META-MEMGRAN: RETHINKING PIXEL-LEVEL VISUAL ENCODERS VIA META-MEMORY 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 global semantic alignment, they often struggle to preserve fine-grained visual details. 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 multi-granularity reasoning. These complementary limitations reveal a fundamental bottleneck: existing visual representations struggle to simultaneously preserve perceptual details and organize them into high-level semantics. We argue that the key challenge lies not merely in the visual encoder, but in the lack of an adaptive mechanism for transforming dense visual evidence into representations at different levels of abstraction This raises a central question: ***How can fine-grained perception be effectively bridged with high-level semantics for multi-granularity multimodal reasoning?*** To address this challenge, we propose `Meta-MemGran`, a `DINOv3`-based MLLM with meta-memory-driven representation learning. By structuring and dynamically aggregating visual information through meta-memory, `Meta-MemGran` bridges pixel-level details, visual structures, and semantic abstractions without explicit granularity selection. Extensive experiments show that `Meta-MemGran` achieves an accuracy improvement of $\sim\mathbf{40}\%\uparrow$ and a hallucination reduction of $\sim\mathbf{30}\%\downarrow$. Code is available at Meta-MemGran.

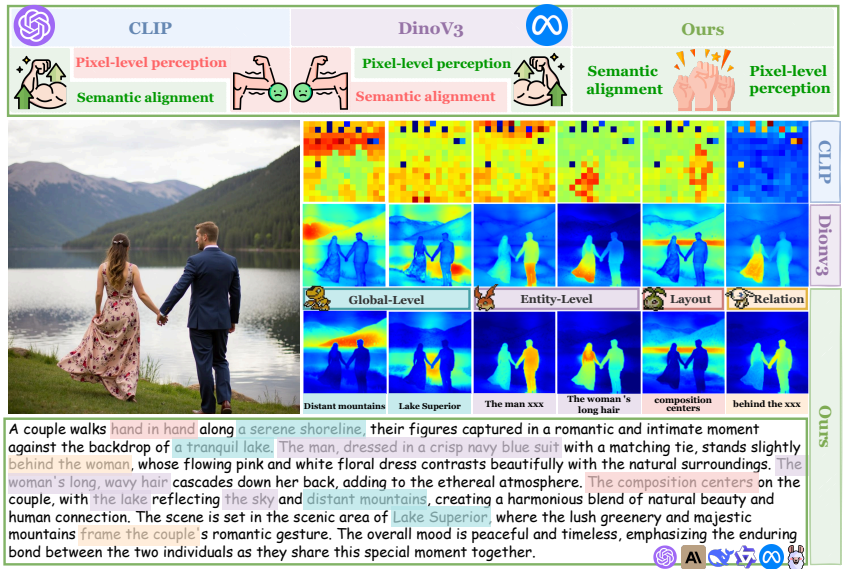


Figure 1: Schematic of `Meta-MemGran`, which organizes visual information from Global, Entity, Layout, and Relation perspectives to bridge pixel-level perception and semantic abstraction for multi-granularity reasoning.

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 substantially improved visual reasoning (Pang et al., 2025; Yang et al., 2026; Dong et al., 2025), captioning (Sarto et al., 2025; Chen et al., 2025a; Yu et al., 2025), and cross-modal understanding (Fu et al., 2025; Zhang et al., 2026b; Wu et al., 2025). Despite this progress, most existing MLLMs still rely heavily on semantically aligned visual representations (Huang et al., 2025; Shen et al., 2026; Zhang et al., 2025c; Jiang et al., 2025a), where visual encoders often struggle to reconcile fine-grained perception with high-level semantic abstraction (Yu et al., 2026; Jiang et al., 2025b; Wang et al., 2025a; Ke et al., 2025b; Liu et al., 2025c; Mao et al., 2025). This limitation restricts the model's ability to perform structured reasoning over complex scenes, as pixel-level tokens lack explicit organization across different semantic granularities (Ji et al., 2025; Williams, 2025; Liu & Lian, 2026).

Existing visual encoders exhibit complementary strengths (Ma et al., 2025; Gao et al., 2026; Tu et al., 2025; Tan et al., 2025; Girdhar et al., 2023; Chen et al., 2024a). Vision-language encoders such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) learn strong global semantic representations through large-scale image-text alignment, but often lose fine-grained details and local structures (Wang et al., 2025c; Liu et al., 2025a; Ke et al., 2025a; Chen et al., 2026). In contrast, pixel-level self-supervised encoders such as DINOv2 and DINOv3 (Oquab et al., 2023; Siméoni et al., 2025), demonstrate superior capability in modeling local structures and texture-level information. However, these encoders inherently lack mechanisms to organize information across multiple semantic levels (Wang et al., 2025d; Liu et al., 2025d; Geetha & Hema, 2025), leading to a gap between pixel-level perception and high-level reasoning (Su et al., 2025; Xu et al., 2026; Liu et al., 2025b). Existing attempts to combine these paradigms either incur significant computational overhead or fail to establish a unified representation across granularity levels (Boné et al., 2026; Luo et al., 2026; Zhang et al., 2025a). Recently, Granulon (Mao et al., 2026) first revealed that DINOv3 exhibits an emerging capability to support semantic abstraction across multiple granularities when guided appropriately, by introducing a text-conditioned granularity controller. However, this design relies on an external LLM to determine granularity. This raises a central question: ***How can fine-grained visual evidence be adaptively organized into high-level semantic representations for multi-granularity multimodal reasoning?***

To address this question, we propose to move beyond fixed granularity control and introduce a ***meta-memory perspective for visual representation learning***, as shown in Figure 1. Specifically, we propose `Meta-MemGran`, a novel architecture that integrates pixel-level perception with structured semantic memory through a Meta-MemBank. Instead of treating tokens as isolated visual units, we organize them into multiple memory banks that capture complementary aspects of visual cognition:

> 💡 ❶ ***Global Bank*** → ***scene-level semantics and overall context*** ❷ ***Entity Bank*** → ***object-centric representations and identity-aware features*** ❸ ***Layout Bank*** → ***spatial arrangements and geometric distributions*** ❹ ***Relation Bank*** → ***interactions and dependencies between entities***

By explicitly modeling these memory components, `Meta-MemGran` enables dynamic and fine-grained control over token aggregation from a meta-cognitive perspective, allowing the model to adaptively construct semantic representations beyond coarse predefined levels. This design eliminates the dependency on external language models for granularity control and establishes a more flexible and scalable pathway from pixel-level perception to semantic-level reasoning. We conduct extensive experiments across diverse benchmarks. Under identical settings, our method achieves significant improvements, boosting accuracy by $\sim\mathbf{40}\%\uparrow$, while reducing hallucination by $\sim\mathbf{30}\%\downarrow$, highlighting the effectiveness of structured memory for scalable and more reliable multimodal reasoning.

The contributions of this paper are summarized as follows:

① We propose `Meta-MemGran`, a novel architecture that enhances pixel-level MLLMs via meta-memory-driven semantic modeling, enabling more effective integration of fine-grained visual perception with high-level semantic reasoning for improved multi-granularity understanding.

② We introduce a structured Meta-MemBank comprising Global, Entity, Layout, and Relation Banks to organize complementary visual information across abstraction levels, enabling dynamic aggregation of dense visual features into task-relevant semantic representations, inspired by human metacognitive processes.

③ Extensive experiments demonstrate that `Meta-MemGran` consistently improves reasoning accuracy while substantially reducing hallucination, outperforming existing visual encoders under identical settings.

2 Related Work

Semantic-level Visual Encoder. Semantic-level visual encoders map images to language-aligned representations via CLIP-style contrastive pre-training (Zhao et al., 2025; Zhang et al., 2024; Chen et al., 2025b; Munasinghe et al., 2025). Recent efforts focus on improving representation quality beyond global alignment. Sigmoid-based objectives (SigLIP (Zhai et al., 2023)) enhance training scalability and alignment efficiency, while region-aware extensions (Alpha-CLIP (Sun et al., 2024a)) introduce finer-grained control over localized semantics. Frameworks such as BLIP(Li et al., 2022) and BLIP-2 (Li et al., 2023c) connect frozen vision encoders with LLMs via lightweight bridging modules, achieving efficient alignment. Building on this design, recent models including LLaVA (Li et al., 2023b), LLaVA-OneVision (Li et al., 2024a), Qwen-VL (Bai et al., 2025), and BLIP-3 (Xue et al., 2025) further refine visual token representations and supporting complex inputs. *However, semantic abstraction can discard fine-grained **visual details** and **spatiotemporal cues**, limiting precise grounding and complex **visual reasoning**.*

Pixel-level Visual Encoder. Early pixel-level representation learning is largely driven by masked image modeling paradigms such as BEiT (Bao et al., 2021) and iBOT (Zhou et al., 2021), which learn fine-grained visual structures by reconstructing or predicting masked tokens. Building on this line, DINO (Caron et al., 2021) reveal that pixel-aligned features can naturally encode semantic segmentation cues without explicit supervision, while large-scale extensions such as DINOv2 (Oquab et al., 2023) and DINOv3 (Siméoni et al., 2025) further improve robustness and scalability through curated data, model scaling, and stabilization strategies. Recent efforts explore structured pixel-level representations, such as CroBo (Lee et al., 2026), which enforces what-is-where compositional encoding via global-to-local reconstruction, and Pixel-SAIL (Zhang et al., 2025b), which unifies pixel-level understanding within a single transformer through early fusion and distillation. PixelPrune (Wang et al., 2026) highlight the redundancy in pixel tokens. *However, dense pixel-level representations remain computationally redundant and often **lack** the **high-level semantics** required for effective multimodal reasoning.*

Visual Encoder in MLLMs. MLLMs typically use semantic-level visual encoders that map images to high-level tokens for LLM alignment (Yin et al., 2024; Ren et al., 2024; He et al., 2024; Huang et al., 2025; Wang et al., 2025b). Representative frameworks such as BLIP-2 (Li et al., 2023c), Flamingo (Alayrac et al., 2022), and LLaVA (Li et al., 2023b) employ pretrained vision encoders and lightweight adapters or query transformers to bridge the modality gap. Subsequent models, including InstructBLIP (Dai et al., 2023), Kosmos-2 (Peng et al., 2023), and PaLM-E (Driess et al., 2023), further enhance this paradigm through instruction tuning and grounded multimodal representations. More recently, InternVL (Chen et al., 2024b) and Qwen-VL series (Wang et al., 2024; Bai et al., 2025) have extend vision encoders and strengthen cross-modal alignment. Granulon (Mao et al., 2026) introduces granularity-aware control to bridge fine-grained perception and high-level reasoning.

