META-COGGRAN: REASONING AS ONLINE LATENT COGNITIVE STATE EVOLUTION FOR MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in multimodal large language models (MLLMs) has largely depended on CLIP-based visual encoders. While these models excel at capturing high-level semantic alignment, they often struggle to preserve fine-grained visual evidence. By comparison, the recent DINOv3 has exhibited emergent capabilities in fine-grained pixel-level perception, but still lacks the higher-level semantic abstraction required for alignment with LLMs and complex multimodal reasoning. These complementary limitations reveal a fundamental bottleneck: existing MLLMs lack a mechanism to dynamically reconcile fine-grained perception with high-level semantics during reasoning. We argue that the key challenge lies not merely in obtaining richer visual representations, but in enabling perceptual evidence and semantic knowledge to continuously interact, update, and converge during reasoning. This raises a central question: ***How can fine-grained perception and high-level semantics be dynamically integrated into evolving cognitive states for multimodal reasoning?*** To address this challenge, we propose `Meta-CogGran`, which formulates multimodal reasoning as online latent cognitive state evolution. It iteratively updates structured cognitive states through interactions with visual evidence and semantic memory, while a semantic equilibrium mechanism terminates reasoning once the state becomes stable and self-consistent. Extensive experiments demonstrate approximately $\mathbf{45}\%\uparrow$ improvement in accuracy and $\mathbf{30}\%\downarrow$ reduction in hallucination. Code is available at Meta-CogGran.

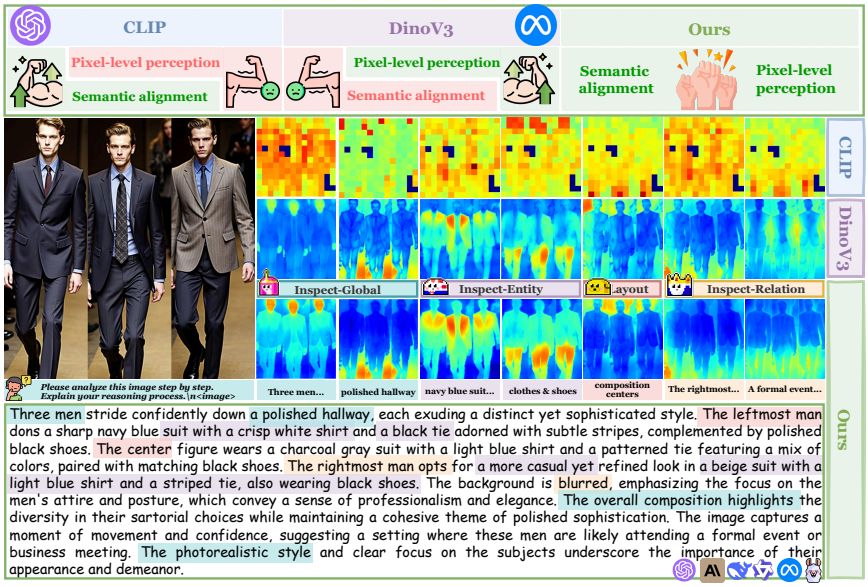


Figure 1: Schematic of `Meta-CogGran`, which models multimodal reasoning as cognitive state evolution by integrating fine-grained visual evidence with semantics across Global, Entity, Layout, and Relation perspectives.

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.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have achieved remarkable progress in visual reasoning (Dong et al., 2025; Yang et al., 2026b; Wang et al., 2026b; Kil et al., 2024), image captioning (Lai et al., 2025; Chen et al., 2025; Yang et al., 2026a), and cross-modal understanding (Zhang et al., 2025b; Li et al., 2025b). Despite these successes, existing MLLMs still predominantly rely on static visual representations, where reasoning is ultimately reduced to processing sequences of visual tokens extracted by vision encoders (Tong et al., 2024; Yao et al., 2026; Fang et al., 2026). Although increasingly powerful visual enco ders provide rich perceptual features, the majority of MLLMs implicitly assume that reasoning can emerge solely from token interactions (Bigverdi et al., 2025; Kuang et al., 2025). Such a representation-centric paradigm overlooks an essential property of intelligent cognition: ***reasoning is inherently a dynamic process that continuously revises internal beliefs according to newly acquired evidence.***

Recent studies have attempted to improve reasoning by introducing intermediate reasoning tokens (Xie et al., 2025; Yao et al., 2025), chain-of-thought (CoT) supervision (Shao et al., 2024; Wu et al., 2026; Jiang et al., 2025a; Zhao et al., 2025), and iterative decoding strategies (Yan et al., 2026; Liu et al., 2026). Nevertheless, these approaches still externalize reasoning as token generation. Internal cognitive variables, such as belief states, uncertainty, reasoning progress, and semantic consistency, remain largely implicit (Li et al., 2025c). Consequently, current MLLMs fail to track known content, uncertainties and semantic interpretive stability. Cognitive science views reasoning as latent cognitive state transitions, not symbolic output generation (Pan et al., 2025; Jaffe et al., 2023; Hitzler et al., 2022; Zhao et al., 2023). Human cognition progressively evolves from coarse perception to semantic understanding through iterative cycles of inspection, verification, and revision until a stable interpretation is reached (Spivey, 2023). Inspired by this perspective, we argue that multimodal reasoning should be formulated as online cognitive state evolution in latent semantic space instead of autoregressive token generation.

In this paper, we propose `Meta-CogGran`, a novel framework that models multimodal reasoning as online latent cognitive state dynamics (as shown in Figure 1). Instead of directly appending retrieved memory tokens to visual representations, `Meta-CogGran` constructs a structured cognitive state and iteratively refines it through a sequence of cognitive transitions. At each reasoning step, the model explicitly decides which cognitive action should be performed, including inspecting entities, revisiting spatial layouts, verifying existing beliefs, revising conflicting interpretations, or exploring uncertain evidence. The cognitive state continuously interacts with both visual observations and semantic memory, producing a reasoning trajectory that gradually converges toward semantic equilibrium.



❶ **Latent Cognitive State** → explicitly models current semantic beliefs & uncertainty
❷ **Cognitive Actions** → determines what reasoning operation should be executed next
❸ **Controlled State Transition** → updates beliefs using visual and memory evidence
❹ **Semantic Equilibrium** → adaptively terminates reasoning when cognition stable

Extensive experiments across diverse benchmarks demonstrate that `Meta-CogGran` consistently improves multimodal reasoning accuracy while significantly reducing hallucination. The results indicate that reasoning as latent cognitive state evolution provides a more scalable and reliable paradigm than conventional token-based reasoning.

The contributions of this paper are summarized as follows:

① We propose `Meta-CogGran`, a novel framework that formulates multimodal reasoning as online latent cognitive state evolution, moving beyond conventional token-centric reasoning paradigms.

② We introduce a cognitive dynamical system consisting of latent cognitive states, action-conditioned transitions, and semantic equilibrium, enabling adaptive and interpretable reasoning trajectories.

③ Extensive experiments demonstrate the proposed `Meta-CogGran` consistently improves reasoning performance and substantially reduces hallucination, highlighting the effectiveness of state-level reasoning for multimodal intelligence.

2 RELATED WORK

Structured Visual Representation. Early vision-language encoders such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) primarily learn globally aligned visual semantics through

large-scale image-text supervision, recent studies increasingly emphasize structured abstractions For instance, M3STR (Zhang et al., 2025a) evaluates MLLMs' ability to understand graph-structured visual knowledge, while Struct2D (Zhu et al., 2026a) demonstrates that object-centric layouts and metadata can effectively support spatial reasoning. Similarly, VPP-LLaVA (Tang et al., 2026) introduces explicit positional structures for visual grounding, and VG-LLM (Zheng et al., 2026) incorporates geometric priors extracted from videos to facilitate 3D scene understanding. Beyond spatial structures, FOCUS (Zhong et al., 2026), MOON (Zhang et al., 2026c), and Quality Token (Jiang et al., 2025b) further exploit structured visual cues to improve fine-grained perception, region localization, and subtle semantic discrimination. More recently, Retrv-R1 (Zhu et al., 2026b) shows that compressed and reasoning-oriented visual structures can serve as efficient intermediates for multimodal retrieval. *Existing works emphasize structured representations but largely treat them as static or one-step reasoning cues, without modeling their evolution throughout reasoning.*

Token-Based Reasoning in MLLMs. Token-based reasoning has become a prevalent paradigm for improving reasoning in large language models and MLLMs. Early approaches, including Chain-of-Thought (CoT) (Wei et al., 2022) and Tree-of-Thoughts (ToT) (Yao et al., 2023), enhance reasoning by generating intermediate reasoning tokens and exploring multiple reasoning paths. Recent advances extend this paradigm to multimodal settings through preference optimization and structured reasoning supervision, as exemplified by MPO (Wang et al., 2024), MindOmni (Xiao et al., 2026), and AtomThink (Xiang et al., 2026). Meanwhile, token-centric frameworks further enrich token functionality for spatial grounding (Ren et al., 2026), adaptive perception (Yu et al., 2025), dynamic visual resolution (Feng et al., 2026), and efficient token scheduling (Zhang et al., 2026d; Xing et al., 2026). *Despite richer reasoning traces and token representations, existing methods still equate reasoning with token generation, leaving the model's cognitive state and reasoning progress implicit.*

Reasoning as State Evolution. Recent research increasingly views intelligent reasoning as a process of state evolution rather than token generation. Cognitive science characterizes cognition as transitions between internal computational states (Barron et al., 2023). Search-based frameworks such as RStar-Math (Guan et al., 2025) model reasoning as iterative exploration of intermediate states, while representational analyses reveal that successful reasoning is accompanied by continuous transitions in latent representations (Zhang et al., 2026e). Related formulations further characterize intelligent behavior through state updates, memory evolution, and trajectory optimization in latent state spaces (Wang et al., 2026a; Zhang et al., 2026b). This state-centric perspective has gradually extended to multimodal intelligence. Recent works emphasize temporal state transitions in sequential visual reasoning (Villegas et al., 2026), continual modality evolution (Yu et al., 2024), and progressive reasoning evolution in multimodal learning (Luo et al., 2025; Lin et al., 2025). Although prior works may define state-like variables, transition-like updates, or both, they still do not unify multimodal reasoning as an explicit online cognitive state evolution process. ***we formulate reasoning as structured latent state initialization, action-conditioned transition, and adaptive convergence toward semantic equilibrium.***

3 METHODS

Overview. As shown in Figure 2, we propose `Meta-CogGran`, which models reasoning as a trajectory in a latent cognitive state space. Given an image-question pair (I, T), the model first constructs a structured cognitive state from visual observations and semantic memory, then iteratively updates this state through a sequence of cognitive transitions until semantic equilibrium is reached. Formally, let

$$\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{M}} * g,\ \boldsymbol{\mathcal{M}} * e,\ \boldsymbol{\mathcal{M}} * l,\ \boldsymbol{\mathcal{M}} * r \tag{1}$$

denote the Global, Entity, Layout, and Relation memory banks inherited from Meta-MemGran. Instead of directly appending retrieved memory tokens to the visual sequence, we treat these banks as an external cognitive workspace that continuously supports state evolution. The overall reasoning process is formulated as

$$\begin{aligned} \mathbf{F} = \boldsymbol{\mathcal{E}}_v(I), \quad \mathbf{z}_0 = \mathrm{InitCog}(\mathbf{F}, T, \boldsymbol{\mathcal{M}}), \\ \mathbf{z}_{t+1} = \mathbf{z} * t + \Delta\mathbf{z} * t, \quad \Delta\mathbf{z}_t = \boldsymbol{\mathcal{G}}(\mathbf{z}_t, \mathbf{F}, \boldsymbol{\mathcal{M}}, T), \end{aligned} \tag{2}$$

where $\mathbf{F}$ denotes visual features extracted by the vision encoder and $\mathbf{z}_t$ denotes the latent cognitive state at reasoning step t. Reasoning terminates when the cognitive dynamics reach a stable semantic equilibrium, $\mathbf{z}^* = \arg, \mathrm{equilibrium}\, \mathbf{z}_t$, after which the equilibrium state $\mathbf{z}^*$ is passed to the MLLMs

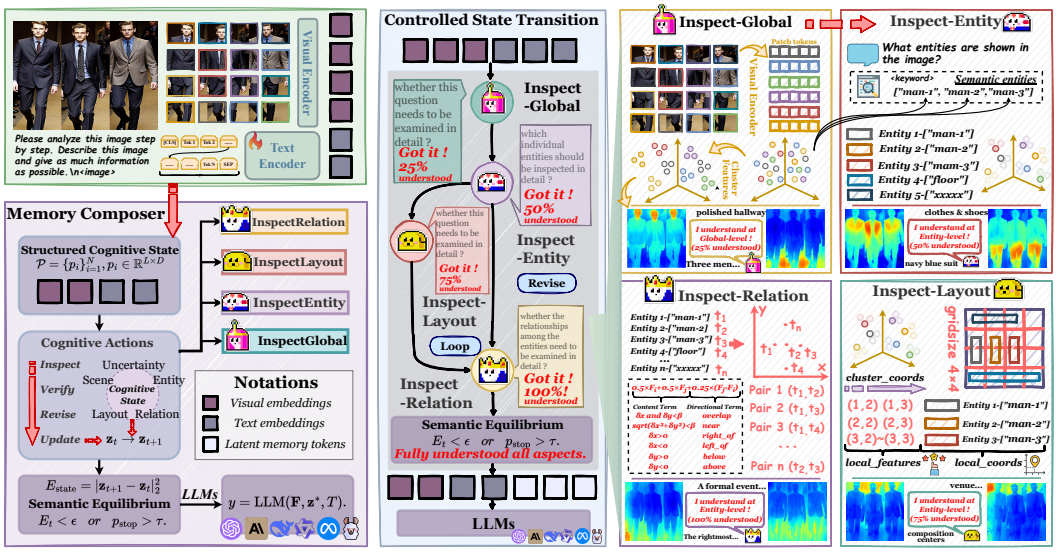


Figure 2: `Meta-CogGran` framework. The model iteratively evolves a memory-augmented latent cognitive state through action-conditioned transitions until semantic equilibrium is reached, and then uses the equilibrium state for answer generation.

for answer generation. Consequently, `Meta-CogGran` transforms memory-augmented reasoning from a static retrieval paradigm into an online cognitive dynamical system.

3.1 Latent Cognitive State

The central component of `Meta-CogGran` is a structured latent cognitive state that explicitly represents the model's current understanding of the visual world. Unlike conventional hidden states that entangle perception and reasoning into a single vector sequence, we decompose cognition into several semantically interpretable components:

$$\mathbf{z}_t = [\mathbf{s} * t, \mathbf{e} * t, \mathbf{l} * t, \mathbf{r} * t, \mathbf{u}_t], \tag{3}$$

where $\mathbf{s} * t$ denotes scene-level understanding, $\mathbf{e} * t$ denotes entity beliefs, $\mathbf{l} * t$ denotes spatial-layout beliefs, $\mathbf{r} * t$ denotes relational beliefs, $\mathbf{u}_t$ denotes the uncertainty landscape. Intuitively, $\mathbf{z}_t$ describes not only what the model currently believes, but also where uncertainty remains. Therefore, the cognitive state acts as a dynamic semantic representation rather than a purely neural hidden variable.

Online Cognitive Initialization. Given visual patch features $\mathbf{F}$, we first construct a compact set of semantic anchors through online clustering, each cluster token summarizes a semantically coherent visual region. Using cluster tokens and their spatial coordinates, we retrieve memory evidence from the four semantic banks,

$$\mathbf{C} = \mathbf{c} * 1, \ldots, \mathbf{c} * K, \quad \mathbf{M}_0 = \mathrm{Retrieve}(\mathbf{C}, \boldsymbol{\mathcal{M}}), \tag{4}$$

which serves as the initial cognitive substrate. The state is then obtained by integrating visual observations and memory evidence, $\mathbf{z}_0 = \mathrm{LayerNorm}\Big(\mathbf{M}_0 + \mathrm{Pool}(\mathbf{F})\Big)$. Thus, reasoning becomes the process of updating semantic beliefs rather than generating intermediate text tokens.

