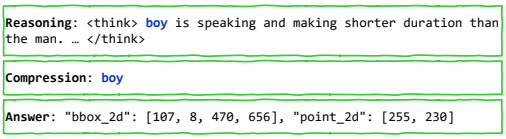
(a) Multimodal grounding



(b) Prediction through Chain-of-Thought (CoT)



(c) Prediction through answer-reward-only RL



(d) Prediction from TCA (Ours)

Figure 1: **Motivation of Think-Compress-Answer.** Given multimodal inputs, (b) standard CoT and (c) answer-reward-only RL methods identify the correct entity but drift during localization. (d) Our Think-Compress-Answer distills reasoning into a concise, faithful cue, yielding accurate grounding. The predicted bounding boxes and center points in (a) and the output from each method (b–d) are matched in color.

referred target. Moreover, conventional supervision, including CoT demonstrations and final task rewards, encourages a correct prediction but does not specify which intermediate evidence should be retained or discarded. Consequently, increasing reasoning length or test-time computation may produce additional trajectories without reliably extracting the target-specific context required for grounding, as illustrated in Fig. 1b and Fig. 1c. The challenge is therefore to transform an extended reasoning path into a concise and sufficient representation before answer generation.

This perspective naturally connects multimodal grounding to the information bottleneck (IB) principle Tishby et al. (2000); Tishby & Zaslavsky (2015), which seeks an intermediate representation that discards non-predictive information while preserving information relevant to the target. From the IB principle, a key limitation of existing reasoning-based MLLMs is their insufficient ability to compress an extended reasoning path into the target-specific context required for localization. Building on this insight, we propose **Think-Compress-Answer (TCA)**, a novel reinforcement learning (RL) framework that decomposes multimodal grounding into three successive phases: free-form reasoning, target-aware compression, and answer generation conditioned on the compressed context. TCA introduces an explicit compressed context that regularizes the transition from extended reasoning to answer generation.

We optimize TCA through a four-stage curriculum based on Group Relative Policy Optimization (GRPO) Shao et al. (2024b). Stage 1 first establishes the structured TCA output protocol using a format reward DeepSeek-AI (2025). Stage 2 minimizes a variational upper bound on the conditional mutual information between the full reasoning path and its compressed context, encouraging the model to remove reasoning-specific information. To complement a potential uninformative compression, Stage 3 introduces a specialized MLLM judge that evaluates whether the compressed context remains concise, target-specific, and sufficient for predicting the answer. Finally, Stage 4 directly optimizes localization accuracy using a task-dependent grounding reward. This staged design separates competing objectives during training and provides a progressive curriculum from structured reasoning to target-preserving compression and, ultimately, accurate multimodal grounding.

We evaluate TCA on referring expression segmentation and referring video object segmentation across six benchmarks. Extensive experiments demonstrate that explicitly compressing intermediate reasoning consistently improves grounding performance, achieving comparable and even better zero-shot performance than supervised methods.

In summary, our contributions are as follows:

- We explore a fundamental limitation of extended reasoning in multimodal grounding and formulate it with the information bottleneck (IB) principle, showing that extended reasoning paths can retain excessive scene-level content that is irrelevant to compact grounding targets.

- We propose Think-Compress-Answer (TCA), a novel RL framework that introduces a target-aware compression stage between reasoning and answer generation, trained with a multi-stage GRPO-based optimization.
- Extensive experiments on two multimodal grounding tasks across six benchmarks demonstrate that TCA consistently outperforms CoT and test-time scaling baselines, while achieving zero-shot performance comparable to or even better than supervised methods on RES and R-VOS benchmarks.

2 Related Work

Chain-of-Thought (CoT) reasoning and multimodal grounding. CoT prompting improves reasoning by generating intermediate steps Wei et al. (2022); Kojima et al. (2022). Multimodal extensions adapt this to vision-language settings via duty-distinct prompting or intermediate spatial verification Zhang et al. (2024b); Zheng et al. (2023a); Shao et al. (2024a); Wu et al. (2025a). In multimodal grounding, this reasoning is integrated either *implicitly*, connecting LLMs to segmentation decoders via latent tokens Lai et al. (2024); Ren et al. (2024); Yan et al. (2024); Zheng et al. (2025), or *explicitly*, generating textual chains before predicting coordinates or masks for visual and audio-visual tasks Kao et al. (2026); Chowdhury et al. (2024); Liu et al. (2025c); You & Wu (2025); Huang et al. (2025); Xu et al. (2026); Zhao et al. (2025). However, extending explicit CoT chains often causes "overthinking" Chen et al. (2024a): the model drifts, producing unfaithful steps Turpin et al. (2023) or accumulating hallucinated content Huang et al. (2024). Standard mitigations like self-consistency Wang et al. (2023), process rewards Lightman et al. (2024), and test-time scaling OpenAI (2024); Snell et al. (2024); Zhang et al. (2025) address this but incur significant computational costs and still assume extensive reasoning is inherently beneficial. DR^{2}Seg He et al. (2026) verifies the self-containment of a target description via answer consistency across two reasoning rollouts. Unlike approaches relying on latent tokens, unconstrained text chains, or answer-consistency verification, our approach introduces a textual compression bottleneck that distills reasoning into a concise and faithful cue before the final grounding stage.