💡 ***We introduce `Meta-MemGran` that unifies pixel-level perception and multi-granularity semantic abstraction, explicitly modeling global context, entities, relations, and spatial structure without external granularity control.***

3 Methods

Overview. Given an image-question pair (I, T), our goal is to transform a static memory bank into a meta-memory system that organizes heterogeneous perception priors for multimodal reasoning. Instead of directly retrieving a single memory item and appending it as an auxiliary token, we formulate memory as a structured set of priors that regulate how much to inspect, where to inspect, and how to compose local evidence. We organize memory into four complementary banks: $\boldsymbol{\mathcal{M}} = \{\boldsymbol{\mathcal{M}}_g, \boldsymbol{\mathcal{M}}_e, \boldsymbol{\mathcal{M}}_l, \boldsymbol{\mathcal{M}}_r\}$, corresponding to global, entity, layout, and relation priors, respectively. Their roles are disentangled:

$\boldsymbol{\mathcal{M}}_g$: scene-level difficulty and granularity bank, $\boldsymbol{\mathcal{M}}_e$: object/part-level saliency bank,
$\boldsymbol{\mathcal{M}}_l$: position-sensitive spatial bias bank, $\boldsymbol{\mathcal{M}}_r$: pairwise interaction and grouping bank.

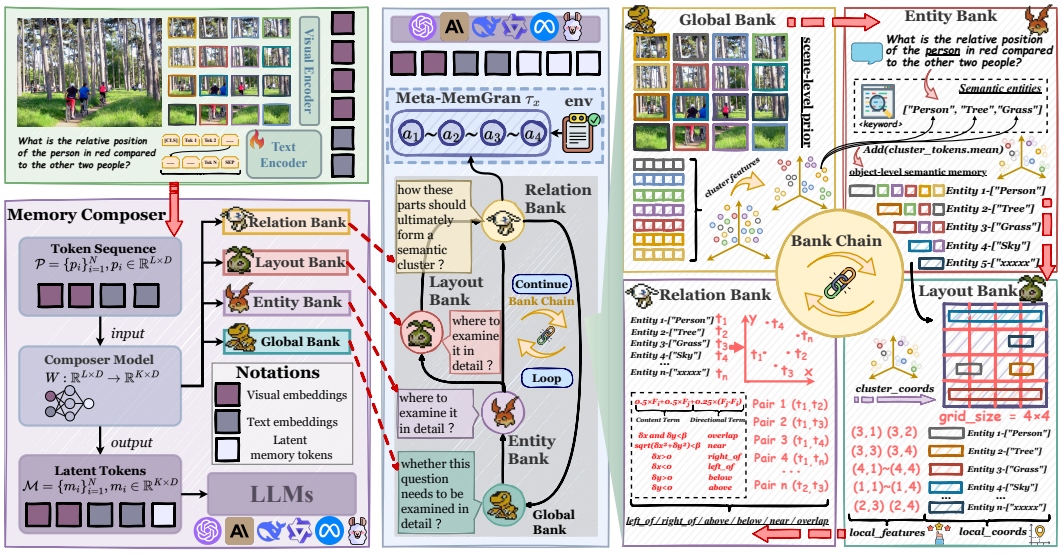


Figure 2: `Meta-MemGran` framework. A router selects and composes global, entity, layout, and relation priors into a meta-memory, which augments visual tokens for multimodal reasoning.

This leads to a hierarchical pipeline:

$$(\boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{C}}) \xrightarrow{\pi_\Theta} \{\boldsymbol{\mathcal{P}}_g, \boldsymbol{\mathcal{P}}_e, \boldsymbol{\mathcal{P}}_l, \boldsymbol{\mathcal{P}}_r\} \longrightarrow \boldsymbol{\mathcal{P}}, \\ \boldsymbol{\mathcal{P}} = \mathrm{Concat}\left(\boldsymbol{\mathcal{P}}_g, \boldsymbol{\mathcal{P}}_e, \boldsymbol{\mathcal{P}}_l, \boldsymbol{\mathcal{P}}_r\right). \tag{1}$$

where $\boldsymbol{\mathcal{F}} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N\} \in \mathbb{R}^{N \times D}$ denotes the visual token sequence extracted by the vision encoder, and $\boldsymbol{\mathcal{C}}$ denotes its clustered representations (defined later). Here, π_Θ denotes an explicit, non-parametric retrieval operator that retrieves prototypes from all four memory banks via similarity-based Top-K selection, rather than a learned routing network. It requires neither routing annotations nor an auxiliary routing loss, and each retrieval decision is directly inspectable through its similarity scores and selected prototypes. $\boldsymbol{\mathcal{P}}$ denotes the resulting meta-memory prior, we adopt a soft k-means clustering procedure to extract semantic anchors:

$$\alpha_{im} = \frac{\exp(\boldsymbol{f}_i^\top \boldsymbol{c}_m / \tau)}{\sum_{m'} \exp(\boldsymbol{f}_i^\top \boldsymbol{c}_{m'} / \tau)}, \quad \boldsymbol{c}_m = \frac{\sum_i \alpha_{im} \boldsymbol{f}_i}{\sum_i \alpha_{im}}, \tag{2}$$

where τ is a temperature parameter. Cluster centers are updated at every forward pass without maintaining external state, let $\boldsymbol{\mathcal{C}} = \mathrm{Cluster}(\boldsymbol{\mathcal{F}}; M)$, computed using soft assignments α_{im}, the vision encoder produces patch-level representations:

$$\boldsymbol{\mathcal{F}} = \mathcal{E}_v(I) \in \mathbb{R}^{N \times D}, \quad \boldsymbol{\mathcal{F}} = \{\boldsymbol{f}_i\}_{i=1}^N, \quad \boldsymbol{f}_i \in \mathbb{R}^D, \\ \boldsymbol{\mathcal{C}} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_M\} \in \mathbb{R}^{M \times D}, \quad \boldsymbol{\Pi} = \{\boldsymbol{p}_m\}_{m=1}^M \subset [0,1]^2, \tag{3}$$

we further derive a compact set of local semantic anchors by online clustering, and $\boldsymbol{c}_m$ is the m-th cluster token and $\boldsymbol{p}_m = (p_x^m, p_y^m)$ is its normalized coordinate, while $\boldsymbol{\Pi}$ provides geometric support for spatial and relational routing.

> ♣ ***Key insight:*** ***Memory is not merely a passive retrieval module, but an active system that dynamically organizes, composes, and controls perceptual representations across different levels of abstraction.***

3.1 META-MEMBANK

Bank Construction Given an input image I, a frozen visual encoder produces N patch-level features $\boldsymbol{\mathcal{F}} = \{\boldsymbol{f}_i\}_{i=1}^N$, $\boldsymbol{f}_i \in \mathbb{R}^D$. We run K-means over $\boldsymbol{\mathcal{F}}$ to obtain M local semantic anchors $\boldsymbol{\mathcal{C}} = \{\boldsymbol{c}_m\}_{m=1}^M$ and their normalized 2D centroids $\{\boldsymbol{p}_m\}_{m=1}^M \subset [0,1]^2$, reconstructed from the patch grid. Pairwise spatial relations are drawn from a fixed geometry-driven set $\mathcal{T}_{\text{rel}} = \{\texttt{left_of}, \texttt{right_of}, \texttt{above}, \texttt{below}, \texttt{overlap}, \texttt{near}\}$.

Per-image features. From each image we derive a global query $\boldsymbol{g}$, an entity-conditioned query $\boldsymbol{e}$, per-anchor layout features $\boldsymbol{l}_m$, and pair-wise relation features $\boldsymbol{r}_{mn}$:

$$\begin{gathered}\boldsymbol{g}=\tfrac{1}{N}\sum_{i=1}^{N}\boldsymbol{f}_i,\quad \boldsymbol{e}=\tfrac{1}{M}\sum_{m=1}^{M}\boldsymbol{c}_m,\quad \boldsymbol{l}_m=\boldsymbol{c}_m \text{ routed by } \pi(\boldsymbol{p}_m),\\ \boldsymbol{r}_{mn}=\operatorname{norm}\big(\tfrac{1}{2}\boldsymbol{c}_m+\tfrac{1}{2}\boldsymbol{c}_n+\tfrac{1}{4}(\boldsymbol{c}_n-\boldsymbol{c}_m)\big)\,,\end{gathered} \tag{4}$$

where $\pi(\boldsymbol{p})=\lfloor Gp_x\rfloor+G\lfloor Gp_y\rfloor\in\{0,\dots,G^2{-}1\}$ quantizes the centroid onto a $G{\times}G$ layout grid, and the relation label $t_{mn}\in\mathcal{T}_{\text{rel}}$ is assigned from $(\boldsymbol{p}_m,\boldsymbol{p}_n)$ by thresholding $(\Delta x,\Delta y)$ and $\|\boldsymbol{p}_n-\boldsymbol{p}_m\|$. The entity query $\boldsymbol{e}$ is further associated with the set of entity names $\mathcal{E}(I)$ parsed from the paired caption, giving tuples $(\boldsymbol{e},n)$ with $n\in\mathcal{E}(I)$.

Dataset-level prototype induction. We aggregate the per-image features over the whole training corpus and run K-means within each group, with a min-count filter that discards sparsely populated entities/cells/relation types. This yields four memory banks:

$$\boldsymbol{\mathcal{M}}_g=\{\boldsymbol{m}_k^g\}_{k=1}^{K_g^{\text{b}}},\boldsymbol{\mathcal{M}}_e=\bigcup_{n\in\mathcal{N}}\{\boldsymbol{m}_k^{e,n}\}_{k=1}^{K_e^{\text{b}}(n)},\boldsymbol{\mathcal{M}}_l=\bigcup_{c=0}^{G^2-1}\{\boldsymbol{m}_k^{l,c}\}_{k=1}^{K_l^{\text{b}}(c)},\boldsymbol{\mathcal{M}}_r=\bigcup_{t\in\mathcal{T}_{\text{rel}}}\{\boldsymbol{m}_k^{r,t}\}_{k=1}^{K_r^{\text{b}}(t)}, \tag{5}$$

where $\mathcal{N}$ is the set of retained entity names. All prototypes live in $\mathbb{R}^D$ and are ℓ_2-normalized. We cap each group at $K_g^{\text{b}}{=}32$, $K_e^{\text{b}}{\le}4$ per entity, $K_l^{\text{b}}{\le}4$ per grid cell, and $K_r^{\text{b}}{\le}8$ per relation type.