3.2 Cognitive Transition Operator

After initialization, `Meta-CogGran` performs iterative reasoning through a cognitive transition operator. Instead of treating refinement as a generic recurrent update, we explicitly model reasoning as a sequence of cognitive actions. At each step, the model first determines which cognitive operation should be executed. We define a discrete action space

$$a_t \in \texttt{InspectGlobal}, \texttt{InspectEntity}, \texttt{InspectLayout}, \texttt{InspectRelation}, \texttt{Verify}, \texttt{Revise}, \texttt{Stop}, \texttt{InspectUncertainty}. \tag{5}$$

Given the current state, an action policy predicts the next reasoning operation: $p(a_t) = \pi_a(\mathbf{z}_t, T)$. As listed in Table 1, different actions induce different memory access:

Table 1: Eight actions for iterative visual belief refinement.

Action	Description
`InspectGlobal`	Revisits global scene semantics.
`InspectEntity`	Focuses on object identities and attributes.
`InspectLayout`	Refines spatial structure.
`InspectRelation`	Updates inter-object relations.
`Verify`	Checks consistency among beliefs.
`Revise`	Corrects conflicting interpretations.
`InspectUncertainty`	Explores unresolved regions.
`Stop`	Requests reasoning termination.

The selected action produces an action-conditioned working memory $\mathbf{m}_t = \mathrm{Compose}(a_t, \boldsymbol{\mathcal{M}})$, which guides subsequent state update. To gather new evidence, the current cognitive state attends both to visual observations and semantic memory, and the transition module then predicts a state variation, and followed by a gated update. Formally, this process is defined as:

$$\begin{aligned} \mathbf{p}_t = \mathrm{Attn} * v(\mathbf{z} * t, \mathbf{F}), \quad \mathbf{h}_t = \mathrm{Attn} * m(\mathbf{z} * t, \mathbf{m}_t), \\ \Delta\mathbf{z}_t = \Phi(\mathbf{z} * t, \mathbf{p} * t, \mathbf{h}_t), \quad \mathbf{g}_t = \sigma\Big(W_g[\mathbf{z} * t; \mathbf{p} * t; \mathbf{h}_t]\Big), \quad \mathbf{z}_{t+1} = \mathbf{z} * t + \mathbf{g} * t \odot \Delta\mathbf{z}_t. \end{aligned} \tag{6}$$

This design yields a controlled cognitive evolution process. Rather than replacing previous beliefs at every step, the model gradually revises its interpretation according to newly acquired evidence, thereby producing a stable reasoning trajectory.

3.3 Semantic Equilibrium

A fundamental question in iterative reasoning is determining when the reasoning process should terminate. Existing recurrent reasoning approaches typically employ a fixed number of refinement steps or rely on simple convergence criteria such as $|\mathbf{z}_{t+1} - \mathbf{z}_t| < \epsilon$. However, cognitive reasoning is not merely a numerical optimization process. A stable reasoning state should simultaneously satisfy three conditions: ❶ the cognitive state no longer changes significantly, ❷ the model is no longer uncertain about the next reasoning action, ❸ the current interpretation is consistent with both visual evidence and semantic memory. Therefore, we formulate reasoning termination as a semantic equilibrium problem rather than a conventional convergence problem. **Semantic Equilibrium Energy.** Given the cognitive trajectory $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_t$, we define an equilibrium energy, and where each term captures a different source of cognitive instability. The first term measures state variation:

$$E_t = \lambda_1 E_{\text{state}} + \lambda_2 E_{\text{action}} + \lambda_3 E_{\text{contr}}, \quad E_{\text{state}} = |\mathbf{z}_{t+1} - \mathbf{z}_t|_2^2. \tag{7}$$

A small value indicates that the cognitive representation has become stable and additional reasoning steps provide diminishing updates. The second term measures action uncertainty: $E_{\text{action}} = H\big(p(a_t)\big)$, where $H(\cdot)$ denotes Shannon entropy and $p(a_t)$ is the action distribution predicted by the cognitive policy. When action entropy remains high, the model is still uncertain whether it should inspect entities, verify relations, revise previous beliefs, or continue exploring unresolved evidence. Conversely, low entropy indicates a confident reasoning direction. The third term evaluates semantic contradiction, intuitively, this quantity measures inconsistency between the current cognitive belief, visual observations, and semantic memory. In practice, we implement it using a lightweight contradiction estimator, let $\bar{\mathbf{z}}_t, \bar{\mathbf{p}}_t, \bar{\mathbf{m}}_t$ denote pooled representations of the current state, visual evidence, and memory evidence, respectively. The overall semantic equilibrium energy becomes

$$\begin{aligned} E_{\text{contr}} = \mathrm{Contradiction}(\mathbf{z}_t, \mathbf{F}, \boldsymbol{\mathcal{M}}). \quad E_{\text{contr}} = \sigma\Big(W_c[\bar{\mathbf{z}}_t; \bar{\mathbf{p}}_t; \bar{\mathbf{m}}_t]\Big), \\ E_t = \lambda_1 |\mathbf{z}_{t+1} - \mathbf{z}_t| * 2^2 + \lambda_2 H(p(a_t)) + \lambda_3 E * \text{contr}. \end{aligned} \tag{8}$$

Equilibrium-Aware Halting. Beyond energy minimization, we explicitly learn a stopping policy that predicts whether the reasoning process has reached semantic equilibrium. A lightweight equilibrium head produces a stopping probability $p_{\text{stop}} = \sigma(W_h \bar{\mathbf{z}}_t)$, where $\bar{\mathbf{z}}_t$ denotes the pooled cognitive state. Reasoning terminates when $E_t < \epsilon \quad or \quad p_{\text{stop}} > \tau$. The first condition corresponds to equilibrium in the cognitive dynamics, while the second condition corresponds to an explicit learned belief that further reasoning is unnecessary. Consequently, `Meta-CogGran` does not reason for a predetermined number of iterations. Instead, it performs adaptive cognitive evolution until a stable semantic interpretation emerges. **Equilibrium State for Answer Generation.** Let $\mathbf{z}^* = \mathbf{z}_{t_{\text{eq}}}$ denote the first state satisfying the equilibrium criterion. Rather than decoding from intermediate reasoning states, only the equilibrium state is exposed to the MLLMs:

$$\mathbf{y} = \mathrm{LLM}(\mathbf{F}, \mathbf{z}^*, T). \tag{9}$$

This design separates reasoning from language generation. The cognitive subsystem is responsible for constructing a stable semantic interpretation of the visual world, whereas the language model only

verbalizes the final equilibrium belief. From this perspective, multimodal reasoning can be viewed as searching for an attractor in latent cognitive space, and answer generation becomes a readout operation from the resulting equilibrium state.

4 EXPERIMENTS

Experiment Setup ❶ **Benchmarks and Evaluation Setting.** We evaluate `Meta-CogGran` on five benchmarks spanning multiple multimodal scenarios. The evaluation covers **VQA** with SEED-Bench (Li et al., 2023) and A-OKVQA (Schwenk et al., 2022), **image captioning** with CC12M_and_Imagenet21K_Recap (Changpinyo et al., 2021), **visual reasoning** with FLUX-Reason (Fang et al., 2025), and **medical vision-language understanding** with SurgVLM (Zeng et al., 2025). These datasets are used to study how different visual encoders affect MLLM performance across heterogeneous domains and task formats. ❷ **Zero-Shot OOD Protocol.** FLUX-Reason is the *only* benchmark involved in model training. SEED-Bench, A-OKVQA, CC12M_and_Imagenet21K_Recap, and SurgVLM are exclusively employed for **zero-shot out-of-distribution (OOD) evaluation**. ❸ **Multimodal Evaluation with GPT-4o.** We employ complementary automatic metrics to assess response quality. Following prior work on hallucination evaluation for MLLMs (Sun et al., 2024; Liu et al., 2024; Jiang et al., 2026; Zhang et al., 2026a; He et al., 2026; Mao et al., 2026; Li et al., 2025a; 2024), GPT-4o serves as an evaluator for **semantic accuracy**, **hallucination**, and **response granularity**. Following Mao et al. (Mao et al., 2026), hallucination is quantified on a 0-80 severity scale, where 0 denotes fully image-grounded content and 80 indicates the most severe visually unsupported generation. We report the mean score over all valid samples. The scoring criteria are summarized in *Appendix Table* 8, and the evaluator prompt is provided in *Prompt* E. ❹ **Human Validation.** To verify the reliability of GPT-based hallucination assessment, we additionally conduct human evaluation on 100 randomly selected examples. The results in *Appendix Table* 9 show strong consistency between human annotations and GPT-based judgments. ❺ **Task Metrics.** For VQA and FLUX-Reasoning evaluation, we report ***BERTScore-P/R/F1***, ***Token-P/R/F1***, ***BLEU-1***, ***Distinct-1/2***, and ***Length Ratio***, capturing semantic correspondence, token-level agreement, generation quality, response diversity, and length consistency, respectively. ❻ **Details of Implementation** are provided in *Appendix* B.2.

4.1 MAIN RESULTS

Table 2: Comparison of visual encoders and LLMs across multiple datasets. Best results are **bold**. GPTscore is used as the **main evaluation metric**.

Encoders	Models	SEED-Bench		A-OKVQA		Image Caption		Reasoning	
		Recall(%)↑	GPTscore(%)↑	Recall(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑
CoT-Augmented Multimodal LLMs									
`Qwen3-VL-8B-Thinking`		–	–	–	–	79.66$_{\uparrow 0.56}$	27.88$_{\uparrow 0.44}$	89.04$_{\uparrow 0.46}$	64.91$_{\uparrow 6.50}$
`InternVL-3.5-8B`		–	–	–	–	79.54$_{\uparrow 0.68}$	24.23$_{\uparrow 4.09}$	89.17$_{\uparrow 0.33}$	64.55$_{\uparrow 6.86}$
Visual Encoder–LLM Combinations									
`CLIP`	Qwen2.5-1.5B	50.91$_{\uparrow 18.57}$	24.40$_{\uparrow 12.69}$	21.79$_{\uparrow 42.68}$	10.30$_{\uparrow 30.80}$	84.35$_{\uparrow 0.43}$	21.54$_{\uparrow 29.48}$	87.79$_{\uparrow 1.20}$	29.36$_{\uparrow 22.05}$
`SigLIP`	Qwen2.5-1.5B	46.72$_{\uparrow 22.76}$	23.10$_{\uparrow 13.99}$	21.89$_{\uparrow 42.58}$	5.10$_{\uparrow 36.00}$	83.92$_{\uparrow 0.86}$	13.59$_{\uparrow 37.43}$	86.51$_{\uparrow 2.48}$	23.59$_{\uparrow 27.82}$
`DINOv2`	Qwen2.5-1.5B	41.40$_{\uparrow 28.08}$	17.70$_{\uparrow 19.39}$	16.67$_{\uparrow 47.80}$	14.10$_{\uparrow 27.00}$	83.40$_{\uparrow 1.38}$	14.36$_{\uparrow 36.66}$	88.15$_{\uparrow 0.84}$	36.67$_{\uparrow 14.74}$
`DINOv3`	Qwen2.5-1.5B	55.74$_{\uparrow 13.74}$	15.40$_{\uparrow 21.69}$	47.43$_{\uparrow 17.04}$	25.60$_{\uparrow 15.50}$	83.90$_{\uparrow 0.88}$	23.97$_{\uparrow 27.05}$	88.26$_{\uparrow 0.73}$	45.31$_{\uparrow 6.10}$
`Granulon`	Qwen2.5-1.5B	58.80$_{\uparrow 10.68}$	35.80$_{\uparrow 1.29}$	57.13$_{\uparrow 7.34}$	35.90$_{\uparrow 5.20}$	84.31$_{\uparrow 0.47}$	31.28$_{\uparrow 19.74}$	88.94$_{\uparrow 0.05}$	49.31$_{\uparrow 2.10}$
Ours	Qwen2.5	**69.48**	**37.09**	**64.47**	**41.10**	**84.78**	**51.02**	**88.99**	**51.41**
`CLIP`	Qwen3-8B	92.16$_{\uparrow 4.98}$	11.94$_{\uparrow 33.87}$	49.87$_{\uparrow 0.90}$	16.67$_{\uparrow 37.56}$	79.31$_{\uparrow 0.91}$	24.74$_{\uparrow 3.58}$	88.05$_{\uparrow 1.45}$	39.74$_{\uparrow 31.67}$
`DINOv3`	Qwen3-8B	94.24$_{\uparrow 2.90}$	28.23$_{\uparrow 17.58}$	47.67$_{\uparrow 3.10}$	38.97$_{\uparrow 15.26}$	79.48$_{\uparrow 0.74}$	22.49$_{\uparrow 5.83}$	88.92$_{\uparrow 0.58}$	58.72$_{\uparrow 12.69}$
Ours	Qwen3-8B	**97.14**	**45.81**	**50.77**	**54.23**	**80.22**	**28.32**	**89.50**	**71.41**
`CLIP`	Llama3.2-3B	42.53$_{\uparrow 10.66}$	21.10$_{\uparrow 38.09}$	25.64$_{\uparrow 39.12}$	25.60$_{\uparrow 39.40}$	82.91$_{\uparrow 1.63}$	7.44$_{\uparrow 18.07}$	87.01$_{\uparrow 1.92}$	28.97$_{\uparrow 33.85}$
`SigLIP`	Llama3.2-3B	40.62$_{\uparrow 12.57}$	17.90$_{\uparrow 41.29}$	17.94$_{\uparrow 46.82}$	6.40$_{\uparrow 58.60}$	**84.89**$_{\uparrow 0.35}$	15.36$_{\uparrow 10.15}$	88.46$_{\uparrow 0.47}$	44.87$_{\uparrow 17.95}$
`DINOv2`	Llama3.2-3B	40.26$_{\uparrow 12.93}$	23.10$_{\uparrow 36.09}$	35.90$_{\uparrow 28.86}$	17.90$_{\uparrow 47.10}$	82.59$_{\uparrow 1.95}$	5.54$_{\uparrow 19.97}$	86.55$_{\uparrow 2.38}$	19.49$_{\uparrow 43.33}$
`DINOv3`	Llama3.2-3B	47.23$_{\uparrow 5.96}$	30.30$_{\uparrow 28.89}$	32.05$_{\uparrow 32.71}$	7.70$_{\uparrow 57.30}$	82.98$_{\uparrow 1.56}$	8.12$_{\uparrow 17.39}$	87.26$_{\uparrow 1.67}$	34.87$_{\uparrow 27.95}$
`Granulon`	Llama3.2-3B	48.98$_{\uparrow 4.21}$	53.80$_{\uparrow 5.39}$	54.61$_{\uparrow 10.15}$	51.30$_{\uparrow 13.70}$	83.82$_{\uparrow 0.72}$	18.08$_{\uparrow 7.43}$	87.33$_{\uparrow 1.60}$	56.67$_{\uparrow 6.15}$
Ours	Llama3.2-3B	**53.19**	**59.19**	**64.76**	**65.00**	84.54	**25.51**	**88.93**	**62.82**

Obs. ❶: Consistent Performance Gains across LLMs and VL Models. As shown in Table 2, our method consistently improves GPTscore across different language backbones. With Qwen2.5, it achieves 37.09% on SEED-Bench, 41.10% on A-OKVQA, 51.02% on image captioning, and 51.41% on reasoning, outperforming conventional visual encoders in overall response quality. With Qwen3-8B and Llama3.2-3B, the reasoning GPTscore further reaches 71.41% and 62.82%, respectively, demonstrating strong generalization across LLMs. In addition, we directly evaluate Qwen3-VL-8B-Thinking and InternVL-3.5-8B without additional fine-tuning. Their captioning and reasoning results further validate the effectiveness of structured meta memory. Overall, these results demonstrate consistent improvements in semantic understanding and reasoning beyond surface-level visual matching. **Obs. ❷: Substantial Improvements in Multi-step Reasoning.** As listed in Table 2, our method consistently improves reasoning performance across different LLM backbones. The reasoning GPTscore reaches 51.41% with Qwen2.5, 71.41% with Qwen3-8B, and 62.82% with Llama3.2-3B, consistently surpassing

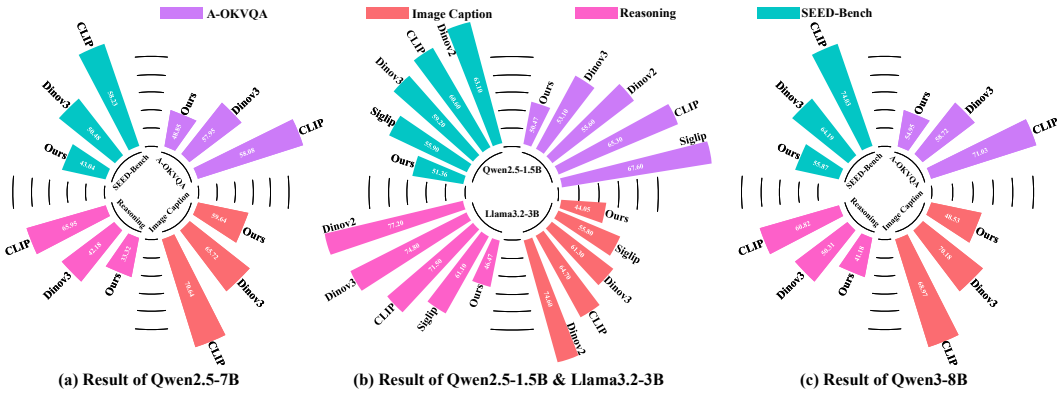
DINOv3 and CLIP. Notably, the gains on GPTscore are substantially larger than those on BERTscore, suggesting improved logical coherence and multi-step reasoning rather than superficial semantic similarity. These results demonstrate that `Meta-CogGran` enables more fine-grained reasoning.