Reinforcement learning from AI feedback and GRPO. RLHF Ouyang et al. (2022) aligns LLMs with human preferences but is expensive to scale. RLAIF Lee et al. (2023); Bai et al. (2022) replaces human annotators with LLM judges, and the LLM-as-judge paradigm Zheng et al. (2023b) shows strong correlation with human evaluation. Group Relative Policy Optimization (GRPO) Shao et al. (2024b) simplifies RL-based training by eliminating the critic, computing advantages relative to a sampled group, making it robust to noisy rewards. Multi-stage RL DeepSeek-AI (2025) shows that sequential objective decomposition improves stability. We combine an LLM-judge reward with multi-stage GRPO, where each stage addresses a distinct requirement of bottleneck learning instead of a single end-to-end task reward.

Information bottleneck in deep learning. The Information Bottleneck (IB) principle Tishby et al. (2000) provides a theoretical framework for extracting compressed, task-relevant representations. Alemi et al. (2017) apply this to neural networks by optimizing variational bounds in continuous latent spaces, inspiring subsequent work on generalization and representation learning Shwartz-Ziv & Tishby (2017). However, applying IB to *discrete* natural language remains challenging since mutual information over text is not directly computable. We sidestep this by using IB as a design guide: we structure the model's output into an observable textual bottleneck, optimize a tractable variational bound on the compression term, and use an LLM judge as a proxy for target-relevant fidelity, enabling IB-inspired GRPO training.

3 Method

3.1 Motivation and overview

Given a text query $\mathbf{q}$ and another modality input $\mathbf{X}$ composed of observations from one or more modalities $\mathcal{M}$ (*e.g.*, image, video, audio, or their combinations), multimodal grounding aims to identify the target entity, event, or region in $\mathbf{X}$ that is semantically referred to by $\mathbf{q}$:

$$\mathbf{y} = \mathcal{F}(\mathbf{X}, \mathbf{q}), \tag{1}$$

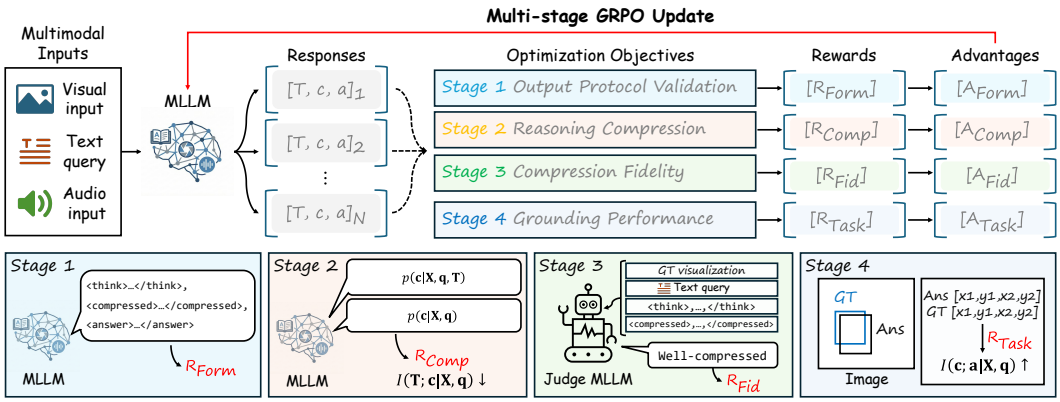


Figure 2: **Overview of Think-Compress-Answer.** The model generates N grouped responses $[\mathbf{T}, \mathbf{c}, \mathbf{a}]$ per input. Stage 1 establishes the required output protocol, Stage 2 encourages reasoning compression, Stage 3 uses a judge MLLM to promote compression fidelity and target relevance, and Stage 4 improves grounding accuracy. The stage-specific rewards are converted into group-relative advantages to progressively update the model.

where $\mathbf{y} \in \mathcal{Y}$ denotes the grounding target predicted by a grounding model $\mathcal{F}$ in a task-dependent output form $\mathcal{Y}$, such as a bounding box Wu et al. (2022), segmentation mask Lai et al. (2024), or center point Liu et al. (2025c).