Meta-Memory Composition Given routed memory types, the Meta-Memory Composer aims to transform heterogeneous memory banks into a unified perception prior that modulates visual token processing. Unlike conventional retrieval that selects a fixed memory item, we formulate composition as a variable inference problem over structured priors. $\mathcal{Z}_e=\sum_k\alpha_k^e\,\boldsymbol{m}_k^e$, $\alpha_k^e=\operatorname{softmax}(\operatorname{sim}(\boldsymbol{c},\boldsymbol{m}_k^e))$. We compute alignment scores using a softmax over similarity functions. Taking the entity branch as an example, we optimize a dual-alignment objective that jointly considers local evidence and memory compatibility and a compositional operator:

$$\underbrace{\mathbb{E}_{\boldsymbol{\mathcal{C}}_m\in\boldsymbol{\mathcal{C}}}\log\operatorname{softmax}_m\big(\operatorname{sim}(\boldsymbol{\mathcal{C}}_m,\boldsymbol{\mathcal{Z}}_e)\big)}_{\textbf{local semantic support}}+\underbrace{\lambda_e\mathbb{E}_{\boldsymbol{\mathcal{M}}_k^e\in\mathcal{M}_e}\log\operatorname{softmax}_k\big(\operatorname{sim}(\boldsymbol{\mathcal{M}}_k^e,\boldsymbol{\mathcal{Z}}_e)\big)}_{\textbf{local semantic support}}, \tag{6}$$

where $\boldsymbol{\mathcal{C}}_m$ is the m-th local semantic anchor from $\boldsymbol{\mathcal{C}}$, and $\boldsymbol{\mathcal{M}}_k^e$ is the k-th entity memory slot from the memory set $\mathcal{M}_e$. We introduce a variable $\boldsymbol{\mathcal{Z}}=\{\boldsymbol{\mathcal{Z}}_g,\boldsymbol{\mathcal{Z}}_e,\boldsymbol{\mathcal{Z}}_l,\boldsymbol{\mathcal{Z}}_r\}$ representing the composed perception prior across memory types. Each component $\boldsymbol{\mathcal{Z}}_\bullet$ serves as a task-adaptive abstraction bridging online features and offline memory prototypes. This objective encourages $\boldsymbol{\mathcal{Z}}$ to capture task-relevant semantics while remaining well-regularized.

$$\hat{\boldsymbol{\mathcal{Z}}}=\Psi(\boldsymbol{\mathcal{Z}})\,,\quad \Psi(\boldsymbol{\mathcal{Z}})=\operatorname{MHCA}(\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{Z}},\boldsymbol{\mathcal{Z}})\,,\quad \boldsymbol{\mathcal{P}}=\Phi(\hat{\boldsymbol{\mathcal{Z}}})=\sigma(W_p\hat{\boldsymbol{\mathcal{Z}}})\odot\boldsymbol{\mathcal{F}}, \tag{7}$$

here the fused representation $\hat{\boldsymbol{\mathcal{Z}}}$ is then transformed into a perception-aware feature $\boldsymbol{\mathcal{P}}=\Phi(\hat{\boldsymbol{\mathcal{Z}}})$ through a gating function, where W_p is a learnable projection matrix, $\sigma(\cdot)$ is the sigmoid activation, and $\odot$ denotes element-wise modulation over the base feature $\boldsymbol{\mathcal{F}}$. enables coordinated interaction among heterogeneous memory components.

3.2 Meta-Memory Training Strategy

To adapt the proposed meta-memory MLLM to downstream multimodal reasoning tasks, we adopt a staged training strategy that separates visual-language alignment from memory-guided reasoning adaptation. The model consists of a vision tower $\mathcal{E}_v$, a multi-modal projector $\mathcal{G}_p$, a router π_Θ, a composer Φ_ϕ, and an autoregressive language model $\mathcal{D}_\psi$. Given an image-question-answer triplet (I,T,Y), the vision tower first extracts visual tokens and the final training objective:

$$\begin{gathered}\Delta\psi=\big\{\Delta W_q^\ell,\Delta W_k^\ell,\Delta W_v^\ell,\Delta W_o^\ell\big\}_{\ell\in\mathcal{S}}\,,\quad W_\star^\ell\leftarrow W_\star^\ell+\frac{\alpha}{r}B_\star^\ell A_\star^\ell,\quad \star\in\{q,k,v,o\}.\\ \boldsymbol{r}=\pi_\Theta\left(\cdot\mid\boldsymbol{\mathcal{F}},\boldsymbol{t}\right),\quad \boldsymbol{\mathcal{P}}=\Phi_\phi\left(\boldsymbol{\mathcal{C}},\{\boldsymbol{\mathcal{M}}_g,\boldsymbol{\mathcal{M}}_e,\boldsymbol{\mathcal{M}}_l,\boldsymbol{\mathcal{M}}_r\},\boldsymbol{r},\boldsymbol{t}\right),\end{gathered} \tag{8}$$

where $\mathcal{S}$ denotes the set of transformer layers where LoRA adapters are applied, and $\Delta\psi$ represents the set of trainable low-rank update parameters of the language model. $W_\star^\ell$ refers to the projection matrices in the ℓ-th attention layer for query, key, value, and output, where $\star\in\{q,k,v,o\}$. $A_\star^\ell\in$

$\mathbb{R}^{r\times d}$ and $B^{\ell}_{\star} \in \mathbb{R}^{d\times r}$ are the low-rank decomposition matrices of LoRA, where r is the rank and α is a scaling factor. **Stage I: projector alignment.** $\theta^{(1)}_{\text{train}} = \{\theta_p\}, \quad \theta_v, \theta_\psi, \theta_\Theta, \theta_\phi$ are frozen in the first stage, we freeze the vision tower and the language model, and only optimize the multi-modal projector. This stage learns a stable mapping from visual tokens to the LLM embedding space without perturbing the pre-trained vision-language backbone. **Stage II: memory-guided LoRA adaptation.** $\theta^{(2)}_{\text{train}} = \{\Delta\psi, \theta_\Theta, \theta_\phi\}, \quad \theta_p, \theta_v$ are frozen After projector alignment, we freeze the multi-modal projector and adapt the language-side reasoning layers with LoRA. Following the implementation, LoRA is applied to the attention projections of selected upper LLM layers.

4 Experiments

Experiment Setup. ❶ **Datasets.** We assess `Meta-MemGran` on **five representative benchmarks** to examine the influence of different visual encoders on MLLM performance across diverse task settings: **Visual Question Answering (VQA)**: SEED-Bench (Li et al., 2023a) and A-OKVQA (Schwenk et al., 2022); **Image Captioning**: CC12M_and_Imagenet21K_Recap (Changpinyo et al., 2021); **Reasoning**: FLUX-Reason (Fang et al., 2025); **Medical Applications**: SurgVLM (Zeng et al., 2025). ❷ **OOD Evaluation Protocol.** Notably, both memory-bank construction and model training use only FLUX-Reason. *All other benchmarks are unseen during training and serve exclusively for **zero-shot OOD evaluation***. ❸ **GPT-based Multimodal Evaluation** To evaluate multimodal capabilities, we adopt multiple complementary metrics. Following recent studies on hallucination evaluation in MLLMs (Sun et al., 2024b; Liu et al., 2024; Jiang et al., 2026; Zhang et al., 2026a; He et al., 2026; Mao et al., 2026; Li et al., 2025; 2024b), GPT-4o is utilized as an automatic evaluator to measure semantic accuracy, hallucination rate, and granularity. Following Mao et al. (Mao et al., 2026), we adopt a 0-80 hallucination severity score, where 0 indicates no visually unsupported content and 80 represents the highest hallucination severity. The reported score is averaged over all valid evaluation samples. Hallucination score interpretations are provided in *Appendix Table* 8, with the corresponding evaluation prompt shown in *Prompt* F. ❹ **Human-Judge Validation.** We further validate the GPT-based hallucination metric through human evaluation on 100 randomly sampled cases. As listed in *Appendix Table* 9, the human and GPT-based judgments show strong agreement, supporting the reliability of the metric. ❺ **Evaluation Metrics.** For VQA tasks (FLUX-Reasoning), we report ***BERTScore-P/R/F1***, ***Token-P/R/F1***, ***BLEU-1***, ***Distinct-1/2***, and ***Length Ratio*** to evaluate semantic alignment, lexical agreement, diversity, and length consistency, respectively. ❻ **Details of Implementation** are included in *Appendix* C.2.