Table 3: OOD Generalization on SurgVLM.

Encoders	Phase Recognition		Instrument Recognition	
	BERTscore(%)↑	Recall(%)↑	BERTscore(%)↑	Recall(%)↑
CLIP	91.64$_{\uparrow 4.89}$	46.15$_{\uparrow 31.96}$	94.44$_{\uparrow 3.49}$	46.92$_{\uparrow 32.12}$
SigLIP	91.98$_{\uparrow 4.55}$	48.72$_{\uparrow 29.39}$	95.71$_{\uparrow 2.22}$	78.20$_{\uparrow 0.84}$
DINOv2	91.67$_{\uparrow 4.86}$	33.33$_{\uparrow 44.78}$	95.93$_{\uparrow 2.00}$	55.21$_{\uparrow 23.83}$
DINOv3	94.71$_{\uparrow 1.82}$	64.10$_{\uparrow 14.01}$	97.41$_{\uparrow 0.52}$	68.89$_{\uparrow 10.15}$
Ours	**96.53**	**78.11**	**97.93**	**79.04**

Obs. ❸: Cross-domain Generalization in Medical Tasks. We further evaluate cross-domain generalization on fine-grained medical tasks, including phase and instrument recognition, as listed in Table 3. Our method achieves the best performance, reaching 96.53% / 78.11% BERTScore/Recall on phase recognition and 97.93% / 79.04% on instrument recognition. Compared with `CLIP`, `SigLIP`, and `DINOv2/v3`, our method consistently improves both semantic accuracy and recognition recall. It surpasses `DINOv3` by 14.01% in phase-recognition Recall and 10.15% in instrument-recognition Recall.

4.2 Robustness and Trustworthiness: Mitigating Hallucinations



(a) Result of Qwen2.5-7B (b) Result of Qwen2.5-1.5B & Llama3.2-3B (c) Result of Qwen3-8B

Figure 3: Hallucination comparison across vision encoders, LLMs and downstream tasks.

As shown in Figure 3 (a), with Qwen2.5-7B, our method reduces the hallucination score to 43.04 on SEED-Bench, compared with 58.23 for CLIP and 50.48 for DINOv3, and to 33.32 on FLUX-Reasoning, compared with 65.95 and 42.18, respectively. In Figure 3 (b), our method further achieves the lowest scores of 51.36 on SEED-Bench and 46.47 on FLUX-Reasoning, outperforming all baselines. Overall, these results demonstrate the consistent effectiveness of our method in mitigating hallucinations across different model scales and benchmarks.

4.3 Ablation Studies

Table 4: Performance comparison across multimodal model configurations. Best results are **bold**.

Encoders	Models	SEED-Bench		A-OKVQA		Image Caption		Reasoning	
		Recall(%)↑	GPTscore(%)↑	Recall(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑
End-to-End Multimodal Models									
`Qwen2.5-VL-3B`		58.79$_{\uparrow 9.79}$	33.54$_{\uparrow 20.20}$	44.50$_{\uparrow 4.05}$	37.89$_{\uparrow 11.01}$	79.04$_{\uparrow 4.67}$	30.55$_{\uparrow 3.28}$	83.88$_{\uparrow 5.01}$	47.72$_{\uparrow 23.70}$
`Qwen2.5-VL-7B`		**68.89**$_{\downarrow 0.31}$	35.64$_{\uparrow 18.10}$	48.33$_{\uparrow 0.22}$	43.77$_{\uparrow 5.13}$	79.43$_{\uparrow 4.28}$	**48.64**$_{\downarrow 14.81}$	84.38$_{\uparrow 4.51}$	54.80$_{\uparrow 16.62}$
Visual Encoder–LLM Combinations									
`CLIP`	Qwen2.5-7B	55.33$_{\uparrow 13.25}$	31.45$_{\uparrow 22.29}$	**48.89**$_{\downarrow 0.34}$	38.21$_{\uparrow 10.69}$	79.85$_{\uparrow 3.86}$	20.64$_{\uparrow 13.19}$	73.92$_{\uparrow 14.97}$	31.67$_{\uparrow 39.75}$
`DINOv3`		60.77$_{\uparrow 7.81}$	39.57$_{\uparrow 14.17}$	47.20$_{\uparrow 1.35}$	42.88$_{\uparrow 6.02}$	79.62$_{\uparrow 4.09}$	23.17$_{\uparrow 10.66}$	86.62$_{\uparrow 2.27}$	63.29$_{\uparrow 8.13}$
Ours		68.58	**53.74**	48.55	**48.90**	**83.71**	33.83	**88.89**	**71.42**

Obs. ❶: Scaling Up with Consistent Multi-Task Gains. As shown in Table 4, our method remains effective when scaling to Qwen2.5-7B. On SEED-Bench, it achieves 68.58% Recall and 53.74% GPTscore, improving over DINOv3 by 7.81 and 14.17 points, respectively. On A-OKVQA, the GPTscore further increases to 48.90%, while the reasoning GPTscore reaches 71.42%, outperforming DINOv3 by 8.13%. Our method also achieves the best BERTScore on image captioning and reasoning, demonstrating consistent gains across multiple tasks. Compared with end-to-end Qwen2.5-VL models, it remains highly competitive while delivering stronger overall reasoning quality. Overall,

the larger gains on GPTscore indicate improved semantic coherence and reasoning capability, confirming effective scalability to larger language backbones. ♣ **Evaluation with Diverse Metrics.** To complement the GPT-based evaluation, we report diverse metrics on FLUX-Reason in Table 7. Our method achieves the best results on BERTScore-P/R/F1, Token-P/R/F1, BLEU-1, and Distinct-1/2. In particular, Token-F1 reaches 49.39%, BLEU-1 reaches 43.34%, and Distinct-2 reaches 94.66%. These improvements across semantic similarity, lexical overlap, and generation diversity show that the gains are robust across evaluation metrics rather than being specific to GPT-based scoring.

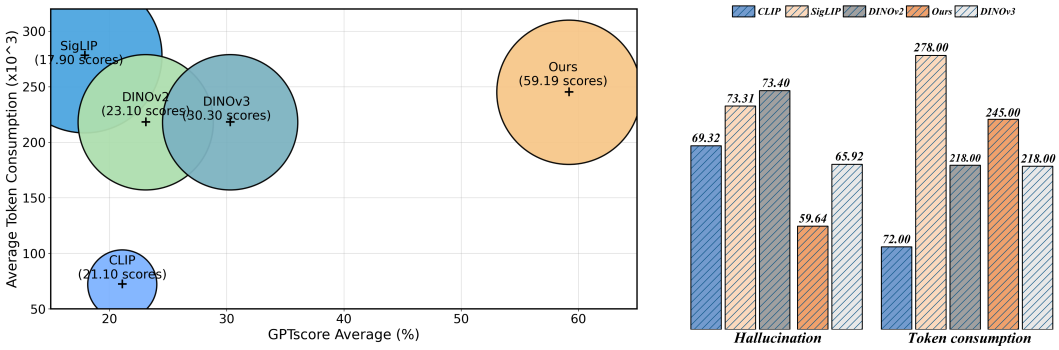


Figure 4: Visualization of Performance Metrics and Input Token Consumption Across Different Setups. Token consumption is measured using Llama 3.2-3B as the language model.

Table 5: Observation memory statistics (Flux).

Config.	Scale	Obs. Memory	Avg. Time/Sample
Qwen2.5 + DINOv3 +Meta-CogGran	1.5B	4492 MiB / 4.39 GiB	4.510 s/Sample
Llama3.2 + DINOv3 +Meta-CogGran	3B	8162 MiB / 7.97 GiB	4.896 s/Sample
Qwen2.5 + DINOv3 +Meta-CogGran	7B	16280 MiB / 15.90 GiB	6.360 s/Sample
Qwen3 + DINOv3 +Meta-CogGran	8B	17008 MiB / 16.61 GiB	7.400 s/Sample

Obs. ❷: Computational Efficiency Analysis in Terms of Performance and Resource Utilization. We further evaluate the computational efficiency of Meta-CogGran in terms of reasoning performance, hallucination, token consumption, and runtime cost, as shown in Figure 4 & Table 5. Compared with single-representation baselines, Meta-CogGran achieves a substantially higher GPTscore of 59.19, outperforming DINOv2 (23.10), DINOv3 (30.30), SigLIP (17.90), and CLIP (21.10). Meanwhile, it reduces the hallucination rate to 59.64, compared with 73.40, 65.92, 73.31, and 69.32 for the respective baselines. Notably, Meta-CogGran maintains token efficiency comparable to existing visual encoders such as DINOv2/DINOv3 and SigLIP, while delivering markedly stronger reasoning performance. We further analyze its practical efficiency across different language-model scales. ♣ **Computational Efficiency Analysis.** As reported in Table 5, the observation-memory overhead scales smoothly from 4.39 GiB for a 1.5B model to 16.61 GiB for an 8B model, while the average inference time increases moderately from 4.510 to 7.400 s/sample. Despite the memory-intensive nature of billion-scale LLM inference (Kwon et al., 2023; Dao et al., 2022), Meta-CogGran maintains a moderate observation-memory footprint with practical inference latency across different backbone scales, demonstrating favorable computational scalability. These results demonstrate that Meta-CogGran provides a favorable trade-off among reasoning accuracy, hallucination mitigation, token consumption, and computational overhead. **Obs. ❸: GPT-based Hallucination Metric Reliability.** We further validate the reliability of the GPT-based hallucination metric through human evaluation. As shown in *Appendix Table 9*, *human and GPT-based evaluations yield highly consistent average hallucination scores* (69.70 vs. 70.00), with an absolute difference of only 0.30. This strong agreement between human judgments and GPT-based scores further demonstrates the reliability of our automated hallucination evaluation, with detailed case-level comparisons provided in *Appendix* B.5.

Table 6: Meta-CogGran for semantic & pixel encoders evaluation.

Encoders	Models	SEED-Bench		A-OKVQA		Image Caption		Reasoning	
		Recall(%)↑	GPTscore(%)↑	Recall(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑
Semantic encoder verification									
CLIP	Qwen2.5	50.91	24.40	21.79	10.30	84.35	21.54	87.79	29.36
+Meta-CogGran		**49.82**	**20.13**	**20.57**	**9.18**	**82.11**	**18.07**	**84.31**	**28.19**
CLIP	Llama3.2	42.53	21.10	25.64	25.60	82.91	7.44	87.01	28.97
+Meta-CogGran		**41.89**	**19.19**	**23.12**	**22.66**	**80.08**	**8.12**	**86.92**	**28.54**
Pixel encoder verification									
DINOv3	Qwen2.5	55.74	15.40	47.43	25.60	83.90	23.97	88.26	45.31
+Meta-CogGran		**69.48**	**37.09**	**64.47**	**41.10**	**84.78**	**51.02**	**88.99**	**51.41**
DINOv3	Llama3.2	47.23	30.30	32.05	7.70	82.98	8.12	87.26	34.87
+Meta-CogGran		**53.19**	**59.19**	**64.76**	**65.00**	**84.54**	**25.51**	**88.93**	**62.82**

Obs. ❹: Generalization ability on Semantic and Pixel Encoders. As listed in Table 6, Meta-CogGran exhibits notably different behaviors across semantic and pixel-level visual encoders. When coupled with CLIP, its contribution remains relatively limited, with most met-

rics showing marginal decreases. This trend suggests that CLIP's highly abstract and globally aligned representations may already suppress much of the fine-grained spatial variation that `Meta-CogGran` is designed to exploit. Substantially stronger gains are observed with DINOv3. For Qwen2.5, the SEED-Bench GPTscore increases from 15.40 to 37.09, while the Image Caption GPTscore improves from 23.97 to 51.02. The effect is more pronounced with Llama3.2, where the A-OKVQA GPTscore rises from 7.70 to 65.00 and the Reasoning GPTscore from 34.87 to 62.82. This suggests that `Meta-CogGran` ***benefits more from fine-grained pixel representations that preserve local and structural cues, while highly semantic features provide less complementary information for further structural refinement.***

Obs. ❺: Ours vs. CoT-Based Reasoning (Best-Case Analysis). As shown in Figure 5, the CoT-based Qwen3-VL-8B-Thinking explicitly performs step-by-step reasoning, but several early visual errors, such as misidentifying the tie and shirt colors, are propagated into subsequent scene-level inference. In contrast, our method first organizes visual understanding from Global, Entity, Layout, and Relation perspectives, leading to more accurate fine-grained recognition and a more complete interpretation of the image. **Obs. ❻: Qualitative Case Study.** To examine our method, we present a qualitative case study in *Appendix* G (Figure 13). Compared with `CLIP`, our model provides a more accurate and fine-grained description, correctly capturing key details such as the gray backdrop, embellished bodice, waist belt, flowing skirt, and bridal accessories. It also better reflects the overall composition and visual emphasis of the image. Despite a few attribute-level errors, this example demonstrates improved semantic alignment and descriptive completeness over the baseline.

Ours vs. CoT-Based VL Models (Best-Case Analysis)



Question: *Please analyze and describe the image in detail.*
`Qwen3-VL-8B-Thinking (CoT)`: *First, I need to look at each part of the image. There are three men in suits walking on a runway. For the left man, I observe a dark navy suit with a black tie. For the middle man, the suit is also dark navy, with a patterned gold-and-dark tie. For the right man, the suit is beige with a dark tie. All three appear to wear light-blue dress shirts. Next, I infer the scene context. The suits are tailored, and the blurred figures in the background look like photographers or an audience. Therefore, this is a fashion-show runway. ... Putting these observations together, ...*

OURS (Based on Qwen3-8B-Base): ***[Global]*** *Three well-dressed men stride confidently through a polished, softly blurred fashion-show environment, conveying a refined and professional atmosphere.* ***[Entity]*** *The left model wears a navy suit with a* ***crisp white shirt and burgundy tie****; the center model wears a charcoal-gray suit, light-blue shirt, and patterned tie; the right model wears a light-brown suit with a blue-toned shirt and striped tie.* ***[Layout]*** *They are arranged left-to-right as three prominent full-body figures, with the center model visually anchoring the composition.* ***[Relation]*** *Despite differences in suit color and accessories, their coordinated tailoring, dark dress shoes, and synchronized forward movement create a cohesive menswear presentation, while the blurred background keeps attention on the garments.*

Figure 5: Qualitative comparison between Qwen3-VL-8B-Thinking (CoT) and our method. Green denotes correct visual understanding, red indicates erroneous or unsupported observations, and gray marks intermediate reasoning statements. Our method integrates Global, Entity, Layout, and Relation information for more complete perception.