Recent approaches Lai et al. (2024); Zhong et al. (2025); Liu et al. (2025c); Li et al. (2025) have defined the grounding model $\mathcal{F}$ as a combination of a multimodal large language model (MLLM) π and a decoder $\mathcal{D}$ that maps the output of π into the target grounding space $\mathcal{Y}$. In existing reasoning-based MLLMs, inference typically follows an autoregressive reasoning-to-answer process: the model generates extended reasoning paths $\mathbf{T} = \{t_l\}_{l=1}^{L}$, where L is the token length, produces an intermediate textual answer $\mathbf{a}$ (*i.e.*, the target bounding-box and center-point coordinates) conditioned on $\mathbf{T}$, and then generates the final segmentation mask $\mathbf{y}$:

$$\mathbf{T} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{q}), \quad \mathbf{a} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{q}, \mathbf{T}), \quad \mathbf{y} = \mathcal{D}(\mathbf{a}, \mathbf{X}). \tag{2}$$

With reinforcement learning (RL), such models are optimized to produce the reasoning paths $\mathbf{T}$ and the answer $\mathbf{a}$ that maximize task reward, encouraging $\mathbf{T}$ to retain information that is useful for predicting the intermediate target $\mathbf{a}$. In multimodal grounding, such reasoning remains important for identifying the correct grounding target. However, not all information contained in $\mathbf{T}$ is required for predicting $\mathbf{a}$. The key mismatch is that the intermediate target $\mathbf{a}$ is highly compact, whereas $\mathbf{T}$ often contains rich scene-level descriptions, relational analysis, and modality-wide contextual details that far exceed what is necessary for localization.

To address this challenge, we propose TCA (Think-Compress-Answer) to decompose the reasoning-to-answer process. We design an intermediate reasoning compression stage to extract a target-specific compressed context $\mathbf{c}$ motivated by the information bottleneck (IB) principle Tishby et al. (2000). Instead of directly mapping the full reasoning paths $\mathbf{T}$ onto the answer (*i.e.*, maximizing $I(\mathbf{T}; \mathbf{a})$), the model learns to retain only the information that is maximally predictive of $\mathbf{a}$, following the IB objective:

$$\min_{\pi} \quad \underbrace{I(\mathbf{T}; \mathbf{c} \mid \mathbf{X}, \mathbf{q})}_{\text{reasoning-specific information}} \quad -\beta \quad \underbrace{I(\mathbf{c}; \mathbf{a} \mid \mathbf{X}, \mathbf{q})}_{\text{task-relevant information}}, \tag{3}$$

where β controls the trade-off between the amount of preserved information from the reasoning paths, $I(\mathbf{T}; \mathbf{c} \mid \mathbf{X}, \mathbf{q})$, and task-relevant information, $I(\mathbf{c}; \mathbf{a} \mid \mathbf{X}, \mathbf{q})$. The overall procedure of TCA is illustrated in Fig. 2.

3.2 THINK-COMPRESS-ANSWER

TCA represents the extended reasoning paths $\mathbf{T}$, target-specific compressed context $\mathbf{c}$, and intermediate answer $\mathbf{a}$ as three explicit textual fields. In addition to (2), TCA introduces an intermediate

process for $\mathbf{c}$ as

$$\mathbf{c} \sim \pi(\cdot \mid \mathbf{X}, \mathbf{q}, \mathbf{T}). \tag{4}$$

The main challenge in training TCA is that no direct supervision is available for the optimal compressed context $\mathbf{c}$. Although $\mathbf{c}$ should be concise, derived from $\mathbf{T}$, and sufficient for predicting the correct grounding target, these properties are only indirectly constrained by the final grounding annotation $\mathbf{y}^*$. We therefore introduce four reward components, each of which targets a distinct property of TCA: a format reward that enforces a valid output structure; a compression-rate reward that removes reasoning-specific information from $\mathbf{c}$; a compression-fidelity reward that preserves the target-relevant evidence required for grounding; and a grounding reward that directly evaluates localization accuracy.

These rewards are optimized through a multi-stage GRPO curriculum Shao et al. (2024b) to avoid jointly optimizing competing objectives from the beginning or shortcut solutions, such as bypassing the bottleneck by encoding information directly in the answer while leaving the compressed context uninformative Alemi et al. (2017). At each training iteration, we sample N responses from the current rollout policy $\bar{\pi}$, compute the rewards R^n corresponding to each stage, and normalize the rewards to estimate a relative advantage for each response:

$$A^n = \frac{R^n - \text{mean}(\{R^1, \ldots, R^N)\})}{\text{std}(\{R^1, \ldots, R^N\})} \tag{5}$$

The model is then updated using the GRPO objective.