4.1 Main Results

Table 1: 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(%)↑
CLIP	Qwen2.5	50.91$_{\uparrow 8.03}$	24.40$_{\uparrow 10.72}$	21.79$_{\uparrow 42.49}$	10.30$_{\uparrow 31.37}$	84.35$_{\uparrow 0.58}$	21.54$_{\uparrow 26.38}$	87.79$_{\uparrow 0.99}$	29.36$_{\uparrow 30.51}$
SigLIP	Qwen2.5	46.72$_{\uparrow 12.22}$	23.10$_{\uparrow 12.02}$	21.89$_{\uparrow 42.39}$	5.10$_{\uparrow 36.57}$	83.92$_{\uparrow 1.01}$	13.59$_{\uparrow 34.33}$	86.51$_{\uparrow 2.27}$	23.59$_{\uparrow 36.28}$
DINOv2	Qwen2.5	41.40$_{\uparrow 17.54}$	17.70$_{\uparrow 17.42}$	16.67$_{\uparrow 47.61}$	14.10$_{\uparrow 27.57}$	83.40$_{\uparrow 1.53}$	14.36$_{\uparrow 33.56}$	88.15$_{\uparrow 0.63}$	36.67$_{\uparrow 23.20}$
DINOv3	Qwen2.5	55.74$_{\uparrow 3.20}$	15.40$_{\uparrow 19.72}$	47.43$_{\uparrow 16.85}$	25.60$_{\uparrow 16.07}$	83.90$_{\uparrow 1.03}$	23.97$_{\uparrow 23.95}$	88.26$_{\uparrow 0.52}$	45.31$_{\uparrow 14.56}$
Granulon	Qwen2.5	58.80$_{\uparrow 0.14}$	**35.80**$_{\downarrow 0.68}$	57.13$_{\uparrow 7.15}$	35.90$_{\uparrow 5.77}$	84.31$_{\uparrow 0.62}$	31.28$_{\uparrow 16.64}$	**88.94**$_{\downarrow 0.16}$	49.31$_{\uparrow 10.56}$
Ours	Qwen2.5	**58.94**	35.12	**64.28**	**41.67**	**84.93**	**47.92**	88.78	**59.87**
CLIP	Llama3.2	42.53$_{\uparrow 10.25}$	21.10$_{\uparrow 35.51}$	25.64$_{\uparrow 39.98}$	25.60$_{\uparrow 34.02}$	82.91$_{\uparrow 1.98}$	7.44$_{\uparrow 15.26}$	87.01$_{\uparrow 1.80}$	28.97$_{\uparrow 32.95}$
SigLIP	Llama3.2	40.62$_{\uparrow 12.16}$	17.90$_{\uparrow 38.71}$	17.94$_{\uparrow 47.68}$	6.40$_{\uparrow 53.22}$	84.89$_{\uparrow 0.00}$	15.36$_{\uparrow 7.34}$	88.46$_{\uparrow 0.35}$	44.87$_{\uparrow 17.05}$
DINOv2	Llama3.2	40.26$_{\uparrow 12.52}$	23.10$_{\uparrow 33.51}$	35.90$_{\uparrow 29.72}$	17.90$_{\uparrow 41.72}$	82.59$_{\uparrow 2.30}$	5.54$_{\uparrow 17.16}$	86.55$_{\uparrow 2.26}$	19.49$_{\uparrow 42.43}$
DINOv3	Llama3.2	47.23$_{\uparrow 5.55}$	30.30$_{\uparrow 26.31}$	32.05$_{\uparrow 33.57}$	7.70$_{\uparrow 51.92}$	82.98$_{\uparrow 1.91}$	8.12$_{\uparrow 14.58}$	87.26$_{\uparrow 1.55}$	34.87$_{\uparrow 27.05}$
Granulon	Llama3.2	48.98$_{\uparrow 3.80}$	53.80$_{\uparrow 2.81}$	54.61$_{\uparrow 11.01}$	51.30$_{\uparrow 8.32}$	83.82$_{\uparrow 1.07}$	18.08$_{\uparrow 4.62}$	87.33$_{\uparrow 1.48}$	56.67$_{\uparrow 5.25}$
Ours	Llama3.2	**52.78**	**56.61**	**65.62**	**59.62**	**84.89**	**22.70**	**88.81**	**61.92**

Obs. ❶: Consistent Performance Gains across Benchmarks. As listed in Table 1, under Qwen2.5, the GPTscore of SEED-Bench increases to 35.12%, indicating stronger reasoning ability. On A-OKVQA, the GPTscore rises significantly to 41.67%, showing better answer quality. For reasoning tasks, the GPTscore reaches 59.87%, exceeding DINOv3 by a large margin. Under Llama3.2, our method achieves the highest GPTscores across all tasks, including 61.92% on reasoning, Compared with Granulon, our method shows better semantic understanding and response quality. Overall, the results indicate that our model improves reasoning quality rather than only matching surface level features, validating the effectiveness of structured meta memory. **Obs. ❷: Substantial Improvements in Multi-step Reasoning.** Our method shows clear advantages on reasoning tasks, as reported in Table 1. Under Qwen2.5, the reasoning GPT score reaches 59.87%, outperforming DINOv3 at 45.31% and CLIP at 29.36%. With Llama3.2, it further improves to 61.92%, exceeding DINOv3 and CLIP by large margins. The gains are more significant on GPT

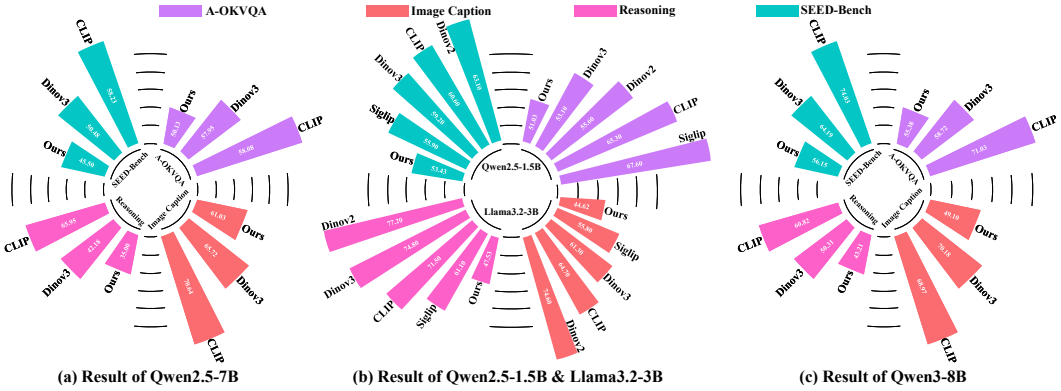
metrics than on BERT score, indicating improved logical coherence and multi step reasoning. This demonstrates that `Meta-MemGran` enables stronger structured and fine grained reasoning.

Table 2: OOD Generalization on SurgVLM.

Encoders	Phase Recognition		Instrument Recognition	
	BERTscore(%)↑	Recall(%)↑	BERTscore(%)↑	Recall(%)↑
CLIP	91.64$_{\uparrow 5.24}$	46.15$_{\uparrow 32.89}$	94.44$_{\uparrow 3.29}$	46.92$_{\uparrow 28.67}$
SigLIP	91.98$_{\uparrow 4.90}$	48.72$_{\uparrow 30.32}$	95.71$_{\uparrow 2.02}$	**78.20**$_{\downarrow 2.61}$
DINOv2	91.67$_{\uparrow 5.21}$	33.33$_{\uparrow 45.71}$	95.93$_{\uparrow 1.80}$	55.21$_{\uparrow 20.38}$
DINOv3	94.71$_{\uparrow 2.17}$	64.10$_{\uparrow 14.94}$	97.41$_{\uparrow 0.32}$	68.89$_{\uparrow 6.70}$
Ours	**96.88**	**79.04**	**97.73**	75.59

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 2, our method achieves the best overall performance, with 96.88% / 97.73% BERTScore and a leading 79.04% Recall in phase recognition. Compared to `CLIP`, `DINOv2/v3`, our method yields consistent gains, particularly in Recall, significantly outperforming `CLIP` and `DINOv2`,

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, our method consistently achieves the lowest hallucination scores across models and datasets. In Figure 3 (a), with Qwen2.5-7B, our method reduces the hallucination score on SEED-Bench from 58.23 with CLIP and 50.48 with DINOv3 to 45.50, demonstrating clear advantages in complex reasoning scenarios. On FLUX-Reasoning, it further reduces the hallucination score to 47.53, compared with 71.50, 77.20, 74.80, and 61.10 for the respective baselines, showing strong generalization across model configurations. Overall, these results demonstrate that our method consistently mitigates hallucinations across different model scales and evaluation benchmarks.

4.3 Ablation Studies

Obs. ❶: Scaling Up: Ablation on 7B/8B Models with Consistent Multi-Task Performance Gains As listed in Table 3, our method remains consistently effective on larger models. For Qwen2.5-7B,

Table 3: Performance comparison on Qwen2.5-7B and Qwen3-8B models. Best results are **bold**.

Encoders	Models	SEED-Bench		A-OKVQA		Image Caption		Reasoning	
		Recall(%)↑	GPTscore(%)↑	Recall(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑
CLIP	Qwen2.5-7B	55.33$_{\uparrow 14.24}$	31.45$_{\uparrow 22.40}$	**48.89**$_{\downarrow 0.68}$	38.21$_{\uparrow 10.89}$	79.85$_{\uparrow 4.48}$	20.64$_{\uparrow 12.23}$	88.17$_{\uparrow 1.29}$	31.67$_{\uparrow 40.64}$
DINOv3	Qwen2.5-7B	60.77$_{\uparrow 8.80}$	39.57$_{\uparrow 14.28}$	47.20$_{\uparrow 1.01}$	42.88$_{\uparrow 6.22}$	79.62$_{\uparrow 4.71}$	23.17$_{\uparrow 9.70}$	88.67$_{\uparrow 0.79}$	63.29$_{\uparrow 9.02}$
Ours	Qwen2.5-7B	**69.57**	**53.85**	48.21	**49.10**	**84.33**	**32.87**	**89.46**	**72.31**
CLIP	Qwen3-8B	92.16$_{\uparrow 4.84}$	11.94$_{\uparrow 30.37}$	49.87$_{\uparrow 0.00}$	16.67$_{\uparrow 29.61}$	79.31$_{\uparrow 5.60}$	**24.74**$_{\downarrow 0.65}$	88.77$_{\uparrow 0.73}$	39.74$_{\uparrow 27.62}$
DINOv3	Qwen3-8B	94.24$_{\uparrow 2.76}$	28.23$_{\uparrow 14.08}$	47.67$_{\uparrow 2.20}$	38.97$_{\uparrow 7.31}$	79.48$_{\uparrow 5.43}$	22.49$_{\uparrow 1.60}$	88.92$_{\uparrow 0.58}$	58.72$_{\uparrow 8.64}$
Ours	Qwen3-8B	**97.00**	**42.31**	**49.87**	**46.28**	**84.91**	24.09	**89.50**	**67.36**

SEED-Bench Recall reaches 69.57%, exceeding DINOv3 by nearly 9 points, while the GPTscore increases from 39.57% to 53.85%. On A-OKVQA, the GPTscore improves to 49.10%. For reasoning, the GPTscore reaches 72.31%, showing a large margin. For Qwen3-8B, SEED-Bench Recall reaches 97.00% with a GPTscore of 42.31%. On A-OKVQA, both metrics achieve the best results. The reasoning GPTscore reaches 67.36%, clearly outperforming baselines. Overall, improvements are larger on GPT metrics than Recall, indicating better semantic coherence. The consistent gains

confirm strong scalability to larger models. ♣ **Evaluation with Diverse Metrics.** To complement the GPT-based evaluation in Table 3, we further report diverse metrics on FLUX-Reason in *Appendix Table 7*. Across both Qwen2.5-7B and Qwen3-8B, our method consistently achieves the best or competitive results on BERTScore, token-level F1, BLEU-1, and Distinct metrics. In particular, Token-F1 improves to 52.01% and 52.65%, while BLEU-1 reaches 46.01% and 46.44%, respectively. These consistent gains across semantic similarity, lexical overlap, and generation diversity confirm that the improvements are not specific to a single evaluation metric.

Table 4: Ablation study of `Meta-MemGran` components.