5 Conclusion & Limitation

We presented `Meta-CogGran`, a novel framework that formulates multimodal reasoning as online latent cognitive state evolution rather than autoregressive token generation. By explicitly modeling cognitive states, action-conditioned transitions, and semantic equilibrium, our approach enables MLLMs to iteratively inspect, verify, and refine their beliefs through interactions with visual observations and semantic memory. Extensive experiments demonstrate that `Meta-CogGran` consistently improves reasoning performance while significantly reducing hallucinations. These findings suggest that explicit cognitive state dynamics provide a more scalable, interpretable, and reliable paradigm for multimodal reasoning and offer a promising direction for future multimodal intelligence systems. **Limitation:** While `Meta-CogGran` demonstrates strong empirical performance, the current implementation adopts a relatively generic design for cognitive state representations and transition operators. Exploring more expressive cognitive structures and task-adaptive transition mechanisms may further improve reasoning efficiency and generalization.

LLM USAGE DISCLOSURE

Large Language Models (LLMs) were used as general-purpose assistance tools during the preparation of this manuscript. Specifically, GPT-5.6 was used to assist with language editing and polishing, including grammar correction, sentence restructuring, and improvements to clarity and readability. It was also used to provide assistance with general writing tasks and coding-related aspects of the research. The use of GPT-5.6 did not introduce new scientific ideas, experimental results, interpretations, or claims. All LLM-assisted content was based on the authors' original ideas, analyses, and experimental results, and was carefully reviewed, verified, and revised by the authors to ensure that it accurately reflects the intended meaning. GPT-5.6 was not used to autonomously conduct the research or complete any part of the manuscript. The authors take full responsibility for all content, analyses, results, and claims presented in this work.

REPRODUCIBILITY STATEMENT

We provide implementation and evaluation details for `Meta-CogGran` in the appendices. Appendix B describes data preprocessing, deterministic data partitioning, training and evaluation splits, model backbones, optimization settings, hardware configuration, and evaluation metrics. The complete GPT-4o judging prompts, inference settings, score aggregation procedure, and human-validation protocol are also provided in Appendices B.4 and E. Source code is included as `code.zip` in the anonymous submission. FLUX-Reason is the sole source of training supervision; SEED-Bench, A-OKVQA, CC12M_and_Imagenet21K_Recap, and SurgVLM are evaluated without benchmark-specific fine-tuning.

ETHICS STATEMENT

This study uses publicly available benchmark datasets and released pretrained models, and collects no new personal data. The human validation is limited to judging 100 model outputs from FLUX-Reason and involves no intervention or collection of sensitive participant information. Because benchmark data may contain incidental depictions of people or sensitive scenes, it should be handled according to the original dataset licenses and terms. GPT-4o is used only as an evaluation model; its scores may reflect limitations of model-based judging, and we report the prompts and evaluation protocol transparently.

REFERENCES

Andrew B Barron, Marta Halina, and Colin Klein. Transitions in cognitive evolution. Proceedings of the Royal Society B: Biological Sciences, 290(2002):20230671, 2023.

Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 3836–3845, 2025.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3558–3568, 2021.

Xiaohui Chen, Satya Narayan Shukla, Mahmoud Azab, Aashu Singh, Qifan Wang, David Yang, ShengYun Peng, Hanchao Yu, Shen Yan, Xuewen Zhang, et al. Compcap: Improving multimodal large language models with composite captions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23582–23592, 2025.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems, 35:16344–16359, 2022.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 9062–9072, 2025.

Rongyao Fang, Aldrich Yu, Chengqi Duan, Linjiang Huang, Shuai Bai, Yuxuan Cai, Kun Wang, Si Liu, Xihui Liu, and Hongsheng Li. Flux-reason-6m & prism-bench: A million-scale text-to-image reasoning dataset and comprehensive benchmark. arXiv preprint arXiv:2509.09680, 2025.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Hao Tian, Shilin Yan, Weihao Yu, Xingyu Zeng, Jifeng Dai, et al. Got: Unleashing reasoning capability of mllm for visual generation and editing. Advances in Neural Information Processing Systems, 38:67680–67708, 2026.

Yuchen Feng, Zhenyu Zhang, Naibin Gu, Yilong Chen, Peng Fu, Zheng Lin, Shuohuan Wang, Yu Sun, Hua Wu, Weiping Wang, et al. Blink: Dynamic visual token resolution for enhanced multimodal understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3511–3521, 2026.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. Rstar-math: Small llms can master math reasoning with self-evolved deep thinking. arXiv preprint arXiv:2501.04519, 2025.

Feng He, Zhenting Wang, Qifan Wang, Qiang Guan, Dongfang Liu, Ruixiang Tang, and Qiankun Li. Hive: Understanding post-hallucination reasoning in vision language models. arXiv preprint arXiv:2607.07507, 2026.

P Hitzler, MK Sarker, TR Besold, AD Garcez, S Bader, H Bowman, P Domingos, P Hitzler, KU Kühnberger, LC Lamb, et al. Neural-symbolic learning and reasoning: A survey and interpretation. Frontiers in artificial intelligence and applications, 342:1–51, 2022.

Paul I Jaffe, Russell A Poldrack, Robert J Schafer, and Patrick G Bissett. Modelling human behaviour in cognitive tasks with latent dynamical systems. Nature Human Behaviour, 7(6):986–1000, 2023.

Jingjing Jiang, Chao Ma, Xurui Song, Hanwang Zhang, and Jun Luo. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3034–3046, 2025a.

Song Jiang, Lijuan Jiao, Yuguang Li, Pei You, Hui Li, Ran Yang, Feng Zhu, Sehwan Ki, and Hyong-Euk Lee. Exploring mllm in fine-grained visual quality comparison with quality token. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3475–3485, 2025b.

Zhiyuan Jiang, Weihao Hong, Xinlei Guan, Tejaswi Dhandu, Miles Q Li, Meng Xu, Kuan Huang, Umamaheswara Rao Tida, Bingyu Shen, Daehan Kwak, et al. Llm-as-judge framework for evaluating tone-induced hallucination in vision-language models. arXiv preprint arXiv:2604.18803, 2026.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Harry Chao. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. Advances in Neural Information Processing Systems, 37:28798–28827, 2024.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. ACM Computing Surveys, 57(8):1–36, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the 29th symposium on operating systems principles, pp. 611–626, 2023.

Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Wenze Hu, Juan Tebar, Zhe Gan, Peter Grasch, et al. Revisit large-scale image-caption data in pre-training multimodal foundation models. In International Conference on Learning Representations, volume 2025, pp. 94493–94512, 2025.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 2757–2791, 2025a.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579, 2024.

Haoxuan Li, Yi Bin, Yunshan Ma, Guoqing Wang, Yang Yang, See-Kiong Ng, and Tat-Seng Chua. Semcore: A semantic-enhanced generative cross-modal retrieval framework with mllms. arXiv preprint arXiv:2504.13172, 2025b.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. Implicit reasoning in large language models: A comprehensive survey. arXiv preprint arXiv:2509.02350, 2025c.

Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. Mind with eyes: from language reasoning to multimodal reasoning. arXiv preprint arXiv:2503.18071, 2025.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9225–9236, 2026.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In International Conference on Learning Representations, volume 2024, pp. 57689–57733, 2024.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Yongbin Li, Minzheng Wang, Pengpeng Zeng, et al. Mmevol: Empowering multimodal large language models with evol-instruct. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 19655–19682, 2025.

Junyuan Mao, Qiankun Li, Linghao Meng, Zhicheng He, Xinliang Zhou, Kun Wang, Yang Liu, and Yueming Jin. Granulon: Awakening pixel-level visual encoders with adaptive multi-granularity semantics for mllm. arXiv preprint arXiv:2603.08800, 2026.

Ti-Fen Pan, Jing-Jing Li, Bill Thompson, and Anne GE Collins. Latent variable sequence identification for cognitive models with neural network estimators. Behavior Research Methods, 57(10): 272, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PmLR, 2021.

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 41171–41181, 2026.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In European conference on computer vision, pp. 146–162. Springer, 2022.

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. Advances in Neural Information Processing Systems, 37:8612–8642, 2024.

Michael J Spivey. Cognitive science progresses toward interactive frameworks. Topics in Cognitive Science, 15(2):219–254, 2023.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In Findings of the association for computational linguistics: ACL 2024, pp. 13088–13110, 2024.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. arXiv preprint arXiv:2410.12784, 2024.

Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. Visual position prompt for mllm based visual grounding. IEEE Transactions on Multimedia, 2026.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9568–9578, 2024.

Danae Sánchez Villegas, Ingo Ziegler, and Desmond Elliott. Imagechain: Advancing sequential image-to-text reasoning in multimodal large language models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5131–5141, 2026.

Chujie Wang, Jianyu Lu, Zhiyuan Luo, Xi Chen, and Chu He. Ovod-agent: A markov-bandit framework for proactive visual reasoning and self-evolving detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 41416–41425, 2026a.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. Advances in neural information processing systems, 38:4350–4376, 2026b.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. arXiv preprint arXiv:2411.10442, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 33577–33587, 2026.

Kun Xiang, Zhili Liu, Terry Jingchen Zhang, Yinya Huang, Yunshuang Nie, Kaixin Cai, Yiyang Yin, Runhui Huang, Hanhui Li, Yihan Zeng, et al. Atomthink: Multimodal slow thinking with atomic step reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

Yicheng Xiao, Lin Song, Yukang Chen, Yingmin Luo, Yuxin Chen, Yukang Gan, Wei Huang, Xiu Li, Xiaojuan Qi, and Ying Shan. Mindomni: Unleashing reasoning generation in vision language models with rgpo. Advances in Neural Information Processing Systems, 38:88786–88810, 2026.

Zhifei Xie, Ziyang Ma, Zihang Liu, Kaiyu Pang, Hongyu Li, Jialin Zhang, Yue Liao, Deheng Ye, Chunyan Miao, and Shuicheng Yan. Mini-omni-reasoner: Token-level thinking-in-speaking in large speech models. arXiv preprint arXiv:2508.15827, 2025.

Yifei Xing, Yijie Yu, Ruiping Wang, Wenjun Huang, Qingfang Zheng, Dongmei Jiang, and Xiangyuan Lan. Dytos: Budget-aware dynamic token scheduling for efficient multi-modal large language models. Neurocomputing, pp. 134181, 2026.

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. Advances in Neural Information Processing Systems, 38:119152–119184, 2026.

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. Captionqa: Is your caption as useful as the image itself? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23741–23750, 2026a.

Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in mllm reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 11694–11702, 2026b.

Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, et al. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 273–283, 2025.

Linli Yao, Long Xing, Yang Shi, Sida Li, Yuanxin Liu, Yuhao Dong, Yi-Fan Zhang, Lei Li, Qingxiu Dong, Xiaoyi Dong, et al. Towards efficient multimodal large language models: A survey on token compression. 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36:11809–11822, 2023.

Jiazuo Yu, Haomiao Xiong, Lu Zhang, Haiwen Diao, Yunzhi Zhuge, Lanqing Hong, Dong Wang, Huchuan Lu, You He, and Long Chen. LLMs can evolve continually on modality for $\mathbb{X}$-Modal reasoning. In Advances in Neural Information Processing Systems, volume 37, pp. 49834–49858, 2024.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Auto-controlled image perception in mllms via visual perception tokens. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21822–21831, 2025.

Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. arXiv preprint arXiv:2506.02555, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 11975–11986, 2023.

Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, Minghui Qiu, et al. Seeing is believing? mitigating ocr hallucinations in multimodal large language models. Advances in Neural Information Processing Systems, 38:74230–74248, 2026a.

Ce Zhang, Jinxi He, Junyi He, Katia Sycara, and Yaqi Xie. Evolving contextual safety in multi-modal large language models via inference-time self-reflective memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 41182–41192, 2026b.

Daoze Zhang, Chenghan Fu, Zhanheng Nie, Jianyu Liu, Wanxian Guan, Yuan Gao, Jun Song, Pengjie Wang, Jian Xu, and Bo Zheng. Moon: Generative mllm-based multimodal representation learning for e-commerce product understanding. In Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining, pp. 924–933, 2026c.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. Advances in Neural Information Processing Systems, 38:25438–25468, 2026d.

Siyuan Zhang, Jialian Li, Yichi Zhang, Xiao Yang, Yinpeng Dong, and Hang Su. Reasoning as state transition: A representational analysis of reasoning evolution in large language models. arXiv preprint arXiv:2602.00770, 2026e.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Min Zhang, Wen Zhang, and Huajun Chen. Abstractive visual understanding of multi-modal structured knowledge: A new perspective for mllm evaluation. In Proceedings of the 33rd ACM International Conference on Multimedia, pp. 12323–12332, 2025a.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 19781–19791, 2025b.

Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. Unsupervised visual chain-of-thought reasoning via preference optimization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2303–2312, 2025.

Shukuo Zhao, Hongzhi You, Ru-Yuan Zhang, Bailu Si, Zonglei Zhen, Xiaohong Wan, and Da-Hui Wang. An interpretable neuro-symbolic model for raven's progressive matrices reasoning. Cognitive Computation, 15(5):1703–1724, 2023.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. Advances in neural information processing systems, 38: 20560–20586, 2026.

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. Advances in neural information processing systems, 36:46595–46623, 2023.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. Focus: Internal mllm representations for efficient fine-grained visual question answering. Advances in Neural Information Processing Systems, 38:96498–96535, 2026.

Fangrui Zhu, Hanhui Wang, Yiming Xie, Jing Gu, Tianye Ding, Jianwei Yang, and Huaizu Jiang. Struct2d: A perception-guided framework for spatial reasoning in mllms. Advances in Neural Information Processing Systems, 38:135805–135837, 2026a.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, and Shiqi Wang. Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. Advances in Neural Information Processing Systems, 38:171848–171876, 2026b.

APPENDIX CONTENTS

Appendix A. Hierarchy-Aware Visual Understanding Analysis 17
Appendix B. More Implementation Details 17
Appendix C. Hierarchy-Aware Analysis of Fine-Grained Visual Understanding 22
Appendix D. Compute Reporting 23
Appendix E. Evaluation Prompts 23
Appendix F. More Layer-wise Alignment Analysis 24
Appendix G. More Qualitative Case Study 25

A Hierarchy-Aware Visual Understanding Analysis

Table 7: Evaluation on FLUX-Reason with diverse metrics. Best results are **bold**.