Stage 1 – Format reward. We first establish the desired output protocol using a simple format reward DeepSeek-AI (2025), encouraging the model to reliably separate free-form reasoning from the structured answer block. Let $\mathbf{o}$ denote the serialized output generated by π:

$$\mathbf{o} = \big[\texttt{<think>}, \mathbf{T}, \texttt{</think>}, \texttt{<compressed>}, \mathbf{c}, \texttt{</compressed>}, \texttt{<answer>}, \mathbf{a}, \texttt{</answer>}\big]. \tag{6}$$

We define the format reward R_{Form} as

$$R_{\text{Form}} = \mathbf{1}[\texttt{<think>}\ldots\texttt{</think>} \in \mathbf{o}] + \mathbf{1}[\texttt{<compressed>}\ldots\texttt{</compressed>} \in \mathbf{o}] + \mathbf{1}[\texttt{<answer>}\ldots\texttt{</answer>} \in \mathbf{o}]. \tag{7}$$

The format reward yields a score in $\{0, 1, 2, 3\}$, normalized to $[0, 1]$, and provides partial credit for outputs that satisfy only a subset of the required constraints.

Stage 2 – Compression-rate reward. In Stage 2, we encourage $\mathbf{c}$ to discard reasoning-path-specific information by minimizing the first term in (3). However, directly evaluating $I(\mathbf{T}; \mathbf{c} \mid \mathbf{X}, \mathbf{q})$ requires the intractable marginal $p(\mathbf{c} \mid \mathbf{X}, \mathbf{q})$. We therefore introduce a context-conditioned variational prior $\rho_\eta(\mathbf{c} \mid \mathbf{X}, \mathbf{q})$ that predicts the compressed context without observing $\mathbf{T}$. The conditional mutual information is upper-bounded Hou et al. (2025) as

$$I_\pi(\mathbf{T}; \mathbf{c} \mid \mathbf{X}, \mathbf{q}) \leq \mathbb{E}\big[D_{\text{KL}}\big(\pi(\mathbf{c} \mid \mathbf{X}, \mathbf{q}, \mathbf{T}) \,\|\, \rho_\eta(\mathbf{c} \mid \mathbf{X}, \mathbf{q})\big)\big]. \tag{8}$$

For a sampled trajectory, we estimate its information rate as

$$\hat{\mathcal{I}}_{\text{Rate}} = \log \bar{\pi}(\mathbf{c} \mid \mathbf{X}, \mathbf{q}, \mathbf{T}) - \log \rho_\eta(\mathbf{c} \mid \mathbf{X}, \mathbf{q}), \tag{9}$$

where $\bar{\pi}$ denotes the fixed rollout policy. For an autoregressive compressed context, each log-likelihood is computed by summing its token-level log-probabilities. The compression-rate reward is then defined as

$$R_{\text{Comp}} = \sigma(-\hat{\mathcal{I}}_{\text{Rate}}), \tag{10}$$

where $\sigma(\cdot)$ is a sigmoid function to force the reward to be $[0, 1]$. Maximizing R_{Comp} encourages $\mathbf{c}$ to contain less information that can only be recovered from the particular realization of $\mathbf{T}$. We instantiate the variational prior using the same rollout policy $\bar{\pi}$, *i.e.*, $\rho_\eta(\mathbf{c} \mid \mathbf{X}, \mathbf{q}) := \bar{\pi}(\mathbf{c} \mid \mathbf{X}, \mathbf{q})$. The parameters of $\bar{\pi}$ are held fixed while computing rewards and performing each GRPO update, not requiring separately optimized variational prior.

Stage 3 – Compression fidelity reward. Minimizing the information rate alone admits a degenerate solution in which the model produces an empty or overly generic compressed context (*e.g.*, "object" or "animal"), merely copies the input query, or generates an input-independent context Skalse et al.

Table 1: Referring expression segmentation on RefCOCO, RefCOCO+, and RefCOCOg. We report the performance (cIoU) evaluated on validation, testA, and testB splits for RefCOCO and RefCOCO+, and validation and test splits for RefCOCOg. * employs a frozen decoder (Dec.) only for mask prediction.