Encoders	Meta-MemGran compositions				SEED-Bench		Reasoning	
	Global	Entity	Layout	Relation	Recall(%)↑	GPTscore(%)↑	BERTscore(%)↑	GPTscore(%)↑
Qwen2.5	✓				57.93$_{\uparrow 1.01}$	29.32$_{\uparrow 5.80}$	88.70$_{\uparrow 0.08}$	52.23$_{\uparrow 7.64}$
		✓			57.60$_{\uparrow 1.34}$	24.74$_{\uparrow 10.38}$	88.64$_{\uparrow 0.14}$	49.87$_{\uparrow 10.00}$
			✓		58.88$_{\uparrow 0.06}$	30.16$_{\uparrow 4.96}$	88.68$_{\uparrow 0.10}$	53.74$_{\uparrow 6.13}$
				✓	56.59$_{\uparrow 2.35}$	19.61$_{\uparrow 15.51}$	88.60$_{\uparrow 0.18}$	47.02$_{\uparrow 12.85}$
		✓	✓	✓	57.84$_{\uparrow 1.10}$	25.79$_{\uparrow 9.33}$	88.65$_{\uparrow 0.13}$	52.93$_{\uparrow 6.94}$
	✓		✓	✓	58.11$_{\uparrow 0.83}$	28.02$_{\uparrow 7.10}$	88.70$_{\uparrow 0.08}$	55.23$_{\uparrow 4.64}$
	✓	✓		✓	57.04$_{\uparrow 1.90}$	23.32$_{\uparrow 11.80}$	88.64$_{\uparrow 0.14}$	50.82$_{\uparrow 9.05}$
	✓	✓	✓		58.51$_{\uparrow 0.43}$	29.01$_{\uparrow 6.11}$	88.74$_{\uparrow 0.04}$	57.01$_{\uparrow 2.86}$
	✓	✓	✓	✓	**58.94**	**35.12**	**88.78**	**59.87**

Obs. ❷: Single-component and leave-one-out (LOO) ablation studies of `Meta-MemGran`. As listed in Table 4, the LOO results reveal complementary contributions of the four memory banks. Under Qwen2.5, removing the Entity bank causes the largest drop in reasoning GPT score (4.64%), while excluding the Layout bank leads to the largest degradation in SEED-Bench GPT score (11.80%). Removing the Layout or Relation bank also consistently reduces performance across the evaluated metrics. Combining all four banks consistently achieves the best performance, reaching 35.12%/59.87% GPT scores with Qwen2.5 and 56.61%/61.92% with Llama3.2 on SEED-Bench/reasoning, respectively. Overall, the LOO results demonstrate that each bank provides complementary information and that their joint integration is essential for robust structured reasoning. **Obs. ❸: Hierarchy-Aware Visual Understanding** We further examine whether our method captures structured visual information beyond improving final-answer accuracy. As shown in *Appendix Table 11*, our model consistently recovers fine-grained information at three levels: *object attributes, inter-object relationships, and global spatial layouts*. These results provide direct qualitative evidence that our method enhances structured visual understanding across multiple levels of granularity. Detailed case-level analysis provided in *Appendix D*.

Obs. ❹: Computational Efficiency Analysis in Terms of Performance and Resource Utilization.

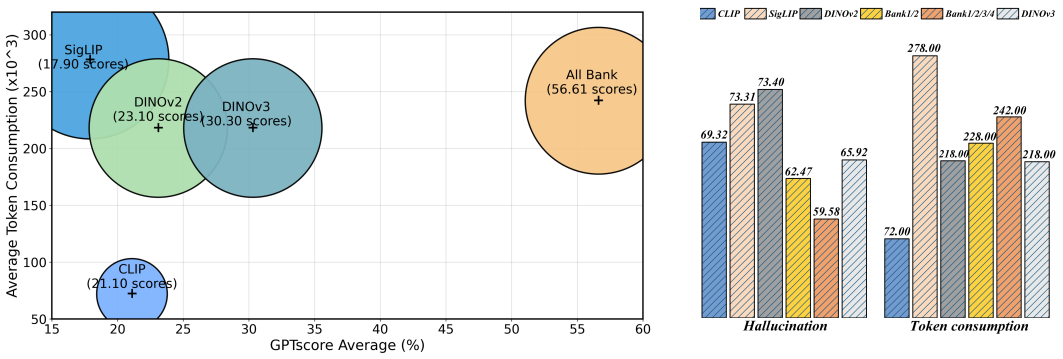


Figure 4: Visualization of Performance Metrics and Input Token Consumption Across Different Setups. Token consumption is measured using Llama 3.2-3B as the language model.

Table 5: Statistics of observation memory (Flux-Reasoning).

Config.	Scale	Bank Dims.	Obs. Memory	Avg. Time /Sample
Qwen2.5 + DINOv3 +`Meta-MemGran`	1.5B	1536	4524 MiB / 4.42 GiB	4.88 s
Llama3.2 + DINOv3 +`Meta-MemGran`	3B	3072	8070 MiB / 7.88 GiB	4.81 s
Qwen2.5 + DINOv3 +`Meta-MemGran`	7B	3584	15916 MiB / 15.54 GiB	5.16 s
Qwen3 + DINOv3 +`Meta-MemGran`	8B	4096	16932 MiB / 16.54 GiB	5.30 s

We further evaluate the computational efficiency of `Meta-MemGran` from the perspective of performance, hallucination, and token consumption, as shown in Figure 4. The single-representation baselines, including CLIP, SigLIP, DINOv2, and DINOv3, exhibit limited reasoning performance, with GPTscores ranging only from 17.90 to 30.30. In contrast, introducing the proposed Meta-MemBank leads to a clear and progressive improvement. When all memory components are integrated, including Global, Entity, Relation, and Layout Banks, `Meta-MemGran` further improves the GPTscore to 56.61. Meanwhile, the hallucination rate is also reduced, indicating that

structured memory provides more reliable visual evidence rather than simply increasing the number of tokens. Although the full memory bank introduces additional visual tokens, the performance gain is disproportionately larger than the increase in token usage, demonstrating a more favorable trade-off between accuracy and computational cost. ♣ **Computational Efficiency Analysis.** As listed in Table 5, `Meta-MemGran` scales efficiently across different LLM sizes. From 1.5B to 8B, observation memory increases from 4.42 to 16.54 GiB, while inference time remains stable at around 4.8–5.3 s/sample, demonstrating favorable scalability and computational efficiency. Despite the memory-intensive nature of billion-scale LLM inference (Kwon et al., 2023; Dao et al., 2022), `Meta-MemGran` maintains a moderate observation-memory footprint and nearly stable inference latency across different backbone scales. **Obs. ❺: GPT-based Hallucination Metric Reliability.** We further validate the reliability of the GPT-based hallucination metric through human evaluation. As listed in *Appendix Table 9*, *human and GPT-based evaluations achieve highly consistent average scores* (34.30 vs. 34.10), with an absolute difference of only 0.20. This close agreement supports the reliability of our GPT-based hallucination evaluation, detailed analysis provided in *Appendix* C.5.

Table 6: `Meta-MemGran` 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-MemGran`		**50.12**	**22.47**	**20.83**	**9.05**	**84.29**	**20.71**	**86.55**	**29.92**
`CLIP`	Llama3.2	42.53	21.10	25.64	25.60	82.91	7.44	87.01	28.97
+`Meta-MemGran`		**42.08**	**20.36**	**25.19**	**22.64**	**82.77**	**8.22**	**87.41**	**27.88**
Pixel encoder verification									
`DINOv3`	Qwen2.5	55.74	15.40	47.43	25.60	83.90	23.97	88.26	45.31
+`Meta-MemGran`		**58.94**	**35.12**	**64.28**	**41.67**	**84.93**	**47.92**	**88.78**	**59.87**
`DINOv3`	Llama3.2	47.23	30.30	32.05	7.70	82.98	8.12	87.26	34.87
+`Meta-MemGran`		**52.78**	**56.61**	**65.62**	**59.62**	**84.89**	**22.70**	**88.81**	**61.92**

Obs. ❻: Generalization ability on Semantic and Pixel Encoders. Table 6 highlights a divergence between semantic and pixel encoders when integrating our method. For semantic encoders, `Meta-MemGran` brings no consistent gains and occasionally causes drops. For instance, with CLIP + Qwen2.5, SEED-Bench Recall decreases (50.91→50.12) and GPTscore drops (24.40→22.47); This suggests that semantic encoders already rely on well-aligned global representations, leaving limited room for additional structured aggregation. In contrast, pixel encoders benefit substantially. With DINOv3 + Qwen2.5, SEED-Bench GPTscore improves dramatically (15.40→35.12) and reasoning GPTscore rises (45.31→59.87). These consistent improvements indicate that `Meta-MemGran` effectively enhances semantic organization. Overall, the results suggest that our hypothesis: *`Meta-MemGran` is more effective for encoders that preserve rich local and structural information, while offering limited gains for encoders like CLIP, whose representations are intrinsically optimized for global semantic alignment and exhibit inherently reduced spatial granularity, making explicit structural modeling less beneficial.*

Obs. ❼: Qualitative Case Study. To better understand the strengths and limitations of our method, we present a qualitative case study in *Appendix* H (Figure 15). Compared with the baseline (`CLIP`), our model produces a more coherent and fine-grained description. While `CLIP` overemphasizes culturally salient but incorrect attributes (e.g., "traditional Chinese wedding attire" and "Grand Canyon"), our method better aligns with the visual content, correctly capturing key elements such as the couple, lakeside setting, grassy foreground, and mountainous background. Furthermore, our model exhibits improved compositional reasoning, organizing the scene into foreground, midground, and background, and providing richer attribute-level details (e.g., clothing and interaction cues) consistent with the ground truth.

5 Conclusion & Limitation

This paper presents `Meta-MemGran`, a meta-memory driven framework that bridges pixel-level perception and semantic reasoning in MLLMs. By organizing visual representations into Global, Entity, Relation, and Layout memory banks, the model enables adaptive and structured composition of visual information. Experimental results show clear improvements in reasoning accuracy and reduced hallucination across multiple benchmarks , demonstrating the effectiveness of treating memory as an active perception mechanism. **Limitation:** While `Meta-MemGran` effectively bridges pixel-level perception and semantic reasoning, its current scope is primarily focused on static image understanding. Human perception encompasses even more dimensions. Future work could extend to dynamic video and explore incorporating additional specialized memory components, such as physics priors or temporal-action banks, to support more complex physical and causal reasoning.