Encoders	Models	Open-ended generation (FLUX-Reason)									
		BERTScore-P(%)↑	BERTScore-R(%)↑	BERTScore-F1(%)↑	Token-P(%)↑	Token-R(%)↑	Token-F1(%)↑	BLEU-1(%)↑	Distinct-1(%)↑	Distinct-2(%)↑	Length Ratio
End-to-End Multimodal Models											
`Qwen2.5-VL-3B`		82.87$_{\uparrow 6.45}$	84.93$_{\uparrow 3.53}$	83.88$_{\uparrow 5.01}$	36.71$_{\uparrow 12.95}$	40.57$_{\uparrow 9.15}$	38.39$_{\uparrow 11.00}$	30.33$_{\uparrow 13.01}$	57.04$_{\uparrow 7.14}$	89.35$_{\uparrow 5.31}$	1.1185
`Qwen2.5-VL-7B`		83.55$_{\uparrow 5.77}$	85.23$_{\uparrow 3.23}$	84.38$_{\uparrow 4.51}$	36.25$_{\uparrow 13.41}$	41.35$_{\uparrow 8.37}$	38.47$_{\uparrow 10.92}$	30.41$_{\uparrow 12.93}$	58.41$_{\uparrow 5.77}$	90.78$_{\uparrow 3.88}$	1.1575
Visual Encoder–LLM Combinations											
`CLIP`	Qwen2.5-7B	70.53$_{\uparrow 18.79}$	77.69$_{\uparrow 10.77}$	73.92$_{\uparrow 14.97}$	17.32$_{\uparrow 32.34}$	18.24$_{\uparrow 31.48}$	17.52$_{\uparrow 31.87}$	12.94$_{\uparrow 30.40}$	15.42$_{\uparrow 48.76}$	23.97$_{\uparrow 70.69}$	1.0540
`DINOv3`	Qwen2.5-7B	87.41$_{\uparrow 1.91}$	85.85$_{\uparrow 2.61}$	86.62$_{\uparrow 2.27}$	46.40$_{\uparrow 3.26}$	37.32$_{\uparrow 12.40}$	39.87$_{\uparrow 9.52}$	32.13$_{\uparrow 11.21}$	59.90$_{\uparrow 4.28}$	90.28$_{\uparrow 4.38}$	0.8338
Ours		**89.32**	**88.46**	**88.89**	**49.66**	**49.72**	**49.39**	**43.34**	**64.18**	**94.66**	**1.0184**

Obs. ❶: Layer-wise Alignment Analysis. As shown in Figure 6 and Figure 7, we measure the layer-wise cosine similarity between visual-conditioned hidden representations and their paired textual references. Ours+Llama generally shows higher similarity than CLIP+Llama across samples and layers. The CLIP baseline shows weak correspondence in early and intermediate layers, with noticeable alignment emerging mainly in deeper layers, suggesting that its visual features capture high-level semantics but are less compatible with the LLM's intermediate representation space. In comparison, Ours+Llama establishes stronger alignment at earlier stages and preserves it throughout the network. The more extensive and stable high-similarity regions further indicate that `Meta-CogGran` produces visual representations that better match the LLM's internal semantic structure.

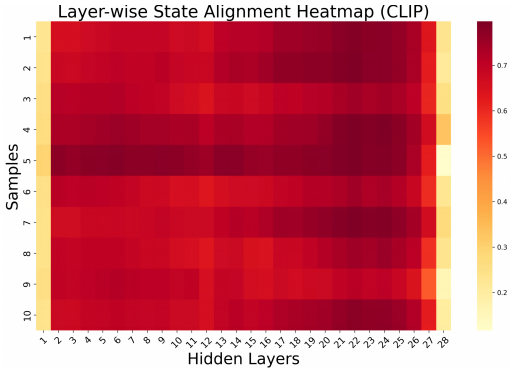


Figure 6: Cosine similarity between the reference state and each corresponding layer state for CLIP + Llama.

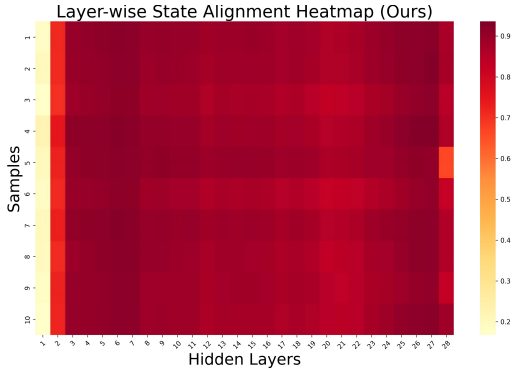


Figure 7: Cosine similarity between the reference state and each corresponding layer state for ours + Llama.

B More Implementation Details

B.1 Dataset Processing

Data Preparation and Experimental Protocol. To facilitate training across datasets with heterogeneous storage formats and annotation conventions, we develop a common data preparation procedure that converts all samples into a standardized multimodal dialogue representation. The same preprocessing strategy is applied to both training and evaluation data unless otherwise specified.

Unified Data Representation. The original datasets are loaded either from local `parquet` files or through the HuggingFace dataset interface. When multiple predefined splits are available, they are first combined before further processing. Each instance is then reorganized into a common structure consisting of a unique sample identifier, a visual input, and its associated textual supervision. Because the source datasets encode images in different ways, including raw byte streams and array-based objects, all visual inputs are converted to regular image files during preprocessing. The processed annotations store only the corresponding relative image path, which makes the resulting data independent of the original storage backend and directly usable by commonly adopted multimodal training pipelines.

Multimodal Dialogue Conversion. After visual normalization, every instance is transformed into a LLaVA-compatible conversational sample. Specifically, a processed instance is represented by (`id`, `image`, `conversations`), where `conversations` contains one user turn followed by one assistant turn. The user instruction is designed to elicit explicit visual reasoning rather than a short

prediction, while the assistant response is generated from the ground-truth annotation provided by the source dataset. The special multimodal marker `<image>` is included in the instruction so that the language context is explicitly associated with the corresponding visual input. Annotation fields often differ substantially across datasets. We therefore map dataset-specific labels or responses into a single textual target and remove formatting inconsistencies before constructing the dialogue. As a result, samples originating from different benchmarks share an identical supervision interface and can be jointly used during multimodal instruction tuning. The final converted instances are stored in JSON format. Each record contains an image path together with a two-turn dialogue, allowing the processed datasets to be consumed directly by multimodal large language model training frameworks without additional dataset-specific adapters.

Reproducible Data Partitioning For datasets requiring a new train-test partition, we use a deterministic slicing strategy instead of random sampling. Let $\mathcal{D} = \{x_0, x_1, \ldots, x_{N-1}\}$ denote the ordered collection of processed examples. Given two boundary indices s and e, samples whose indices satisfy $s \leq i \leq e$ are assigned to the held-out set, whereas all other instances are retained for training. Since the partition is determined entirely by sample indices, identical data splits can be reproduced across different runs without fixing an additional random seed.

Training Data and Evaluation Settings. We consider both general-domain multimodal reasoning and domain-specific visual reasoning in our experiments. For the general reasoning setting, the model is trained using 200K examples selected from FLUX-Reason (Fang et al., 2025). FLUX-Reason provides diverse forms of multimodal supervision, covering compositional reasoning, multi-stage inference, and fine-grained visual grounding. In addition to the training subset, we reserve 1K non-overlapping examples from the same corpus for validation. The validation instances undergo exactly the same preprocessing procedure as the training data, providing an in-domain measurement under a consistent data distribution. We further investigate whether the learned reasoning capability transfers to unseen datasets. To isolate such generalization, FLUX-Reason is the only source of supervision used during training, and no examples from downstream evaluation benchmarks are introduced during optimization. Zero-shot evaluation is performed on SEED-Bench (Li et al., 2023), A-OKVQA (Schwenk et al., 2022), and CC12M_and_Imagenet21K_Recap (Changpinyo et al., 2021). For each benchmark, we uniformly sample 1K instances to construct the evaluation subset. These benchmarks provide complementary perspectives on multimodal capability, including knowledge-intensive visual question answering, compositional reasoning, semantic image understanding, and high-level image description. In addition to the general-domain experiments, we construct a separate medical-domain setting using SurgVLM (Zeng et al., 2025). A total of 1K examples is reserved exclusively for testing. This partition allows us to study visual reasoning in the surgical domain. All reported experiments use the data partitions described above. Evaluation datasets are never used for additional supervised adaptation, and no benchmark-specific fine-tuning is conducted prior to testing. Consequently, results on SEED-Bench, A-OKVQA, SurgVLM, and CC12M_and_Imagenet21K_Recap reflect zero-shot transfer from FLUX-Reason.

B.2 Implementation Details

We conduct all multimodal experiments using different language model backbones, *Qwen-2.5-Instruct*, *Llama-3.2-3B*, *Qwen2.5-7B-Instruct*, *Qwen3-8B-Base*, together with several alternative visual encoders to assess the robustness of our framework across model configurations. In addition, we include several publicly available vision-language models as multimodal reasoning baselines, namely *Qwen2.5-VL-3B-Instruct*, *Qwen2.5-VL-7B-Instruct*, and *InternVL3.5-8B*. Unlike the models trained within our framework, these vision-language models are evaluated directly using their released pretrained or instruction-tuned checkpoints. No additional fine-tuning, task-specific adaptation, or parameter update is performed on these baselines. Their performance therefore reflects their native multimodal reasoning capability under a zero-shot evaluation setting. This comparison allows us to contrast our trained multimodal configurations with strong off-the-shelf vision-language models while avoiding potential performance gains introduced by benchmark-specific supervision. All models are optimized with DeepSpeed under the ZeRO-2 strategy. Training is carried out for two epochs on eight NVIDIA A800-SXM4 GPUs, each equipped with 80 GB of memory. We set the local batch size to 8 samples per device and accumulate gradients over two consecutive steps, yielding a total effective batch size of 128 across all GPUs. Optimization starts from a learning rate of 2×10^{-5}. The learning rate is adjusted using cosine decay, preceded by a warmup phase covering the first 5% of the training schedule. To improve computational efficiency, we enable both bf16 mixed-precision arithmetic and tf32 operations. Gradient checkpointing is additionally employed to reduce activation memory

and accommodate multimodal model training within the available GPU capacity. The input pipeline uses parallel data-loading workers together with pinned host memory to minimize I/O overhead. During optimization, checkpoints are written at predefined intervals, while only the three most recent checkpoints are preserved to limit storage consumption. We do not run validation at intermediate training steps; model evaluation is instead performed after completion of the training procedure.

B.3 Evaluation Metrics.

To obtain a comprehensive assessment of visual encoder quality on both the FLUX-Reason test set and zero-shot downstream benchmarks, we adopt a two-level evaluation framework that combines conventional text-based similarity metrics with multimodal model-based judging. We first report BERTScore as a reference metric for measuring semantic overlap between generated responses and the corresponding textual annotations. However, across different visual encoders, we observe that BERTScore values remain relatively concentrated and often fail to reflect substantial differences that are apparent upon qualitative inspection. This behavior indicates that representation-based text similarity alone is insufficient for evaluating multimodal reasoning, since a generated response may be linguistically similar to the reference while still containing visually unsupported statements, missing important image content, or reaching an incorrect conclusion. We therefore complement BERTScore with a multimodal evaluation protocol based on GPT-4o (Zheng et al., 2023; Li et al., 2024; 2025a; Tan et al., 2024). Through manual examination of model predictions, we find that GPT-4o judgments are substantially more consistent with human assessment than BERTScore, particularly in cases where the quality difference is determined by visual evidence rather than textual overlap. Importantly, GPT-4o receives access to the source image during evaluation, allowing it to directly verify whether the generated response is supported by the observed scene.

This image-grounded evaluation is particularly important for FLUX-Reason. We deliberately avoid treating the provided textual annotation as an exhaustive target description, because the reference may omit visually valid objects, attributes, relations, or reasoning details that are nevertheless present in the image. As a result, a model response can be factually correct and semantically richer than the annotation while receiving an artificially low score under strict reference-based comparison. We observe this behavior in outputs produced by `Meta-CogGran`, which frequently include additional image-grounded details beyond those explicitly recorded in the reference. For this reason, GPT-4o evaluates FLUX-Reason predictions primarily with respect to the original visual input, examining whether the identified evidence, intermediate reasoning, and final interpretation remain consistent with the image rather than requiring exact agreement with a potentially incomplete textual annotation. The judging criteria are adapted according to the downstream task. For visual question answering benchmarks, GPT-4o primarily assesses whether the predicted answer is factually correct and supported by the image, thereby avoiding the brittleness of exact or lexical matching when semantically equivalent answers are expressed differently. For image-description and captioning tasks, the evaluator considers several complementary factors, including coverage of salient visual content, semantic correctness, relevance to the scene, and the presence of unsupported or hallucinated details. Consequently, the model-based evaluator captures aspects of multimodal output quality that are difficult to characterize using text-only similarity metrics.

Overall, the two-level protocol serves distinct purposes: BERTScore provides a reproducible reference-based similarity measure, whereas GPT-4o offers a visually grounded assessment of factuality, completeness, and reasoning faithfulness. Their combination enables a more informative comparison of different visual encoders, especially in settings where the reference annotation represents only one valid textual description of the underlying image.

B.4 GPT-Based Hallucination Evaluation

In addition to standard quantitative measures such as Accuracy, Recall, and BERTScore, we introduce GPT-4o as a multimodal judge to provide a more detailed evaluation of hallucination behavior. Recent multimodal research has increasingly adopted model-based evaluators to capture aspects such as factual consistency, visual faithfulness, and unsupported content that are difficult to characterize using reference-based metrics alone (Sun et al., 2024; Liu et al., 2024; Jiang et al., 2026; Zhang et al., 2026a; He et al., 2026). In our setting, GPT-4o is provided with both the source image and the generated response and is instructed to assess whether the textual content can be justified by the available visual evidence. Following the evaluation protocol of Mao et al. (Mao et al., 2026), hallucination severity is measured on a continuous 0–80 scale, where smaller values correspond to stronger visual grounding. A score of 0 denotes a response with no detectable visually unsupported

information, while a score of 80 represents the most severe level of hallucination. As summarized in the corresponding table, the full range is organized into eight qualitative severity levels, as listed in Table 8. These levels are used as semantic reference points for holistic judgment rather than as rigid rules determined by counting individual erroneous claims. This design allows the evaluator to consider not only the presence of unsupported content, but also its semantic importance and its overall impact on the correctness of the generated response. Each prediction is scored independently, after which the final hallucination metric is computed as the mean of all valid sample-level scores. Responses that do not satisfy the required evaluation format are excluded from aggregation. To facilitate reproducibility, we report the complete judging prompt together with the main inference settings, including the GPT-4o configuration, decoding parameters, input image resolution, score aggregation procedure, and the total number of valid evaluated samples.

Table 8: Hallucination severity levels corresponding to different Hallucination Score ranges.

Hallucination Score	Severity
0–10	Negligible hallucination. The response is highly grounded in the image, with little or no visually unsupported content.
11–20	Very low hallucination. The response remains strongly grounded, although a small degree of unsupported content may be present.
21–30	Low hallucination. The response is generally consistent with the image, but contains some noticeable unsupported information.
31–40	Moderate hallucination. Visually unsupported content is clearly present, while a substantial part of the response remains grounded.
41–50	Moderately high hallucination. Unsupported content becomes prominent and meaningfully affects the factual reliability of the response.
51–60	High hallucination. A large portion of the generated description lacks sufficient visual support.
61–70	Very high hallucination. The response is predominantly unsupported or inconsistent with the image, with limited visually grounded content.
71–80	Extreme hallucination. The response is almost entirely unsupported by, or substantially contradictory to, the image.

B.5 Human Validation of the GPT-Based Hallucination Metric

To further validate the reliability of the GPT-based hallucination metric, we conducted an additional human evaluation on 100 randomly sampled model outputs from the FLUX-Reason benchmark, together with their corresponding ground-truth descriptions. Human evaluators assessed each output using the same scoring scale and evaluation criteria as the GPT-based evaluator, enabling a direct comparison between human and automated judgments under a consistent evaluation protocol. As listed in Table 9, the human and GPT-based evaluations exhibit strong agreement both at the aggregate level and across representative individual cases. Over the 100 sampled cases, the average scores assigned by the human and GPT-based evaluators are 69.70 and 70.00, respectively, with an absolute difference of only 0.30. More importantly, the representative examples demonstrate that the two evaluators respond similarly to different types and degrees of hallucination. For example, in Case 001, where the model incorrectly identifies a Halloween-themed Publix candy display as a "Candyland" store in the Bronx and introduces several unsupported details, both evaluators assign a score of 75. Similarly, Cases 003, 004, and 100 receive identical scores from the human and GPT-based evaluators despite involving different semantic errors, such as incorrect character identities, object categories, brands, and product attributes. Even in Case 002, where the model misidentifies a Painted Bunting as a Northern Cardinal while preserving several visual attributes, the two scores remain very close (60 from the human evaluator versus 65 from the GPT-based evaluator).