Method	MLLM	Dec.	RefCOCO			RefCOCO+			RefCOCOg	
			val	tA	tB	val	tA	tB	val	test
Supervised, Fine-Tuning										
Qwen2.5-Omni-7B Qwen Team (2025)	Qw2.5OM-7B	SAM2*	76.2	78.0	71.3	69.4	75.1	62.7	71.5	70.5
Supervised, Embedding-Guided										
LISA Lai et al. (2024)	LLaVA-7B	SAM	74.1	76.5	71.1	62.4	67.4	56.5	67.9	70.6
PixelLM Ren et al. (2024)	LLaVA-7B	Pix.Dec.	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5
GSVA Xia et al. (2024)	Vicuna-7B	SAM	76.4	77.4	72.8	64.5	67.7	58.6	71.1	72.0
GLaMM Rasheed et al. (2024)	Vicuna-7B	SAM	79.5	**83.2**	76.9	72.6	78.7	64.6	74.2	74.9
OMG-LLaVA Zhang et al. (2024a)	IntLM2-7B	SAM	78.0	80.3	74.1	69.1	73.1	63.0	72.9	72.9
F-LMM Wu et al. (2025b)	LLaVA-7B	SAM	75.2	79.1	71.9	63.7	71.8	54.7	67.1	68.1
Sa2VA Yuan et al. (2025)	IntVL2.5-8B	SAM2	**81.9**	–	–	**76.5**	–	–	**78.9**	–
UniPixel Liu et al. (2025b)	Qw2.5VL-7B	SAM2	80.8	83.0	**77.4**	75.3	**80.1**	**70.0**	76.4	**77.1**
Supervised, Reasoning-Guided										
SAM4MLLM Chen et al. (2024b)	LLaVA1.6-7B	SAM*	79.6	82.8	76.1	73.5	77.8	65.8	74.5	75.6
Seg-Zero Liu et al. (2025c)	Qw2.5VL-7B	SAM2*	–	80.3	–	–	76.2	–	–	72.6
DR^{2}Seg-7B He et al. (2026)	Qw2.5VL-7B	SAM2*	–	78.7	–	–	75.4	–	–	72.2
SAM-R1 Huang et al. (2025)	Qw2.5VL-7B	SAM2*	–	79.2	–	–	74.7	–	–	73.1
Zero-Shot, Reasoning-Guided										
Qwen2.5-Omni-7B Qwen Team (2025)	Qw2.5OM-7B	SAM2*	74.5	75.8	69.1	66.5	71.2	60.5	71.8	71.2
Seg-R1-7B You & Wu (2025)	Qw2.5VL-7B	SAM2*	74.3	78.7	67.6	62.6	70.9	57.9	71.0	71.4
VideoSeg-R1 Xu et al. (2026)	Qw2.5VL-7B	SAM2*	78.2	**82.3**	**75.1**	71.8	76.1	64.7	73.1	74.1
TCA (**Ours**)	Qw2.5OM-7B	SAM2*	**79.4**	81.5	73.6	**74.0**	**78.7**	**66.1**	**75.3**	**76.0**

(2022). We employ a specialized judge $\mathcal{J}$ (*i.e.*, MLLM) to evaluate whether the compressed context $\mathbf{c}$ is well-aligned with $\mathbf{T}$, while preserving sufficient information for grounding. We first generate $\mathbf{X}^*$ by visualizing the ground-truth (*e.g.*, segmentation mask, bounding box, or center point) on the input $\mathbf{X}$. The judge $\mathcal{J}$ takes $\mathbf{X}^*$, $\mathbf{q}$, $\mathbf{T}$, and $\mathbf{c}$ as inputs, and produces the fidelity reward as:

$$R_{\text{Fid}} = \mathcal{J}(\mathbf{X}^*, \mathbf{q}, \mathbf{T}, \mathbf{c}), \tag{11}$$

where $R_{\text{Fid}} \in \{0, 1\}$ returns 1 if $\mathbf{c}$ is "well-compressed" from $\mathbf{T}$ and semantically aligned with the ground-truth target highlighted on $\mathbf{X}^*$, and 0 otherwise. For more details on the prompt to generate the fidelity reward, please refer to the supplementary material.

In Stage 2 and 3, GRPO updates are applied primarily to the tokens in $\mathbf{c}$, while the distribution of $\mathbf{T}$ is stabilized using a KL constraint to the Stage 1 checkpoint. This isolates compression learning from changes in the reasoning path distribution.

Stage 4 – Grounding reward. Finally, we optimize a task-driven surrogate for the task-relevant term $I(\mathbf{c}; \mathbf{a} \mid \mathbf{X}, \mathbf{q})$ in (3) using a reward on the intermediate answer and resulting grounding target. For segmentation-based grounding tasks, $\mathbf{a}$ consists of the predicted bounding-box coordinates $\mathbf{b}$ and the center point $\mathbf{p}$ of the target. From the ground-truth label $\mathbf{y}^*$, we construct the corresponding intermediate ground truth $\mathbf{a}^*$, consisting of the bounding box $\mathbf{b}^*$ and center point $\mathbf{p}^*$. We define the task reward as the sum of the bounding-box IoU, bounding-box ℓ_1, and point ℓ_1 rewards Liu et al. (2025c):

$$R_{\text{Task}} = \mathbf{1}[\text{IoU}(\mathbf{b}, \mathbf{b}^*) > \tau_{\text{iou}}] + \mathbf{1}[\|\mathbf{b} - \mathbf{b}^*\|_1 < \tau_b] + \mathbf{1}[\|\mathbf{p} - \mathbf{p}^*\|_1 < \tau_p]. \tag{12}$$

The bounding-box IoU reward evaluates the overlap between $\mathbf{b}$ and $\mathbf{b}^*$, while the bounding-box and point ℓ_1 rewards evaluate their distances from the corresponding ground truth. GRPO updates in Stage 4 are applied to both $\mathbf{c}$ and $\mathbf{a}$, while the reasoning field remains constrained by the previous-stage policy. We set τ_{iou}=0.5, τ_b=10, and τ_p=100 in our experiments.