LLM USAGE DISCLOSURE

During the preparation of this manuscript, GPT-5.6 was employed as an auxiliary tool to support language refinement and selected writing-related tasks. Its use primarily included improving grammar, sentence fluency, wording, and overall readability, as well as providing limited assistance with coding-related work. All scientific ideas, methodological choices, experimental designs, analyses, and conclusions were developed by the authors. GPT-5.6 was not used to generate independent scientific contributions or to make autonomous research decisions. Any text or code produced with LLM assistance was subsequently examined, validated, and, when necessary, modified by the authors to ensure correctness and consistency with the intended content. The authors remain fully responsible for the accuracy, interpretation, and presentation of all material included in this manuscript.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 3 describes the Meta-MemGran architecture, including memory-bank construction, prototype retrieval, memory composition, and the staged training strategy. Appendix B provides dataset preprocessing procedures, experimental settings, training hyperparameters, and evaluation metrics, while Appendix D reports the hardware and computational configuration. The GPT-4o evaluation prompt and scoring procedure are presented in Appendix F, with the hallucination scoring criteria and human validation described in Appendices F and Appendices C.5.

ETHICS STATEMENT

Meta-MemGran aims to improve visual grounding and reduce hallucinations in multimodal large language models. Our experiments use existing datasets and pretrained models, whose inherited biases and potentially sensitive content may affect model behavior. Although our approach reduces hallucination scores on the evaluated benchmarks, it does not eliminate unsupported statements or guarantee reliable performance across populations and domains. In particular, the SurgVLM experiments assess performance on surgical image understanding tasks and do not establish clinical safety or suitability for autonomous medical decision-making. Improved visual understanding may also be misused for privacy-invasive analysis or surveillance. Downstream applications should therefore respect privacy, applicable data and model usage restrictions, and incorporate safeguards appropriate to their intended use.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.

Aleix Boné, Alejandro Aguirre, David Álvarez, Pedro J Martinez-Ferrer, and Vicenç Beltran. A task-based data-flow methodology for programming heterogeneous systems with multiple accelerator apis. Future Generation Computer Systems, pp. 108383, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660, 2021.

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.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26540–26550, 2024a.

Shangjing Chen, Feng Xu, Xin Lyu, Dafa Wang, and Xin Li. Clip-driven feature disambiguation and cross-modal synergy for few-shot semantic segmentation. Expert Systems with Applications, 306: 130892, 2026.

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, 2025a.

Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. Mimo: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 24732–24741, 2025b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 24185–24198, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems, 36:49250–49267, 2023.

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.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023.

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.

Shengwei Fu, Mingyang Yu, Kaichen OuYang, Qingsong Fan, and Haisong Huang. Mllms-mr: Multi-modal recognition based on multi-modal large language models. Knowledge-Based Systems, pp. 114717, 2025.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pp. 4095–4103, 2026.

T Geetha and C Hema. Deep learning-based joint analysis of diabetic retinopathy and glaucoma in retinal fundus images. Scientific Reports, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15180–15190, 2023.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13504–13514, 2024.

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.

Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 3779–3787, 2025.

Xiaofeng Ji, Faming Gong, Nuanlai Wang, Yanpu Zhao, Yuhui Ma, and Zhuang Shi. Pixel-level semantic parsing in complex industrial scenarios using large vision-language models. Information Fusion, 116:102794, 2025.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. arXiv preprint arXiv:2505.17316, 2025a.

Xin Jiang, Hao Tang, Yonghua Pan, and Zechao Li. Rethinking vision transformer for large-scale fine-grained image retrieval. IEEE Transactions on Multimedia, 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.

Wenjun Ke, Yifan Zheng, Yining Li, Hengyuan Xu, Dong Nie, Peng Wang, and Yao He. Large language models in document intelligence: A comprehensive survey, recent advances, challenges, and future trends. ACM Transactions on Information Systems, 44(1):1–64, 2025a.

Xiao Ke, Yuhang Cai, Baitao Chen, Hao Liu, and Wenzhong Guo. Multi-granularity interaction and feature recombination network for fine-grained visual classification. Pattern Recognition, 166: 111632, 2025b.

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.

Seokmin Lee, Yunghee Lee, Byeonghyun Pak, and Byeongju Woo. Pixel-level scene understanding in one token: Visual states need what-is-where composition. arXiv preprint arXiv:2603.13904, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024a.

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, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36: 28541–28564, 2023b.

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, 2025.

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, 2024b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pp. 19730–19742. PMLR, 2023c.

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.

Hai Liu, Yu Song, Tingting Liu, Hao Zheng, Lin Chen, Zhaoli Zhang, and You-Fu Li. Homllm: exploiting semantic homology relationship for fine-grained bird image classification via large language models. IEEE transactions on neural networks and learning systems, 2025a.

Peng Liu, Haozhan Shen, Chunxin Fang, Zhicheng Sun, Jiajia Liao, and Tiancheng Zhao. Vlm-fo1: Bridging the gap between high-level reasoning and fine-grained perception in vlms. arXiv preprint arXiv:2509.25916, 2025b.

Xu Liu and Zhouhui Lian. Rsunivlm: A unified vision-language model for remote sensing via granularity-oriented moe. Pattern Recognition, pp. 113717, 2026.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. arXiv preprint arXiv:2509.18094, 2025c.

Yide Liu, Haijiang Sun, Xiaowen Zhang, Qiaoyuan Liu, Zhouchang Chen, and Chongzhuo Xiao. Ttrd3: Texture transfer residual denoising dual diffusion model for remote sensing image super-resolution. IEEE Transactions on Geoscience and Remote Sensing, 2025d.

Zhenghui Luo, Min Meng, and Jigang Wu. Dynamic patch selection and dual-granularity alignment for cross-modal retrieval. Neurocomputing, pp. 132999, 2026.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 4744–4754, 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.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. Spatiallm: Training large language models for structured indoor modeling. arXiv preprint arXiv:2506.07491, 2025.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 19036–19046, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193, 2023.

Yuqi Pang, Bowen Yang, Haoqin Tu, Yun Cao, and Zeyu Zhang. Language models can see better: Visual contrastive decoding for llm multimodal reasoning. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE, 2025.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824, 2023.

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.

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

Sara Sarto, Marcella Cornia, and Rita Cucchiara. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. arXiv preprint arXiv:2503.14604, 2025.

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.

Lingdong Shen, Xiaoshuang Huang, Fangxin Shang, Xudong Zhang, Yehui Yang, Bin Fan, and Shiming Xiang. From image to pixels: towards fine-grained medical vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. arXiv preprint arXiv:2508.10104, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13019–13029, 2024a.

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, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pp. 7184–7192, 2025.

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.

Weijie Tu, Weijian Deng, and Tom Gedeon. Toward a holistic evaluation of robustness in clip models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

Fengxiang Wang, Deying Yu, Liang Huang, Yalun Zhang, Yongbing Chen, and Zhiguo Wang. Fine-grained ship image classification and detection based on a vision transformer and multi-grain feature vector fpn model. Geo-Spatial Information Science, 28(3):923–944, 2025a.

Jingchao Wang, Zhijian Wu, Dingjiang Huang, Yefeng Zheng, and Hong Wang. Unlocking the potential of mllms in referring expression segmentation via a light-weight mask decoder. arXiv preprint arXiv:2508.04107, 2025b.

Nan Wang, Zhiwei Jin, Chen Chen, and Haonan Lu. Pixelprune: Pixel-level adaptive visual token reduction via predictive coding. arXiv preprint arXiv:2604.00886, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

Yaxiong Wang, Lianwei Wu, Lechao Cheng, Zhun Zhong, Yujiao Wu, and Meng Wang. Beyond general alignment: Fine-grained entity-centric image-text matching with multimodal attentive experts. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 792–802, 2025c.

Zenghui Wang, Songlin Du, Yaping Yan, Guobao Xiao, and Xiaobo Lu. Tex2sem: Learning from textures to semantics for robust semantic correspondence. IEEE Transactions on Circuits and Systems for Video Technology, 2025d.

Christopher KI Williams. Structured generative models for scene understanding. International Journal of Computer Vision, 133(5):2845–2867, 2025.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747, 2025.

Yu Xu, Yuxin Zhang, Juan Cao, Lin Gao, Chunyu Wang, Oliver Deussen, Tong-Yee Lee, and Fan Tang. Beyond pixels: Visual metaphor transfer via schema-driven agentic reasoning. arXiv preprint arXiv:2602.01335, 2026.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. Blip-3: A family of open large multimodal models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6124–6135, 2025.

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, 2026.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. National Science Review, 11(12):nwae403, 2024.

Ying Yu, Wei Wei, Cairong Zhao, Jin Qian, and Enhong Chen. Structural feature enhanced transformer for fine-grained image recognition. Pattern Recognition, 169:111955, 2026.

Zhipeng Yu, Qianqian Xu, Yangbangyan Jiang, Pinci Yang, and Qingming Huang. Mgvc: Mllm-guided video captioning for the intentvc challenge. In Proceedings of the 33rd ACM International Conference on Multimedia, pp. 13815–13821, 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.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence, 46(8):5625–5644, 2024.

Shenglin Zhang, Sibo Xia, Wenzhao Fan, Binpeng Shi, Xiao Xiong, Zhenyu Zhong, Minghua Ma, Yongqian Sun, and Dan Pei. Failure diagnosis in microservice systems: A comprehensive survey and analysis. ACM Transactions on Software Engineering and Methodology, 35(1):1–55, 2025a.

Tao Zhang, Xiangtai Li, Zilong Huang, Yanwei Li, Weixian Lei, Xueqing Deng, Shihao Chen, Shunping Ji, and Jiashi Feng. Pixel-sail: Single transformer for pixel-grounded understanding. arXiv preprint arXiv:2504.10465, 2025b.

Wei Zhang, Miaoxin Cai, Yaqian Ning, Tong Zhang, Yin Zhuang, Shijian Lu, He Chen, Jun Li, and Xuerui Mao. Earthgpt-x: A spatial mllm for multilevel multisource remote sensing imagery understanding with visual prompting. IEEE Transactions on Geoscience and Remote Sensing, 63: 1–21, 2025c.

Xu Zhang, Jiabin Fang, Zhuoming Ding, Jin Yuan, Xuan Liu, Qianjun Zhang, and Zhiyong Li. Cross-modal context-aware learning for visual prompt guided multimodal image understanding in remote sensing. IEEE Transactions on Geoscience and Remote Sensing, 2026b.

Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A survey. Medical Image Analysis, 102: 103551, 2025.

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.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832, 2021.