These results indicate that the GPT-based evaluator not only produces nearly identical average scores to human evaluation, but also captures the relative severity of hallucinations in a manner that is highly consistent with human judgment at the instance level. This agreement across both overall statistics and representative error patterns provides additional evidence for the reliability of the GPT-based hallucination metric and strengthens the validity of the hallucination-related analyses and conclusions reported in our experiments.

Table 9: Comparison of human and GPT-based hallucination scores on 100 sampled FLUX-Reason cases.

Case	Model output	Ground Truth	Human Judge	GPT-Based
001	Vibrant and eye-catching, the "Candyland" candy store in the Bronx, New York, presents a colorful and inviting display of confectionery delights. The central feature is a large, red candy counter adorned with a bold, black-and-white sign that reads "Candyland," drawing immediate attention to the store's playful theme. ...	Vibrant and festive, a Halloween-themed candy display stand dominates the center of a brightly lit Publix store aisle in Clearwater, Florida. The stand, labeled "HALLOWEEN COOL'E PTS," features multiple orange shelves adorned with playful black ghost and pumpkin silhouettes. ...	75	75
002	Vibrant and striking, a male Northern Cardinal perches gracefully on a slender, rust-colored branch. Its plumage is a vivid contrast of deep red, bright green, and rich blue. The cardinal's head is a bold shade of crimson, complemented by a black mask and an orange-red beak. ...	Perched gracefully on a slender, earth-toned branch, a vibrant Painted Bunting displays its breathtaking plumage. The bird's head is adorned with vivid electric blue, its wings and back are bright green, while its chest glows with a fiery red. ...	60	65
003	Two iconic superheroes, Captain America and Iron Man, stand side by side. Captain America is clad in his classic red, white, and blue costume, while Iron Man dons futuristic armor characterized by a red and gold color scheme and a gold helmet. ...	Boldly standing side by side, Thor and Captain America showcase their iconic redesigns. Thor, on the left, wears sleek black armor with gold accents and a flowing red cape, while Captain America stands on the right in a vibrant blue suit with his signature shield. ...	65	65
004	Presented in a sleek display case, a meticulously crafted 1:18 scale model of the iconic 1960s Volkswagen Beetle stands out with its vibrant green body and pristine white accents. The display case features the text "1960s Volkswagen Beetle" and the brand name "Volkswagen." ...	Detailed and meticulously crafted, a vibrant green Austin Mini Cooper 1/43 scale model takes center stage, capturing the timeless charm of the classic 1966 design. The model features a pristine white roof and the iconic "MINI 11 MINI" racing livery on its side. ...	50	50
.....				
100	Elegantly positioned against a pristine white background, the "L'Oréal Paris" perfume bottle stands as a striking centerpiece. The body of the bottle is a rich, vibrant pink, while the brand name "L'Oréal Paris" is prominently displayed on the front. ...	Elegantly presented against a pristine white backdrop, the Balconi perfume bottle exudes sophistication and luxury. The transparent glass body contains a vibrant pink liquid, with the word "BALCONI" prominently displayed in bold black capital letters. ...	70	70
AVG Score	——	——	**69.70**	**70.00**

B.6 Generalization to Unseen Entities and Out-of-Distribution Data

One possible limitation of a prototype-based memory mechanism is that its stored representations are derived from the training distribution. This raises the question of whether the memory module may become overly dependent on previously observed semantic patterns and consequently behave poorly when encountering novel entities or samples drawn from substantially different domains. In the most unfavorable case, an unseen entity could be attracted toward an unrelated prototype by the memory compatibility objective, leading to incorrect semantic alignment under distribution shift.

Table 10: Comparison of model performance across in-distribution and zero-shot OOD benchmarks.

Dataset	Model training	Evaluation setting	Result
FLUX-Reason (Fang et al., 2025)	✓	In-distribution	SOTA
SEED-Bench (Li et al., 2023)	×	Zero-shot OOD	SOTA
A-OKVQA (Schwenk et al., 2022)	×	Zero-shot OOD	SOTA
SurgVLM (Zeng et al., 2025)	×	Zero-shot OOD	SOTA
CC12M (Changpinyo et al., 2021)	×	Zero-shot OOD	SOTA

As listed in Table 10, our evaluation protocol is specifically designed to examine this possibility. The optimization of the proposed model rely solely on FLUX-Reason. None of the samples, annotations, entity information, or prototype statistics from the downstream benchmarks are exposed during

training. In particular, the memory module is never constructed or updated using data from SEED-Bench, A-OKVQA, SurgVLM, CC12M, or the remaining OOD evaluation sets. All OOD benchmarks are evaluated without any additional adaptation, prototype augmentation, or benchmark-specific fine-tuning. This creates a stringent test for the proposed memory mechanism: if successful prediction depended on retrieving a prototype corresponding exactly to each test-time entity, performance should deteriorate considerably once the model encounters concepts or domains absent from FLUX-Reason. A similar degradation would also be expected if the compatibility objective systematically projected novel instances onto semantically mismatched training prototypes. Empirically, we observe the opposite behavior. The proposed method maintains strong performance across all four zero-shot OOD benchmarks and achieves state-of-the-art results under these distribution shifts, including on SurgVLM, whose surgical imagery and domain-specific semantics differ substantially from those present in the general-domain training corpus. Such results suggest that the benefit of the memory module is not tied to memorizing a finite inventory of training entities.

Instead, we interpret the learned entity prototypes as **semantic reference anchors** in the representation space. These anchors provide structural guidance during optimization, encouraging semantically related representations to form coherent and stable organization without requiring an exact correspondence to any prototype at inference time. Consequently, previously unseen entities can still be represented according to their semantic relationships with the learned representation structure, rather than being restricted to a closed set of predefined categories. This allows the memory-enhanced model to retain its generalization capability when encountering novel entities, visual concepts, or substantial domain shifts at test time.

C Hierarchy-Aware Analysis of Fine-Grained Visual Understanding

To investigate whether the proposed method improves not only final-answer accuracy but also the underlying structure of visual understanding, we conduct a hierarchy-aware qualitative analysis at three complementary levels: object, relationship, and layout. Specifically, the object level evaluates the recovery of fine-grained entities and their visual attributes; the relationship level focuses on interactions, relative positions, and spatial dependencies among entities; and the layout level examines whether the model can preserve the global organization of multiple visual elements.

Table 11: Qualitative analysis of hierarchy-aware fine-grained visual understanding.

Hierarchy	Case	Key Visual Information in GT	Information Recovered by `Meta-CogGran`
Object Level	Case 0508	Compartmentalized storage box; multiple colorful beads; beads stored in separate compartments	Correctly identifies the "storage box," "colorful beads," and that the beads are organized in "separate compartments."
Object Level	Case 0526	Pink princess dress; white hanger; sleeveless bodice; bow at the waist; layered tulle skirt	Correctly identifies the "pink princess dress," "white hanger," "sleeveless bodice," waist "bow," and "voluminous tulle layer."
Attribute Level	Case 0506	Dark gray cargo pants; multiple utility pockets; silver-button closures; elastic waistband; black athletic shoes	Recovers the cargo-pants category, multiple functional pockets, dark coloration, and the pairing with black shoes, while preserving the utilitarian appearance.
Attribute Level	Case 0509	Doll with wavy brown hair; white flower hairpin with yellow center; lace dress; plain neutral background	Correctly recovers the "wavy brown hair," white flower accessory with a yellow center, the lace dress, and the clean neutral background.
Relationship Level	Case 0502	The couple walks hand-in-hand; the woman leads slightly ahead; the man follows closely; the lake reflects the surrounding mountains	Correctly captures that the couple walks "hand-in-hand," that the woman "walks ahead" while the man "follows closely," and that the water mirrors the surrounding mountains.
Relationship Level	Case 0529	A wooden spoon rests inside the bowl; cucumber slices surround the bowl; crackers are placed nearby; the bowl occupies the foreground	Correctly captures a "small white bowl" in the foreground, a wooden spoon "inserted into" the dip, and cucumber slices and crackers distributed around the central bowl.
Layout Level	Case 0525	Fruits arranged in concentric circles; blueberries occupy the center; strawberries and kiwis form surrounding rings; the overall arrangement is symmetric	Correctly recognizes the "concentric circles," recovers the central blueberry cluster, and captures the surrounding strawberry–kiwi ring structure and balanced composition.
Layout Level	Case 0520	Three figures arranged from left to right; navy suit on the left, dark gray in the center, light brown suit on the right; background audience is blurred	Correctly distinguishes the "leftmost man," "center figure," and "rightmost man," associates them with distinct suit colors, and recognizes the blurred background.

Table 11 presents representative cases by comparing the key visual information in the ground truth with the corresponding information recovered by our model. The examples reveal a clear progression from local visual details to more structured scene-level representations. At the object level, the model retains detailed characteristics of individual entities, including color, material, clothing structure, and constituent components. At the relationship level, it goes beyond isolated object recognition and correctly captures how entities interact or are spatially related, such as containment, relative ordering, and directional relations. At the layout level, the model further recovers global configurations involving multiple elements, including center–surround organization, concentric arrangements, and left-to-right spatial structures.

These observations indicate that the gains introduced by the proposed memory mechanism are not limited to producing more accurate final descriptions. Rather, the model preserves visual information across multiple levels of abstraction, from fine-grained object attributes to inter-object relations and global scene organization. This hierarchy-aware evidence therefore provides a more direct and interpretable view of how the proposed method strengthens structured fine-grained visual understanding.

D Compute Reporting

All models are trained in a distributed environment using eight NVIDIA A800-SXM4 GPUs, each with 80 GB of memory. We employ `DeepSpeed` with ZeRO Stage-2 to partition optimizer states across devices and reduce the memory footprint associated with large-scale multimodal optimization. This configuration allows different visual encoders and language-model backbones to be trained under a unified computational setting. Each experiment is run for two epochs. The local batch size is set to 8 examples per GPU, and gradients are accumulated over two update steps. With eight devices, this corresponds to an effective batch size of 128 samples for each optimization step. The learning rate is initialized at 2×10^{-5} and follows a cosine decay schedule, with the first 5% of training steps used for linear warmup. To improve memory and computational efficiency, training is performed with bf16 precision while tf32 computation is enabled for supported operations. We additionally activate gradient checkpointing to trade additional computation for lower activation-memory requirements. The input pipeline uses parallel data-loading workers and pinned host memory to reduce data-transfer latency and prevent I/O from becoming a major training bottleneck. Checkpointing is performed periodically throughout optimization, with only the three most recent model states retained. This policy limits disk usage while preserving sufficient intermediate states for recovery and model selection. We do not perform validation during the training trajectory; therefore, computational resources are concentrated on parameter optimization and cross-modal representation learning rather than repeated inference on held-out data.

Overall, the training configuration is chosen to provide a practical balance between memory efficiency, throughput, and scalability. In particular, the combination of ZeRO Stage-2, mixed-precision computation, gradient checkpointing, and optimized data loading enables stable multimodal training across multiple model configurations without requiring changes to the underlying hardware environment.

E Evaluation Prompts

We adopt a multimodal evaluator to automatically assess the quality of generated image descriptions. Rather than relying on a single aggregate score, the evaluation considers two complementary dimensions: *accuracy* and *hallucination*. Together, these metrics characterize whether a generated description is both visually correct and free from unsupported content.

For each evaluation instance, the judge receives the original image together with the description produced by the target model. It is instructed to ground its assessment exclusively in the visual evidence contained in the image and to evaluate the textual response accordingly. The accuracy score is defined on a scale from 0 to 100, where larger values indicate stronger correspondence between the generated description and the observed visual content, including objects, attributes, actions, and their relationships. Hallucination is scored independently on a scale from 0 to 80 and measures the amount of information introduced by the description that cannot be verified from the image. Thus, a lower hallucination score indicates a more visually faithful response. To facilitate large-scale automatic evaluation, the judge is required to return its decision in a constrained machine-readable format. Specifically, each response must contain a valid JSON object with two integer-valued fields corresponding to the accuracy and hallucination scores. Responses that violate the prescribed schema

are treated as invalid and excluded from subsequent aggregation. This structured-output requirement minimizes parsing ambiguity and enables the complete evaluation pipeline to operate without manual intervention. At inference time, each image is converted into a base64 representation and submitted jointly with the scoring instruction and the candidate description as a multimodal request. Samples are evaluated independently using a low-temperature decoding configuration to reduce stochastic variation in the judge outputs and improve score consistency across repeated evaluations. After all samples have been processed, valid scores are collected and averaged over the evaluation set. We report the mean accuracy and mean hallucination score as the final summary statistics. This evaluation procedure provides two complementary signals: the former measures semantic agreement with the image, while the latter explicitly penalizes visually unsupported statements. By combining visually grounded instructions, explicit scoring definitions, deterministic decoding, and a constrained response schema, the protocol aims to provide a reproducible and scalable assessment of multimodal generation quality.

GPT-4o Evaluator Prompt

You are an objective evaluator for AI image descriptions.
Evaluate the following **MODEL OUTPUT** against the **PROVIDED IMAGE**, and assign two independent scores.

1. Accuracy Score (0–100): how well the text matches the actual image content.
2. Hallucination Score (0–80): how much content in the text is not present in the image.

MODEL OUTPUT: `{pred}`

Return **ONLY** a valid JSON object:

```
{"accuracy_score": int, "hallucination_score": int}
```

F More Layer-wise Alignment Analysis

We further investigate the layer-wise evolution of visual features under two different LLM backbones. As illustrated in Figures 8, 9, 10, 11, `Meta-CogGran` exhibits smoother and more coherent representation dynamics across successive network layers compared with alternative visual encoders. The reduced fluctuation in feature transitions indicates that the proposed method is better able to retain semantically relevant visual information during progressive cross-modal processing. Such representational stability facilitates more consistent semantic alignment with the language model and provides a stronger foundation for downstream multimodal reasoning.