Overall, the four-stage optimization first establishes a valid TCA protocol, then learns a minimum sufficient compressed context, and finally improves grounding accuracy without allowing the compressed context to recover the full information content of the original reasoning paths.

Table 2: Referring video object segmentation on Ref-YouTube-VOS, Ref-DAVIS17, and MeViS. We report $\mathcal{J}\&\mathcal{F}$ evaluated on the validation set of each benchmark. * employs a frozen decoder (Dec.) only for mask prediction. † additionally employs an external XMem tracker for temporal mask propagation.

Method	MLLM	Dec.	RY-VOS	RD17	MeViS
Supervised, Fine-Tuning					
Qwen2.5-Omni-7B Qwen Team (2025)	Qw2.5OM-7B	SAM2*	69.3	76.8	53.3
Supervised, Embedding-Guided					
VISA Yan et al. (2024)	ChatUniVi-7B	SAM†	61.5	69.4	43.5
ViLLa Zheng et al. (2025)	Vicuna-7B	E2E	67.5	74.3	49.4
Sa2VA Yuan et al. (2025)	IntVL2.5-8B	SAM2	70.7	75.2	46.0
VideoLISA Bai et al. (2024)	LLaVA-Phi-3.8B	SAM2	63.7	68.8	44.4
UniPixel Liu et al. (2025b)	Qw2.5VL-7B	SAM2	71.0	76.4	**55.8**
Supervised, Reasoning-Guided					
VideoSeg-R1 Xu et al. (2026)	Qw2.5VL-7B	SAM2*†	**81.3**	**79.8**	55.3
Zero-Shot, Reasoning-Guided					
Qwen2.5-Omni-7B Qwen Team (2025)	Qw2.5OM-7B	SAM2*	63.5	65.2	49.1
PARSE-VOS Zhao et al. (2025)	Llama3-8B	SAM2*	72.1	75.5	52.4
Omni-R1 Zhong et al. (2025)	Qw2.5OM-7B	SAM2*	71.0	77.6	52.3
TCA (**Ours**)	Qw2.5OM-7B	SAM2*	**72.3**	**78.9**	**55.8**

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate TCA on two multimodal grounding tasks across six benchmarks, including the RefCOCO series Yu et al. (2016); Mao et al. (2016) for referring expression segmentation (RES), Ref-YouTube-VOS (RY-V) Seo et al. (2020), Ref-DAVIS17 (RD17) Khoreva et al. (2018), and MeViS Ding et al. (2023) for referring video object segmentation (R-VOS). We train TCA on a mixed set of data consisting of 842 samples from RefAVS Wang et al. (2024) and 1,258 videos from ReVOS Yan et al. (2024) for one epoch. Note that we *do not* use the RES and R-VOS benchmarks during training, but only for zero-shot performance evaluation. We provide details on each benchmark and results on the benchmarks used for training in the supplementary material.

Baselines. We consider two types of baselines according to the output of the MLLM π: *Embedding-guided methods* Lai et al. (2024); Ren et al. (2024); Rasheed et al. (2024); Yuan et al. (2025); Liu et al. (2025b) in which the decoder $\mathcal{D}$ takes a grounding latent token from π. These baselines train the MLLM and the decoder simultaneously for multimodal grounding, typically showing better performance than the following reasoning-guided baselines. *Reasoning-guided methods* Liu et al. (2025c); Huang et al. (2025); He et al. (2026); You & Wu (2025); Xu et al. (2026); Zhao et al. (2025); Zhong et al. (2025) employ an off-the-shelf pre-trained decoder (*e.g.*, SAM2 Ravi et al. (2024)) so that the MLLM π provides the intermediate answer $\mathbf{a}$ in a form required by the decoder. TCA follows the reasoning-guided setting.

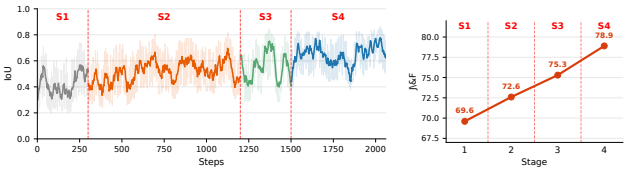
Backbone. We employ Qwen2.5-Omni-7B Qwen Team (2025) as an MLLM π and SAM2 Ravi et al. (2024) as a decoder $\mathcal{D}$. The MLLM predicts bounding-box coordinates and a center point, collectively denoted as $\mathbf{a}$, and the decoder takes $\mathbf{a}$ as input to produce a segmentation mask $\mathbf{y}$. Although we mainly use Qwen2.5-Omni-7B and SAM2, our TCA can be readily adapted to other MLLMs and decoders, such as SAM Kirillov et al. (2023) and SAM3 Carion et al. (2025). We provide the results with a smaller MLLM (*i.e.*, Qwen2.5-Omni-3B) and different decoders (*i.e.*, SAM and SAM3) in the supplementary material. In our framework, we apply low-rank adapters (LoRA) Hu et al. (2022) to the LLM of π and train only the LoRA layers, while freezing all the pre-trained parameters in π and $\mathcal{D}$. Please refer to the supplementary material for additional implementation details.