APPENDIX CONTENTS

Appendix A. Hierarchy-Aware Visual Understanding Analysis 17
Appendix B. Layer-wise Alignment Analysis 17
Appendix C. More Implementation Details 18
Appendix D. Hierarchy-Aware Analysis of Fine-Grained Visual Understanding 21
Appendix E. Compute Reporting 22
Appendix F. Evaluation Prompts 22
Appendix G. More Layer-wise Alignment Analysis 23
Appendix H. More Qualitative Case Study 24

A HIERARCHY-AWARE VISUAL UNDERSTANDING ANALYSIS

Table 7: Evaluation on FLUX-Reason across visual encoders and LLMs. 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$_{\uparrow 7.03}$	84.93$_{\uparrow 4.19}$	83.88$_{\uparrow 5.62}$	36.71$_{\uparrow 16.10}$	40.57$_{\uparrow 12.55}$	38.39$_{\uparrow 14.26}$	30.33$_{\uparrow 16.11}$	57.04$_{\uparrow 9.36}$	89.35$_{\uparrow 6.33}$	1.1185
Qwen2.5-VL-7B		83.55$_{\uparrow 6.35}$	85.23$_{\uparrow 3.89}$	84.38$_{\uparrow 5.12}$	36.25$_{\uparrow 16.56}$	41.35$_{\uparrow 11.77}$	38.47$_{\uparrow 14.18}$	30.41$_{\uparrow 16.03}$	58.41$_{\uparrow 7.99}$	90.78$_{\uparrow 4.90}$	1.1575
Visual Encoder–LLM Combinations											
CLIP	Qwen2.5-7B	88.77$_{\uparrow 1.17}$	87.59$_{\uparrow 1.41}$	88.17$_{\uparrow 1.29}$	45.36$_{\uparrow 7.61}$	46.18$_{\uparrow 5.63}$	45.48$_{\uparrow 6.53}$	39.34$_{\uparrow 6.67}$	59.99$_{\uparrow 5.14}$	91.90$_{\uparrow 3.53}$	1.036
DINOv3	Qwen2.5-7B	89.10$_{\uparrow 0.84}$	88.26$_{\uparrow 0.74}$	88.67$_{\uparrow 0.79}$	48.79$_{\uparrow 4.18}$	48.65$_{\uparrow 3.16}$	48.47$_{\uparrow 3.54}$	43.06$_{\uparrow 2.95}$	64.36$_{\uparrow 0.77}$	94.85$_{\uparrow 0.58}$	1.0099
Ours		**89.94**	**89.00**	**89.46**	**52.97**	**51.81**	**52.01**	**46.01**	**65.13**	**95.43**	**0.9980**
CLIP	Qwen3-8B	89.35$_{\uparrow 0.55}$	88.21$_{\uparrow 0.91}$	88.77$_{\uparrow 0.73}$	49.15$_{\uparrow 3.66}$	49.07$_{\uparrow 4.05}$	48.84$_{\uparrow 3.81}$	43.86$_{\uparrow 2.58}$	61.33$_{\uparrow 5.07}$	93.29$_{\uparrow 2.39}$	1.0132
DINOv3	Qwen3-8B	89.33$_{\uparrow 0.57}$	88.53$_{\uparrow 0.59}$	88.92$_{\uparrow 0.58}$	49.85$_{\uparrow 2.96}$	49.66$_{\uparrow 3.46}$	49.36$_{\uparrow 3.29}$	43.32$_{\uparrow 3.12}$	65.13$_{\uparrow 1.27}$	95.66$_{\uparrow 0.02}$	1.017
Ours		**89.90**	**89.12**	**89.50**	**52.81**	**53.12**	**52.65**	**46.44**	**66.40**	**95.68**	**1.0205**

B LAYER-WISE ALIGNMENT ANALYSIS.

As shown in Figure 5 and Figure 6, we compute the cosine similarity between the visual-conditioned hidden states and the corresponding reference textual states at each LLM layer. Our method achieves consistently higher cosine similarity than the CLIP baseline across different samples and layers. For CLIP+Llama, the similarity is relatively low in shallow and middle layers, and only gradually increases in deeper layers. This indicates that the visual representations produced by the CLIP baseline are mainly aligned with the LLM at a coarse semantic level, while lacking sufficient fine-grained structure to support early and middle-stage semantic integration. In contrast, Ours+Llama exhibits stronger alignment from the shallow layers and maintains high similarity through the middle and higher layers. The heatmap shows a broader and more stable high-similarity region, suggesting that `Meta-MemGran` provides visual features that are more compatible with the internal semantic space of the LLM.

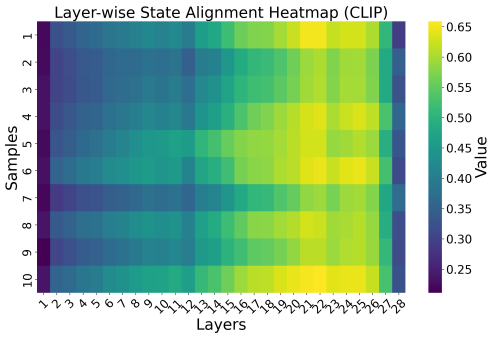


Figure 5: Cosine similarity between the reference state and each corresponding layer state for CLIP + Llama.

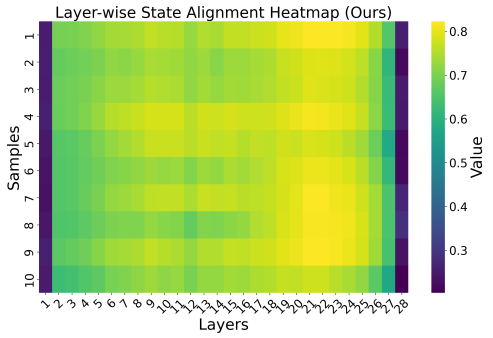


Figure 6: Cosine similarity between the reference state and each corresponding layer state for ours + Llama.

C More Implementation Details

C.1 Dataset Processing

We describe the preprocessing pipeline used to construct training and evaluation data for `Meta-MemGran`. The pipeline standardizes heterogeneous multimodal data into a unified conversational format. **Data Construction** Raw data are collected from `parquet` files or HuggingFace datasets, with all available splits merged into a single pool. Each sample is represented as a structured record containing an identifier, image input, and textual annotations. **Image Handling** Image inputs may appear in different formats (e.g., binary streams or array representations). We unify them into standard image files and reference them via relative paths, ensuring compatibility with multimodal training frameworks. **Conversation Formatting** Each sample is converted into a LLaVA-style tuple $(\texttt{id}, \texttt{image}, \texttt{conversations})$. The conversation consists of a human prompt requiring step-by-step visual reasoning and a corresponding response derived from ground-truth annotations. The prompt is appended with the special token `<image>` to ensure proper multimodal grounding. **Data Standardization** We normalize textual annotations by selecting a single target response per sample and enforcing a consistent format across datasets. This yields a unified schema that supports joint training over diverse benchmarks. **Dataset Splitting** We adopt a deterministic index-based split. Given $\mathcal{D} = \{x_i\}_{i=0}^{N-1}$, the test set is defined as $\{x_i \mid s \leq i \leq e\}$, and the remaining samples are used for training. This avoids randomness and ensures reproducibility across runs. **Output Format** The processed data are serialized into JSON files following a consistent schema, with each sample containing an image reference and a two-turn conversation. The resulting format is directly compatible with multimodal large language model training. **Data Usage and Evaluation Protocol** Beyond preprocessing, we specify the data allocation for training and evaluation. For general multimodal reasoning, we adopt a 200K subset of the FLUX-Reason (Fang et al., 2025) corpus as the training set. This subset retains broad coverage of reasoning patterns, including multi-step inference, compositional understanding, and detailed visual grounding. An additional 1K samples are held out from the same source as a validation set, constructed under identical preprocessing to ensure consistent in-domain evaluation. To examine out-of-distribution generalization, models are trained solely on FLUX-Reason without incorporating external supervision. Evaluation is conducted in a zero-shot manner on four benchmarks, each consisting of 1K uniformly sampled instances: SEED-Bench (Li et al., 2023a), A-OKVQA (Schwenk et al., 2022), SurgVLM (Zeng et al., 2025), and CC12M_and_Imagenet21K_Recap (Changpinyo et al., 2021). These datasets collectively probe factual knowledge, reasoning composition, and high-level captioning ability and domain-specific medical visual reasoning. This setup enables targeted assessment of medical visual reasoning without cross-domain leakage. All experiments strictly follow the above data partitions. No additional fine-tuning is performed on evaluation benchmarks, ensuring a clean zero-shot protocol.

C.2 Implementation Details

We incorporate multiple visual encoders and evaluate them with two language backbones, namely *Qwen-2.5-Instruct* and *Llama-3.2*. The multimodal training is conducted using DeepSpeed with ZeRO Stage-2 optimization. Training is performed on 8 NVIDIA A800-SXM4-80GB GPUs for a total of 2 epochs. Each GPU uses a per-device batch size of 8 with gradient accumulation steps of 2, resulting in an effective global batch size of 128. We adopt a cosine learning rate scheduler with an initial learning rate of 2×10^{-5} and a warmup ratio of 0.05. Mixed precision training is enabled with bf16 and tf32. Gradient checkpointing is applied to reduce memory usage, and data loading is optimized with multiple workers and pinned memory. Model checkpoints are saved periodically during training, with a maximum of 3 checkpoints retained. No intermediate evaluation is performed during training.

C.3 Evaluation Metrics.

To systematically evaluate the effectiveness of different visual encoders on the FLUX-Reason test set as well as zero-shot benchmarks, we employ a two-stage evaluation protocol. As an initial step, we compute BERTScore to measure textual similarity between generated outputs and reference annotations. Nevertheless, we observe that BERTScore exhibits only minimal variation across models, even when their outputs show noticeable qualitative differences. This suggests an inherent limitation of embedding-based metrics, which often fail to adequately reflect factual correctness, visual grounding, or the completeness of multimodal reasoning. To address this issue, we further introduce GPT-4o as an automated evaluator (Zheng et al., 2023; Li et al., 2024b; 2025; Tan et al., 2024). Based on extensive manual inspection, GPT-4o demonstrates a substantially higher consistency

with human judgments compared to BERTScore. Notably, for the FLUX-Reason benchmark, we do not rely on direct comparison with the provided textual ground truth. Although the annotations are detailed, they are not exhaustive in covering all objects, attributes, and relationships within an image. In practice, we find that `Meta-MemGran` often generates accurate and well-grounded descriptions that go beyond the scope of the reference annotations. Therefore, we instead evaluate model outputs against the visual content itself, where GPT-4o determines whether the reasoning process, visual grounding, and final conclusions are consistent with the image. For VQA tasks, GPT-4o is used to assess the factual accuracy of model predictions beyond simple lexical matching. For image captioning benchmarks, the evaluation focuses on semantic completeness, visual relevance, and the ability to capture key elements within the scene.