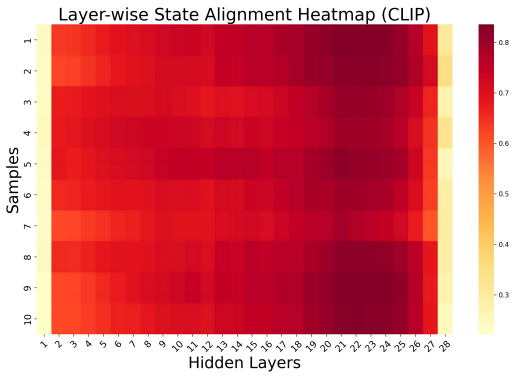


Figure 8: Cosine similarity between the reference state and the state of each corresponding layer for Clip + qwen2.5-7B

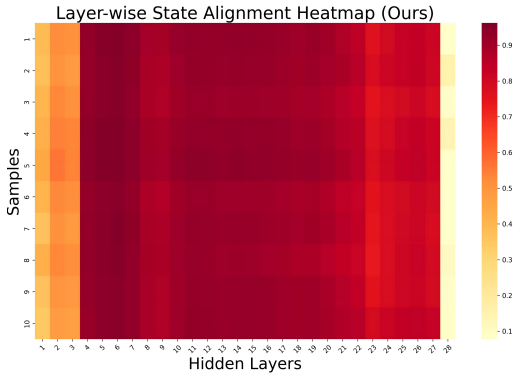


Figure 9: Cosine similarity between the reference state and the state of each corresponding layer for Ours + qwen2.5-7B

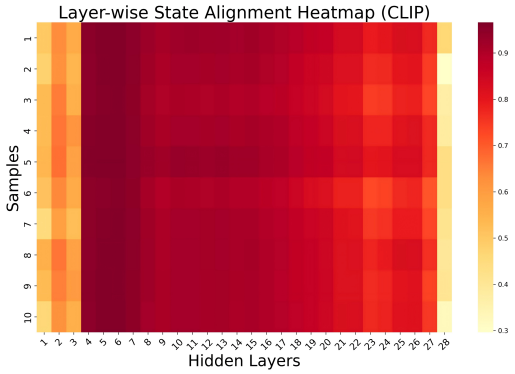


Figure 10: Cosine similarity between the reference state and the state of each corresponding layer for Clip + qwen3-8B

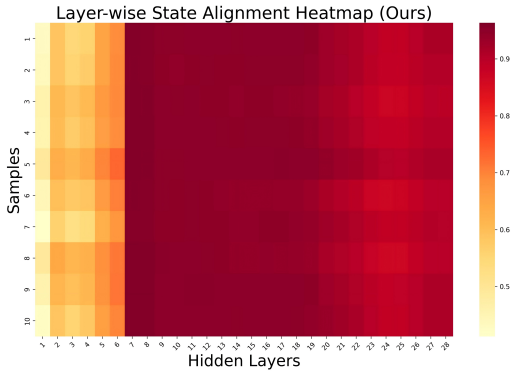


Figure 11: Cosine similarity between the reference state and the state of each corresponding layer for Ours + qwen3-8B

G More Qualitative Case Study

To provide a more intuitive understanding of the generation behavior of `Meta-CogGran`, we further examine a collection of representative qualitative examples from different evaluation benchmarks, as illustrated in Figure 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23. These cases cover diverse visual scenes and reasoning requirements, allowing us to inspect not only whether the generated descriptions are correct, but also how faithfully different methods preserve the semantic structure of the input image.

Across these examples, `Meta-CogGran` generally produces descriptions that remain closely grounded in the observed visual evidence. The generated responses capture the principal objects, their attributes, and the semantic relationships among them with relatively high consistency. More importantly, the model tends to preserve the global interpretation of the scene even when individual fine-grained details are imperfect. The remaining errors are typically localized to subtle properties, such as specific object attributes, small-scale visual cues, or ambiguous details that require precise local discrimination. Such mistakes therefore have limited influence on the overall semantic correctness of the generated description, indicating that `Meta-CogGran` maintains a comparatively stable representation of the dominant visual concepts. A different error pattern can be observed for approaches based on `CLIP`. Although these methods are often able to identify several salient concepts in the image, their generated descriptions exhibit weaker correspondence with the complete visual context and are more susceptible to hallucination. In particular, an initially incorrect prediction is frequently followed by additional details that are semantically plausible with respect to the mistaken concept but are not actually supported by the image. This creates a cascading error effect: once the representation is biased toward an incorrect semantic interpretation, subsequent generation tends to elaborate on that interpretation, progressively introducing more unsupported objects, attributes, or relationships. The qualitative comparison therefore reveals an important distinction between isolated recognition errors and semantic drift. For `Meta-CogGran`, prediction errors are more often confined to specific fine-grained details while the global scene representation remains largely intact. In contrast, errors produced by `CLIP`-based representations can propagate through the generation process and alter the semantic trajectory of the entire response. As a consequence, the latter may produce descriptions that appear linguistically coherent yet gradually deviate from the actual visual evidence.

These observations are consistent with the quantitative improvements in both description accuracy and hallucination control. They suggest that the advantage of `Meta-CogGran` is not limited to recognizing more visual concepts, but also lies in maintaining a more reliable visual-semantic representation throughout downstream language generation. By providing more stable and discriminative visual evidence to the language backbone, the proposed representation reduces the likelihood that an early semantic error develops into a sequence of mutually reinforcing hallucinated details. This property is particularly important for multimodal reasoning tasks, where faithful grounding must be preserved over multiple stages of interpretation and generation.

Case Study 1: Reasoning Answer



CLIP: *Two charming dolls, each dressed in elegant attire, stand side by side on a wooden table, their presence captured in a serene indoor setting. The first doll, adorned in a pastel pink dress with delicate floral patterns, features a matching pink bonnet and a golden tiara, exuding a classic and timeless charm. Her long, blonde hair cascades gracefully, complementing her refined appearance. The second doll, dressed in a soft yellow gown with intricate gold embroidery, wears a matching yellow bonnet and a golden tiara, embodying a more modern and sophisticated style. Her dark hair is styled in a neat bob, adding to her poised demeanor. Both dolls are positioned symmetrically, with the pink-dressed doll slightly to the left and the yellow-dressed doll to the right, creating a balanced composition. The background is softly lit, with bookshelves and a cozy indoor environment providing a warm and inviting backdrop. The overall scene is photorealistic, capturing the intricate details of the dolls' clothing and accessories, while the natural lighting enhances the textures and colors, making the image both visually appealing and nostalgic.*

OURS: *Question: Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible.\n<image>*

Answer: *This photograph showcases two Barbie dolls dressed in elegant, pastel-inspired outfits, set against a warm, neutral background that highlights their vibrant colors. The first doll, with long, curly blonde hair styled in a classic updo adorned with a delicate white headband, wears a soft pink dress featuring intricate floral patterns in shades of pink, white, and cream. The dress is adorned with ruffles at the neckline and sleeves, adding a touch of whimsy and femininity. The second doll, with shorter, wavy blonde hair styled in a matching updo, wears a matching pink dress with similar floral patterns, but with a slightly different design. Both dresses are completed with white aprons that feature a subtle floral pattern, adding a cohesive and harmonious touch to the overall ensemble. The dolls stand side by side, their outfits blending seamlessly into a harmonious palette of soft pinks and whites, emphasizing their elegance and sophistication. The composition is centered, with the dolls positioned symmetrically, and the background features a bookshelf with various titles, suggesting a cozy, homey setting. The overall style is photorealistic, capturing the intricate details of the dresses and the dolls' accessories, such as their hair and jewelry, in a way that feels both realistic and stylized.*

Ground Truth: *Two American Girl dolls, Marie Grace and Cecile, stand side by side in a warm, cozy indoor setting, each adorned in meticulously crafted historical attire that reflects elegance and period-inspired design. Marie Grace, on the left, wears a charming white dress embellished with delicate floral patterns in shades of yellow, pink, and green, complemented by a matching white collar and ruffled sleeves. Her ensemble is completed with a soft pink bonnet featuring a scalloped edge, adding a touch of innocence and charm. To her right, Cecile is dressed in a regal yellow gown with a fitted bodice adorned with decorative buttons and a lace-trimmed collar, exuding a sense of royalty. Her dark hair is crowned with a golden tiara, further enhancing her princess-like appearance. Both dolls are positioned against a backdrop of wooden bookshelves filled with books, suggesting a homey and intellectual environment. The natural lighting and wooden floor create a serene and inviting atmosphere, highlighting the intricate details of their costumes and the timeless beauty of their historical outfits. The composition emphasizes the contrast between the two characters while celebrating the rich heritage and craftsmanship of American Girl dolls.*

Figure 12: Qualitative case study comparing CLIP and our method.

Case Study 2: Reasoning Answer



CLIP: *The image depicts a model standing in front of a white backdrop, wearing a stunning wedding dress with intricate details and a flowing train. The dress features a fitted bodice with delicate lace and a sweetheart neckline, complemented by a full skirt that falls in soft folds to the floor. The model's pose exudes elegance and poise, highlighting the dress's versatility and beauty. The overall composition emphasizes the dress's design, texture, and movement, making it a quintessential representation of trendy wedding fashion.*

OURS: *Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible.\n<image>*
Answer: *Elegant and poised, a woman stands against a neutral gray backdrop, showcasing a stunning white wedding dress that exudes timeless sophistication. The dress features a delicate sweetheart neckline adorned with sheer lace detailing, which gracefully frames her face and adds a touch of romantic charm. The bodice is embellished with intricate beadwork and a subtle waist cinch, accentuated by a sparkling belt that catches the light, enhancing the dress's luxurious appeal. The skirt cascades into a voluminous, flowing train, crafted from soft, flowing fabric that drapes elegantly around her feet, creating a sense of movement and grace. Her hair is styled in loose waves, complementing the dress's ethereal design, while she wears matching pearl earrings and a bracelet, adding a refined touch to her ensemble. The overall composition is minimalist yet striking, with the dress taking center stage against the plain background, highlighting its intricate details and flowing silhouette. This image captures the essence of bridal elegance, blending classic elements with modern refinement.*
Ground Truth:*Modelled elegantly in Justin Alexander Style 88122, the Venice Tulle Ball Gown exudes timeless sophistication against a minimalist gray backdrop. The gown features an off-the-shoulder design with delicate lace detailing at the neckline, seamlessly transitioning into a fitted bodice adorned with intricate embellishments that highlight its artisanal craftsmanship. A gold belt cinches the waist, adding a touch of opulence and drawing attention to the figure before cascading into a voluminous tulle skirt that flows gracefully to the floor. The skirt's layered ruffles create a dramatic yet ethereal silhouette, enhancing the ball gown's romantic appeal. The model accessorizes with subtle yet striking jewelry, including pearl earrings and a delicate bracelet, which complement the gown's refined aesthetic. The overall composition is balanced and poised, capturing the essence of bridal elegance with a blend of classic and contemporary elements. The photograph is rendered in a photorealistic style, emphasizing the gown's texture and the soft, diffused lighting that accentuates its intricate details.*

Figure 13: Qualitative case study comparing CLIP and our method.

Case Study 3: Reasoning Answer



CLIP: *A vintage 1970 Ford Bronco, painted in a classic red hue, is prominently displayed in a residential setting, showcasing its iconic design and rugged charm. The vehicle is parked on a paved street, with a clear view of the surrounding environment, including palm trees, a white house with a tile roof, and a clear blue sky with fluffy white clouds. The Bronco's rear window is open, revealing a spacious interior with a wooden dashboard and a prominent F̈ordL̈ogo on the rear panel. The license plate reads B̈-1-1-1,ädding a personal touch to the vehicle's vintage appeal. The composition highlights the Bronco's sturdy frame, complete with black rubber tires and a prominent grille, while the lush greenery and clear blue sky enhance the warm, sunny atmosphere of the scene. The overall aesthetic is photorealistic, capturing the essence of a bygone era with meticulous attention to detail, making it a compelling and nostalgic image.*

OURS: *Question: Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible.\n<image>*

Answer: *Parked on a quiet suburban street under a clear blue sky, a classic 1970s Ford Bronco stands out with its pristine white exterior and bold red accents, exuding timeless charm and rugged appeal. The vehicle's rear end is prominently visible, showcasing its iconic design with a sleek, boxy profile and a distinctive F̈ORDëmblem on the license plate. The license plate reads Ï234 ABC, adding a touch of authenticity to the scene. The Bronco's black wheels and robust tires emphasize its off-road capabilities, while the surrounding environment features palm trees, a row of houses with red roofs, and a subtle hint of a garage door in the background, suggesting a residential neighborhood. The composition captures the car from a slightly angled perspective, highlighting its proportions and detailing, with the sun casting a bright, natural light that enhances the vehicle's metallic finish. The overall atmosphere is one of nostalgia and simplicity, perfectly encapsulating the essence of a vintage classic in a serene suburban setting.*

Ground Truth: *Parked along a sunlit street in San Diego, California, a pristine 1966 Ford Bronco (CC-1020675) exudes vintage charm with its sleek silver exterior and classic design. The vehicle, identified by its iconic boxy shape and black grille, stands out against the backdrop of palm trees and suburban homes, capturing the essence of mid-century Americana. The California license plate, reading Ö11865,ädds authenticity to the scene, while the car's well-maintained condition reflects its timeless appeal. The rear view highlights the Bronco's robust build, featuring round taillights, a spare tire mounted on the back, and a distinctive black grille with horizontal bars. The surrounding environment, with neatly trimmed hedges, red-tiled roofs, and clear blue skies, enhances the nostalgic atmosphere, suggesting a serene neighborhood where history meets modernity. The image, rendered in photorealistic detail, emphasizes the Bronco's enduring legacy and its seamless integration into the sunny, palm-lined streets of San Diego.*

Figure 14: Qualitative case study comparing CLIP and our method.

Case Study 4: Reasoning Answer



CLIP: *The image presents a beautifully crafted glass apothecary jar, its clear transparent body showcasing a delicate floral pattern etched into the glass, adding a touch of elegance and vintage charm. The jar, with a rounded base and a slender neck, is positioned on a wooden table, which provides a warm and natural background. The jar's lid, adorned with a silver handle, is slightly ajar, revealing a subtle gradient of light within, highlighting its clarity and craftsmanship. To the left of the jar, a small wooden frame with a subtle floral pattern frames a delicate pink flower, its petals softly unfolding towards the jar, creating a harmonious balance of form and nature. The overall composition is serene and inviting, with the jar's transparency and the flower's gentle beauty drawing the viewer's attention to the intricate details of the apothecary's design. The soft, warm lighting enhances the textures and colors, creating a cozy and nostalgic atmosphere reminiscent of traditional apothecary practices.*

OURS: *Question: Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible.\n<image>*

Answer: *Elegant and minimalist, the image showcases a clear glass vase with a polished, cylindrical shape, featuring a sleek, transparent design that highlights its simple yet sophisticated form. The vase, adorned with a subtle gold rim, stands prominently on a wooden table, its smooth surface reflecting the soft, ambient lighting that bathes the scene in a warm, inviting glow. To the left, a delicate white flower with vibrant pink petals rests on the table, adding a touch of natural beauty and contrast to the clean, modern aesthetic. In the background, a framed artwork with intricate patterns and a rich, earthy color palette complements the vase, creating a harmonious blend of art and functionality. To the right, a beige cylindrical object, likely a decorative item or another vase, adds depth and balance to the composition. The overall arrangement exudes a sense of calm and elegance, with the clear glass vase serving as a focal point that draws the eye while the surrounding elements enhance its visual appeal. The photograph captures a serene, well-curated space that emphasizes the beauty of simplicity and the interplay of textures and colors.*

Ground Truth: *Centered on a warm, polished wooden surface, an elegant, transparent glass bottle with a stopper stands as the focal point, its smooth, reflective surface capturing the ambient light. To its left, a delicate white flower with soft petals and a vibrant orange center adds a touch of natural beauty, while to its right, a cylindrical lamp with subtle gold accents provides a sophisticated contrast. The background features green wallpaper adorned with intricate golden leaf patterns, enhancing the overall aesthetic with a harmonious blend of nature-inspired design. The composition exudes a sense of refined elegance, reminiscent of the classic Avon Coleman Oil Lantern Deep Woods Cologne, where the empty bottle suggests a moment of pause or anticipation. The arrangement is meticulously balanced, with each element contributing to a serene and inviting atmosphere, characteristic of a carefully curated interior space. The image's photorealistic style captures the textures and reflections with precision, emphasizing the timeless appeal of the objects and their thoughtful placement within the scene.*

Figure 15: Qualitative case study comparing CLIP and our method.