4.2 Referring Expression Segmentation

Given an image and a natural language referring expression, referring expression segmentation (RES) aims to generate a pixel-wise mask for the object described by the expression.

Table 3: Performance with respect to variants of the stages on R-VOS benchmarks. We report $\mathcal{J}\&\mathcal{F}$ evaluated on the validation set of each benchmark.

S1	S2	S3	S4	RY-V	RD17	MeViS
×	×	×	×	63.5	65.2	49.1
✓	×	×	×	65.6	69.6	50.6
✓	✓	×	×	69.0	72.6	52.3
✓	✓	✓	×	71.1	75.3	53.0
✓	×	×	✓	68.6	72.8	51.5
✓	✓	×	✓	70.8	76.1	53.8
✓	✓	✓	✓	**72.3**	**78.9**	**55.8**



(a) Grounding IoU (b) Zero-shot $\mathcal{J}\&\mathcal{F}$

Figure 3: (a) Grounding IoU over training steps across all four stages. (b) $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS17 after each stage.

4.2.1 Results.

Tab. 1 shows RES performance evaluated on three benchmarks. The results show that our TCA achieves comparable performance to the fully-supervised approaches Liu et al. (2025b); Chen et al. (2024b); Yuan et al. (2025) or even better performance than most existing methods Rasheed et al. (2024); Zhang et al. (2024a) across all evaluation splits. Compared with Seg-Zero Liu et al. (2025c) on the three reported splits, TCA improves cIoU by 1.2, 2.5, and 3.4 points on RefCOCO testA, RefCOCO+ testA, and RefCOCOg test, respectively.

4.3 Referring Video Object Segmentation

The goal of referring video object segmentation (R-VOS) is to segment a referred object instance in each frame, producing temporally consistent masks across a given video sequence. R-VOS extends RES by adding motion, occlusion, and temporal appearance changes.

4.3.1 Results.

In Tab. 2, we compare TCA with state-of-the-art methods on three R-VOS benchmarks. Despite being evaluated in a zero-shot setting, TCA ties for the best performance on MeViS and achieves the second-best performance on RD17. In particular, TCA achieves state-of-the-art zero-shot performance while outperforming most supervised methods Yan et al. (2024); Zheng et al. (2025); Yuan et al. (2025); Bai et al. (2024); Liu et al. (2025b), showing $\mathcal{J}\&\mathcal{F}$ scores of 72.3, 78.9, and 55.8 on RY-V, RD17, and MeViS, respectively. Compared with the zero-shot backbone baseline, *i.e.*, Qwen2.5-Omni Qwen Team (2025), TCA yields substantial gains of 8.8, 13.7, and 6.7 points across the benchmarks.

4.4 Additional Analysis

4.4.1 Optimization stage.

In Tab. 3, we conduct stage-wise ablations under an equal optimization budget and evaluate $\mathcal{J}\&\mathcal{F}$ on the R-VOS validation sets. The comparison between the models without any optimization and with only the format reward shows that a simple format reward consistently improves performance across RY-V, RD17, and MeViS. The model trained up to Stage 2 further improves the scores to 69.0, 72.6, and 52.3, demonstrating the effectiveness of the compression step. Stages 3 and 4 yield additional performance gains, reaching 72.3, 78.9, and 55.8 with all four stages. Skipping the intermediate stages (rows 5–6) consistently leads to lower performance than the full progressive optimization, confirming the effectiveness of the progressive reward schedule.

In Fig. 3, we illustrate the grounding IoU on ReVOS and RefAVS according to the training steps, and the zero-shot grounding performance on Ref-DAVIS17 evaluated after each training stage. The results show that the grounding IoU and the grounding performance are monotonically increased by going through training. Interestingly, the grounding IoU increases even before optimizing the model with the grounding reward. This result suggests that optimizing the model with compression-rate and compression-fidelity rewards in our TCA can directly improve the grounding performance.



Figure 4: Qualitative comparisons for referring video object segmentation on Ref-YouTube-VOS.

4.4.2 Qualitative results.

In Fig. 4, we present qualitative comparisons between TCA and Omni-R1 Zhong et al. (2025) on Ref-YouTube-VOS. The example in Fig. 4 shows that, although both methods reason over broad scene context, only TCA converges to the correct target, indicating that our framework enables the model to yield target-focused reasoning for grounding. We provide additional analyses for TCA's components, including the backbone, decoders, and optimization stages, together with further qualitative examples and a failure case analysis in the supplementary material.