Case Study 5: Reasoning Answer



CLIP: *The image features a beautifully crafted vanilla cake, its smooth, creamy surface adorned with a delicate swirl pattern of white frosting, creating a visually appealing contrast. The cake is placed on a pristine white cake stand, which contrasts sharply with the dark, textured surface beneath, enhancing its elegance. In the background, a lush green succulent plant adds a touch of natural beauty, while a few other plants are visible through the window, suggesting a bright and airy setting. The overall composition is clean and minimalist, with the cake as the central focus, drawing attention to its intricate design and the harmonious balance of colors. The scene is set in a home environment, with a subtle hint of a kitchen or dining area, emphasizing the cake's role as a centerpiece for a special occasion or celebration. The image exudes a sense of sophistication and refinement, capturing the essence of a well-crafted dessert.*

OURS: *Question: Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible.\n<image>*

Answer: *Vibrant and inviting, a beautifully crafted lemon cake takes center stage on a pristine white cake stand, its surface adorned with smooth, creamy frosting that transitions seamlessly from a bright yellow base to a soft green top, creating a striking gradient effect. The cake's edges are slightly uneven, adding a rustic charm, while the frosting is meticulously piped, forming delicate swirls and patterns that enhance its visual appeal. Set against a neutral backdrop of a speckled kitchen counter, the cake is complemented by two potted plants in white ceramic pots, their lush greenery providing a fresh and natural contrast. The overall composition is clean and balanced, with the cake's vibrant colors drawing the eye while the plants and background elements add depth and context. The image exudes a sense of warmth and creativity, capturing the essence of a homemade treat that is both visually appealing and inviting.*

Ground Truth: *Lemon Ombre Layer Cake, elegantly displayed on a pristine white cake stand, captures the eye with its vibrant gradient of yellow and white frosting, transitioning seamlessly from a bright, sunny yellow at the base to a pure white at the top. The cake's surface is adorned with a delicate spiral pattern of Swiss meringue buttercream frosting, adding texture and sophistication. The layers within suggest a delightful lemon sponge cake filled with silky smooth lemon curd, promising a burst of citrus flavor. The setting is warm and inviting, with the cake positioned on a speckled countertop, complemented by the soft green hues of potted plants in the background, which enhance the cozy and natural ambiance. This beautifully crafted dessert, showcased by BirdsParty.com (@birdsparty), exemplifies meticulous cakedesign and a perfect balance of color and texture, making it a visual treat as well as a culinary delight. #lemoncake #layercake #ombrecake #lemonombrecake #lemoncurdcake #lemoncurd #cakedesign #cakerecipe.*

Figure 16: Qualitative case study comparing CLIP and our method.

Case Study 1: Visual Question Answering (VQA)



CLIP: *The scene unfolds on a busy street in London, where a cameraman, dressed in casual attire and carrying a backpack, is positioned at the edge of the road, likely preparing to capture a shot. To his left, a red double-decker bus drives forward, its presence marked by a clear path and a prominent sign indicating the route. Behind the bus, a row of buildings with red brick facades stretches across the horizon, their windows and doors visible through the misty veil of the morning. The cameraman, positioned near the center of the frame, is seen walking towards the bus, his path illuminated by the soft glow of streetlights and the distant glow of the cityscape. The background is dotted with other vehicles, including a black car and a blue van, adding to the dynamic and bustling atmosphere. The overall composition emphasizes the cameraman's role as a stationary observer, poised to capture the movement and activity of the street, while the cityscape behind him suggests a broader narrative of urban life and activity. The image is rendered in a photorealistic style, capturing the detailed textures and colors of the scene, including the cameraman's clothing and the vehicles' metallic surfaces. \n\nAnswer: (B) drive straight.*

OURS: *Please answer the question based on the image and explain your reasoning step by step.\nQuestion: If the cameraman were driving what do they have to do from this position?\n Choices:\n (A) turn left\n (B) drive straight\n (C) reverse course\n (D) turn right\n <image>*

Answer: *If the cameraman were driving from this position, they would need to* ***turn right****. The image depicts a busy street with multiple lanes, including a dedicated bike lane marked by a white line and a bicycle symbol. The cameraman is positioned near the bike lane, facing a group of people walking and a parked car. To the right, there is a clear path for vehicles to continue straight or turn right, while the left side is obstructed by parked cars and a bus stop. The presence of a bus stop and the alignment of the bike lane suggest that turning right would be the most logical and safe option for the cameraman to proceed without interfering with pedestrians or other vehicles. This decision is further supported by the absence of any clear indication to turn left or drive straight, making turn rightthe most appropriate choice among the given options.*

Ground Truth: *Reasoning: The would have to turn right because the lane has right turn arrows painted on it. The arrow on the street indicates that this lane can only go in one direction at the intersection. The sign on the road says to turn right.\nFinal answer: turn right*

Figure 17: Qualitative case study comparing CLIP and our method.

Case Study 2: Visual Question Answering (VQA)



CLIP: *Image Credit: Getty Images\nThe poster for the film T̈he Grand Budapest Hotelf̈eatures a prominent image of Ralph Fiennes as Gustave H, the charming concierge of the titular hotel, standing in front of a grand, ornate façade. He is dressed in a stylish outfit, including a white shirt, black trousers, and a black vest, with a red pocket square adding a pop of color. His hands are clasped together, conveying a sense of confidence and authority. In the background, the hotel's imposing structure is visible, with intricate details such as balconies, windows, and a clock tower, emphasizing the grandeur and luxury of the setting. The image is set against a neutral gray background, allowing the subject to take center stage. The title T̈he Grand Budapest Hotelïs prominently displayed in bold, white letters at the top of the poster, while the tagline Ä Comedy of Mannersïs written in smaller text below, hinting at the film's themes of elegance and wit. The overall design is classic and sophisticated, capturing the essence of the movie's whimsical and visually stunning world.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\nQuestion: Where would one most likely see the show advertised in the poster?\n Choices:\n (A) theater\n (B) tv\n (C) internet\n (D) cinema\n <image>*

Answer: *The poster prominently features a man in a red coat, standing confidently on a sidewalk, holding a black briefcase, with a newspaper in hand. The background suggests a bustling urban environment, with buildings and streetlights visible, indicating a public space such as a city street. The man's attire—formal yet casual—suggests he is likely a professional or someone who values both style and practicality. The presence of a newspaper implies that the show advertised in the poster is a significant event, possibly a theatrical performance or a major cultural happening. Given the setting and the man's demeanor, it is most likely that one would see the show advertised in a* ***theater*** *(A), as theaters are common venues for such events and often feature posters in high-traffic areas like sidewalks. The composition of the image, with the man positioned centrally and the newspaper in his hand, further emphasizes the importance of the advertisement, making the theater the most plausible location for its display.*

Ground Truth: *Reasoning: Most likely the sign is depicting a broadway show at a theater. The poster would be in a theater. This type of display with a character and the added writing is usually used to advertise either a movie or play both of which could be viewed in a theater.\nFinal answer: theater*

Figure 18: Qualitative case study comparing CLIP and our method.

Case Study 3: Visual Question Answering (VQA)



CLIP: *The blue car, positioned in the center of the frame, is surrounded by a group of vehicles, including a bus, a motorcyclist, and a cyclist, all moving in different directions. The bus, with its bright blue and white livery, is heading straight ahead, while the motorcyclist and cyclist are moving in opposite directions, their paths crossing near the blue car. The cyclist, wearing a helmet and riding a bike with a basket, is positioned on the left side of the frame, and the motorcyclist, wearing a helmet and riding a bike with a sidecar, is on the right. The blue car, with its sleek design and vibrant color, stands out as the central object, while the surrounding vehicles create a dynamic and chaotic scene, highlighting the potential dangers of road traffic. The image captures a moment of high tension and urgency, emphasizing the importance of safety and awareness on the road.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\n Question: Which object is in the greatest danger?\n Choices:\n (A) bus\n (B) motorcyclist\n (C) blue car\n (D) cyclist\n <image>*

Answer: *A cyclist, wearing a helmet and riding a bicycle, navigates through a wet, rainy environment, with water splashing up from the road as they move forward. The scene is set against a backdrop of a large, gray building with a sign reading Ï00änd a red car parked nearby. The cyclist's path is partially obscured by the building, suggesting a busy urban setting. The water on the road reflects the surroundings, adding to the sense of motion and the wet conditions. The choices provided (A) bus, (B) motorcyclist, (C) blue car, and (D) cyclist indicate that the cyclist is the object in the greatest danger due to their exposure to the elements and the potential hazards of navigating through a wet, possibly slippery road. The cyclist's vulnerability is highlighted by their need to wear a helmet and the presence of water, which could increase the risk of accidents.*

Ground Truth: *Reasoning: The water is very deep for the bike. The person on the bicycle might get splashed by water. The cyclist is most likely to have problems because of the rain.\nFinal answer: cyclist*

Figure 19: Qualitative case study comparing CLIP and our method.

Case Study 4: Visual Question Answering (VQA)



CLIP: *The scene depicts a dramatic collision between a blue car and a motorcyclist, with the latter seen splashing into the water, creating a powerful splash that reflects off the wet surface. The motorcyclist, identified by a helmet, is positioned near the center of the image, while the blue car, with its license plate clearly visible as "B 1 1 1," is partially submerged in the water, its front end submerged and its rear end still above the surface. The bus, marked with the number B̈ 1 1 1änd a logo, is seen in the background, its presence highlighted by the blurred motion of the water. The overall composition emphasizes the intense movement and danger of the scene, with the motorcyclist's precarious position and the car's partial submersion creating a sense of urgency and risk. The image is rendered in a photorealistic style, capturing the dynamic interaction between the vehicles and the water.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\n Question: The shade is causing the motorcyclists to turn what on?\n Choices:\n (A) windshield wipers\n (B) headlights\n (C) radio\n (D) turning signal\n <image>*

Answer: *Motorcyclists are seen turning on their* ***headlights*** *as they navigate through a shaded area, casting long shadows across the pavement. The scene is set in a residential neighborhood with neatly arranged houses and trees lining the street, creating a serene yet functional environment. The motorcyclists, dressed in protective gear, are positioned in front of parked cars, with one rider holding a helmet and another adjusting their visor. The sunlight filters through the trees, highlighting the contrast between the shaded and sunlit areas, while the pavement reflects the movement of the riders. The choices provided—(A) windshield wipers, (B) headlights, (C) radio, (D) turning signal—are all plausible actions for motorcyclists, but the context of the shaded area and the need for visibility suggests that the most appropriate action is turning on the* ***headlights****. This decision ensures safety and visibility in low-light conditions, which is particularly important when riding in shaded areas.*

Ground Truth: *The shade is causing the motorcyclists to turn on their headlights because they need to see clearly to ride. One can see that the motorcycle lights are on. There are lights visibly on in the front of each motorcycle. the sun is clearly out but there may be enough shade for them to turn the lights on out of precaution.\nFinal answer: headlights*

Figure 20: Qualitative case study comparing CLIP and our method.

Case Study 5: Visual Question Answering (VQA)



CLIP: *The contest described is a trivia competition, where participants are engaged in a lively discussion, likely answering questions related to the theme of the event. The setting is a large, well-lit room with tables and chairs arranged in rows, suggesting a structured format for the competition. The participants, dressed in casual attire, are seated at their designated tables, engaged in animated conversations and laughter, indicating a relaxed yet focused atmosphere. The event is likely a charity fundraiser or a community event, given the mention of C̈harityänd the presence of a large crowd. The overall tone is festive and engaging, with participants actively participating in the trivia contest, suggesting a fun and interactive experience.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\n Question: What type of contest is being held?\n Choices:\n (A) spelling\n (B) running\n (C) trivia\n (D) eating\n <image>*

Answer: *At a lively outdoor event, a group of people gathers around a colorful table adorned with vibrant pastel-colored donuts, each topped with an array of sweet and colorful toppings. The scene captures a moment of casual interaction as individuals, dressed in casual attire, engage in conversation while enjoying the treats. A young person, seated at the table, holds a cup of a beverage, likely coffee or tea, as they chat with others. The atmosphere is festive, with a backdrop of a tent and a sign reading "Donut Contest," suggesting a friendly competition among participants. The table is surrounded by various beverages in blue cups, adding to the lively ambiance. The overall composition highlights the joy and camaraderie of the event, with the focus on the donuts and the people enjoying them, creating a warm and inviting scene.*

Ground Truth: *Reasoning: This is an eating contest between the man and the woman. Both people sitting at the table have food in front of them. The eating contest is held.\nFinal answer: eating*

Figure 21: Qualitative case study comparing CLIP and our method.

Case Study 1: Image Captioning



CLIP: *The image depicts a close-up portrait of a man dressed in a festive Santa Claus costume, set against a warm, wooden backdrop. The man's face is illuminated by soft, golden light, accentuating his bright red beard and matching hat, which is adorned with a white pom-pom. His eyes are cast downward, giving a hint of introspection or contemplation, while his eyebrows are slightly raised, suggesting a mix of curiosity and amusement. The costume is a vibrant red, complementing the warm tones of the wooden structure, which appears to be a rustic wooden frame or a wooden box. The overall composition is simple yet evocative, capturing the essence of the holiday season with a focus on the man's expressive facial features and the cozy, inviting atmosphere of the wooden setting. The image exudes a sense of nostalgia and warmth, evoking feelings of joy and festivity associated with Christmas.*

OURS: *Question: Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible.\n<image>*

Answer: *Wearing a traditional red Santa hat adorned with a white pom-pom and a matching red suit, a man with a long, flowing beard and a thoughtful expression sits inside a wooden cabin, his gaze directed downward with a sense of contemplation. The cabin's interior is dimly lit, casting a warm, rustic glow that highlights the natural wood grain and the soft texture of the man's clothing. The background features a simple, uncluttered space, emphasizing the focus on the figure within. The composition is centered, with the man positioned slightly off-center to the right, creating a balanced yet intimate feel. The overall style is photorealistic, capturing the details of the man's facial features, the texture of his beard, and the subtle shadows that add depth to the scene. The image evokes a sense of quiet reflection, suggesting a moment of introspection or perhaps a pause in the midst of holiday preparations. The use of warm tones and the rustic setting contribute to a cozy, inviting atmosphere, while the man's attire and expression hint at a deeper narrative or emotional state.*

Ground Truth: *Rare Exports: A Christmas Tale*

Figure 22: Qualitative case study comparing CLIP and our method.

Case Study 2: Image Captioning



CLIP: *The image presents a charming wooden wall display, showcasing a rustic yet vibrant arrangement of stacked logs, each roughly cylindrical in shape, forming a layered structure. The logs are arranged in a staggered pattern, with some leaning against each other, creating a visually appealing and dynamic composition. At the center of the display, a small wooden shelf or ledge supports a pair of ceramic pots, one containing a bright red apple and the other a green leaf, adding a pop of color and natural elements to the arrangement. The shelf itself is adorned with a small, weathered wooden block, which serves as a base for the pots. The overall design is rustic and earthy, with the logs providing a natural, textured backdrop for the vibrant accents of the apple and leaf.* $\backslash n \backslash n$*The arrangement is framed by a wooden frame, likely made from the same rustic material as the logs, which adds to the overall aesthetic of the display. The frame is positioned at a slight angle, creating a sense of depth and visual interest. The surface beneath the display is covered in a layer of fallen leaves, adding a touch of autumnal charm to the scene. The arrangement is situated in a garden or outdoor setting, with the surrounding foliage and the presence of a few scattered leaves suggesting a natural, outdoor environment.*

OURS: *Question: Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible.*$\backslash n$*<image>*

Answer: *A rustic wooden structure, likely a shed or storage area, is prominently featured in the composition, with its weathered wooden planks and natural textures creating a warm, earthy aesthetic. The structure is adorned with various decorative elements, including a vibrant red and green vase filled with flowers, a small ceramic pot with a red flower, and a wooden block with a carved design, all strategically placed to add visual interest and a touch of color. The arrangement of the wooden planks forms a stable and inviting framework, with some pieces leaning slightly against each other, suggesting a handcrafted, organic feel. The background is dominated by a pile of neatly stacked wooden logs, their rough surfaces and varied shades of brown adding depth and texture to the scene. A few leaves and twigs are scattered on the ground, enhancing the natural, outdoor ambiance. The overall composition is balanced, with the decorative elements drawing the eye to the center while the wooden structure and logs provide a solid foundation. The image captures a harmonious blend of nature and human craftsmanship, evoking a sense of tranquility and simplicity. The lighting is soft and natural, highlighting the textures and colors of the materials, and the perspective is slightly elevated, offering a clear view of the entire setup.*

Ground Truth: *Woods log in different sizes put together making beautiful art background and very unique wall. Chipped firewood is on the heap. Woodpile of firewood close-up. A stock images*

Figure 23: Qualitative case study comparing CLIP and our method.