5 Conclusion

We presented Think-Compress-Answer (TCA), a novel information-bottleneck-inspired RL framework that makes explicit reasoning compatible with multimodal grounding. We optimize TCA with a staged GRPO using practical rewards, progressively enforcing a valid output protocol, compression rate, compression fidelity, and end-task grounding performance. Without architectural modification or dense intermediate annotations, TCA achieves comparable and even better zero-shot performance than the supervised methods on RES and R-VOS benchmarks.

5.1 Broader impact.

TCA offers a practical framework to study the interaction between long-form reasoning and multimodal grounding. By explicitly optimizing compression via staged RL, TCA enables the community to (1) diagnose when extended reasoning becomes counterproductive, (2) measure information retained in the compressed context, and (3) align inference with compact grounding objectives. This perspective shifts the focus from "more reasoning" to "selective reasoning," informing future work on test-time scaling, information bottlenecks, and multimodal evaluation. The release of our training recipes is expected to facilitate ablations on reward design, judge selection, and compression capacity, as well as standardized comparisons across datasets and modalities.

AI Use Statement

We used generative AI tools to assist with experimental code assistance and LaTeX formatting. All AI-assisted code and manuscript formatting were reviewed and verified by the authors. We take responsibility for the final content of this work, including all experimental results, text, and artifacts.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. *ICLR*, 2017.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *NeurIPS*, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Xingyu Chen et al. Do not think that much for 2+3=? on the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*, 2024a.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. SAM4MLLM: Enhance multi-modal large language model for referring expression segmentation. *ECCV*, 2024b.

Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. *ECCV*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. *ICCV*, 2023.

Yulin He, Wei Chen, Zhikang Jian, Tianhang Guo, Wenjuan Zhou, Minglong Li, Shaowu Yang, and Wenjing Yang. DR^{2}Seg: Decomposed two-stage rollouts for efficient reasoning segmentation in multimodal large language models. *arXiv preprint arXiv:2601.09981*, 2026.

Yue Hou, He Zhu, Ruomei Liu, Yingke Su, Junran Wu, and Ke Xu. Redundancy-aware test-time graph out-of-distribution detection. *NeurIPS*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. SAM-R1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. *NeurIPS*, 2025.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Computing Surveys*, 2024.

Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. CoT-RVS: Zero-shot chain-of-thought reasoning segmentation for videos. *ICLR*, 2026.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. *ACCV*, 2018.

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. *ICCV*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, 2022.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. *CVPR*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *EMNLP*, 2023.

Yifan Li, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Xin Wang, and Yanwei Fu. ReVSeg: Incentivizing the reasoning chain for video segmentation with reinforcement learning. *arXiv preprint arXiv:2512.02835*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *ICLR*, 2024.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? Assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Shan Ying, and Chang Wen Chen. UniPixel: Unified object referring and segmentation for pixel-level visual reasoning. *NeurIPS*, 2025b.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025c.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. *CVPR*, 2016.

OpenAI. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Qwen Team. Qwen2.5-Omni: Technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. GLaMM: Pixel grounding large multimodal model. *CVPR*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. *CVPR*, 2024.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. *ECCV*, 2020.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *NeurIPS*, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *NeurIPS*, 2022.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *ITW*, 2015.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *NeurIPS*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ICLR*, 2023.

Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. *ECCV*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. *CVPR*, 2022.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025a.

Size Wu et al. F-LMM: Grounding frozen large multimodal models. *CVPR*, 2025b.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: Generalized segmentation via multimodal large language models. *CVPR*, 2024.

Zishan Xu, Yifu Guo, Yuquan Lu, Fengyu Yang, Junxin Li, and Lihua Cai. VideoSeg-R1: Reasoning video object segmentation via reinforcement learning. *AAAI*, 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. *ECCV*, 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. *ECCV*, 2016.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with LLaVA for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *ICCV*, 2025.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Change Loy Chen, and Shuicheng Yan. OMG-LLaVA: Bridging image-level, object-level, pixel-level reasoning and understanding. *NeurIPS*, 2024a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024b. arXiv preprint arXiv:2302.00923.

Bingrui Zhao, Lin Yuanbo Wu, Xiangtian Fan, Deyin Liu, Lu Zhang, Ruyi He, Jialie Shen, and Ximing Li. Unleashing Hierarchical Reasoning: An LLM-Driven Framework for Training-Free Referring Video Object Segmentation. *arXiv preprint arXiv:2509.05751*, 2025.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *NeurIPS*, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *NeurIPS*, 2023b.

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. ViLLa: Video reasoning segmentation with large language model. *ICCV*, 2025.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-R1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *NeurIPS*, 2025.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *ICLR*, 2024.