C.4 GPT-Based Hallucination Evaluation

In addition to conventional metrics (Accuracy, Recall, and BERTScore), we employ GPT-4o as an LLM-based judge for a fine-grained assessment of hallucination. Such model-based evaluation has been increasingly adopted in recent multimodal studies to assess factual consistency and visual grounding beyond reference-based metrics (Sun et al., 2024b; Liu et al., 2024; Jiang et al., 2026; Zhang et al., 2026a; He et al., 2026). Specifically, GPT-4o examines the generated response against the corresponding image and evaluates the extent to which its content is unsupported by or inconsistent with the visual evidence.

Following Mao et al. (Mao et al., 2026), as listed in Table 8, we adopt a 0–80 hallucination severity scale, where lower scores indicate better visual grounding and less hallucination. A score of 0 indicates no detected visually unsupported content, whereas 80 represents the highest hallucination severity. The scale is divided into eight qualitative levels that serve as holistic scoring anchors rather than deterministic thresholds based on the number of unsupported claims:

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.

The final hallucination score is obtained by averaging the sample-level scores over all valid evaluation samples. For reproducibility, we additionally provide the complete evaluation prompt and implementation details, including the GPT-4o configuration, decoding settings, image resolution, aggregation rule, and number of valid samples.

C.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 and 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 GPT-based judgments.

As listed in Table 9, the average hallucination scores from human and GPT-based evaluations were 34.30 and 34.10, respectively, with an absolute difference of only 0.20. Representative cases also show similar judgments between the two evaluators. These results demonstrate that the GPT-based

metric is highly consistent with human evaluation at the aggregate level, providing additional support for the reliability of our hallucination-related evaluation and conclusions.

Table 9: Comparison of human and GPT-based hallucination scores on 100 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, showcasing a colorful array of treats perfect for the spooky season. The stand, labeled "HALLOWEEN COOL'E PTS," features multiple orange shelves adorned with playful black ghost and pumpkin silhouettes, creating a cohesive and eye-catching design.	70	70
002	Couples stroll hand-in-hand along a serene lakeside path, their figures framed by the tranquil expanse of Lake Louise, where the calm waters mirror the surrounding mountains and lush greenery. The woman, dressed in a flowing floral dress with a delicate floral pattern, gracefully walks ahead, her long hair cascading down her back, while.....	Couple walking hand-in-hand along the grassy shore of Lake Dillon, their figures framed by the tranquil expanse of serene waters and the majestic backdrop of snow-capped mountains. The woman, dressed in a flowing floral dress with a delicate blush hue and red accents, her hair cascading down her back, leads slightly ahead, while the man, clad in a sharp navy blue suit, follows closely, their hands intertwined in a gesture of affection and unity. The lush.....	15	20
.....				
100	Vibrant and iconic, a classic 1969 Ford Mustang Shelby GT500 takes center stage in this meticulously crafted photograph, its sleek red body gleaming under the soft, diffused lighting of a late afternoon. The car's signature design elements are prominently displayed: the bold, chrome grille featuring the	Shiny and vibrant, a classic 1967 Ford Mustang commands attention as it is parked prominently in the center of a quiet urban street, bathed in soft natural light that accentuates its polished red exterior. The car's iconic design is immediately recognizable, featuring a sleek, low-slung body with sharp lines and curves that epitomize the muscle car era. The front grille proudly displays the Mustang emblem, while the	25	20
AVG Score	——	——	**34.30**	**34.10**

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

A potential concern regarding our memory-based design is whether the learned prototypes may limit the model's ability to generalize to unseen entities or out-of-distribution (OOD) samples. In particular, since the prototypes in the memory banks are constructed from the training data, one may hypothesize that an unseen entity without a corresponding prototype could be forcibly aligned with an inappropriate known prototype through the memory compatibility objective, thereby introducing systematic errors under distribution shift.

We emphasize that our experimental setting provides a direct empirical test of this concern (As listed in Table 10). Specifically, **all memory-bank construction and model training are performed exclusively on FLUX-Reason**. No images, annotations, entities, or prototypes from any of the other evaluation benchmarks are used during either memory construction or model training. Therefore, the memory banks contain no dataset-specific prototypes derived from SEED-Bench, A-OKVQA, SurgVLM, CC12M. Evaluation on these datasets is conducted in a strictly zero-shot manner.

Table 10: Generalization performance on in-distribution and zero-shot OOD benchmarks.

Dataset	Bank construction	Model training	Evaluation setting	Result
FLUX-Reason (Fang et al., 2025)	✓	✓	In-distribution	SOTA
SEED-Bench (Li et al., 2023a)	×	×	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

Importantly, the memory banks contain no dataset-specific prototypes from these OOD benchmarks. If the memory compatibility term forced unseen entities to match inappropriate known prototypes, systematic degradation would be expected under such distribution shifts. Instead, our method consistently achieves state-of-the-art performance across zero-shot OOD benchmarks, including the specialized surgical domain of SurgVLM. These results indicate that the Entity Bank does not act as a closed-set dictionary requiring an exact prototype for every test-time entity. Rather, the learned prototypes serve as **transferable semantic anchors** that regularize the representation space, enabling the model to generalize to unseen entities and distributions without requiring one-to-one prototype correspondence.

D Hierarchy-Aware Analysis of Fine-Grained Visual Understanding

To further examine whether the proposed method genuinely captures structured visual information beyond improving final-answer accuracy, we conduct a hierarchy-aware qualitative analysis at three levels: object, relationship, and layout. These three levels characterize complementary aspects of fine-grained visual understanding: object-level understanding focuses on recognizing detailed visual entities and their attributes; relationship-level understanding examines interactions and spatial relations among entities; and layout-level understanding evaluates the model's ability to recover the global spatial organization of multiple visual elements.

Table 11 presents representative examples together with the key information contained in the ground truth and the corresponding information recovered by our model.

Table 11: Qualitative analysis of hierarchical visual understanding.

Hierarchy	Case	Key Information in GT	Information Recovered by `Meta-MemGran`
Object Level	Case 0526	Pink princess dress, white hanger, sleeveless lace bodice, bow at the waist, layered tulle skirt	Correctly identifies the "pink princess dress," "white hanger," "sleeveless bodice," "lace detailing," "bow," and "voluminous tulle layer."
Object Level	Case 0528	Compartmented storage box, colored beads, beads placed in individual compartments, multiple colors	Correctly identifies a "transparent storage box," "colored beads," "individual compartments," and multiple colors such as red, purple, yellow, and pink.
Relationship Level	Case 0502	The couple walks hand-in-hand; the woman walks slightly ahead; the man follows closely behind; the lake reflects the surrounding mountains	Correctly captures that the couple "stroll 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	The wooden spoon is located in the bowl; cucumber slices and cookies are distributed around the bowl; the bowl is in the foreground	Correctly captures that a "small white bowl takes center stage," that "a wooden spoon is inserted into the dip," and that cucumber slices and cookies surround the bowl.
Layout Level	Case 0525	The fruits are arranged in concentric circles; blueberries are in the center; strawberries, kiwis, and apricot slices form different ring layers	Correctly recognizes the "concentric circles," identifies "a cluster of blueberries" at the center, and recovers the surrounding ring structure formed by different fruits.
Layout Level	Case 0520	Three figures are arranged from left to right; the left figure wears a navy blue suit, the middle figure a dark gray suit, and the right figure a light brown suit; the background is blurred	Correctly distinguishes the "leftmost man," "center figure," and "rightmost man," associates them with the corresponding clothing colors, and also recognizes the blurred background.

These examples provide complementary evidence that the proposed memory mechanism does not merely improve the correctness of the final response. At the object level, the model preserves fine-grained attributes such as color, material, clothing structure, and object components. At the relationship level, it captures interactions, relative positions, and directional relations between entities. At the layout level, it recovers structured spatial organizations involving multiple objects, such as center–surround arrangements, concentric structures, and left-to-right configurations.

Taken together, this hierarchy-aware analysis suggests that the performance gains of our method are accompanied by improved recovery of fine-grained object attributes, inter-object relationships, and global spatial layouts, providing more direct and interpretable evidence for its structured visual-understanding capability.

E COMPUTE REPORTING

All experiments were conducted using `DeepSpeed` with ZeRO Stage-2 optimization on 8 NVIDIA A800-SXM4-80GB GPUs. We evaluated multiple visual encoders together with two language backbones, namely *Qwen-2.5-Instruct* and *Llama-3.2*. Training was performed for 2 epochs with a per-device batch size of 8 and gradient accumulation steps of 2, resulting in an effective global batch size of 128. Optimization was carried out using a cosine learning rate scheduler with an initial learning rate of 2×10^{-5} and a warmup ratio of 0.05. Mixed-precision training with bf16 and tf32 was enabled to improve computational efficiency. Gradient checkpointing was further applied to reduce GPU memory consumption, while data loading was accelerated through multiple workers and pinned memory. The overall compute setup was designed to balance training efficiency and scalability for MLLMs. Distributed training with ZeRO Stage-2 significantly reduced memory overhead and enabled efficient utilization of multi-GPU resources. In addition, checkpoint management was optimized by periodically saving model states while retaining at most three checkpoints, thereby controlling storage consumption during training. Since no intermediate evaluation was conducted, the compute budget was primarily dedicated to model optimization and multimodal alignment.

F EVALUATION PROMPTS

To obtain reliable and consistent evaluation scores, we design a structured prompt that guides a multimodal model to assess the quality of generated image descriptions. The evaluation focuses on two complementary aspects, namely accuracy and hallucination, which together reflect both the correctness and the faithfulness of the model output with respect to the visual content. Given an image and a corresponding model-generated description, the evaluator is instructed to compare the textual output directly against the visual evidence. The prompt explicitly defines the scoring criteria to reduce ambiguity. The accuracy score ranges from 0 to 100 and measures how well the description aligns with the actual content of the image. A higher score indicates that the description correctly captures objects, attributes, and relationships present in the image. In parallel, a hallucination score ranging from 0 to 80 quantifies the extent to which the description introduces content that is not supported by the image. A higher hallucination score therefore reflects a greater degree of unsupported or fabricated details. The prompt enforces a strict output format by requiring a valid JSON object with two integer fields, which correspond to the accuracy score and the hallucination score. This constraint ensures that the evaluation results can be parsed automatically without post-processing ambiguity. In addition, the evaluation is conducted under a low temperature setting to encourage deterministic and stable scoring behavior across samples. During inference, each image is encoded in base64 format and provided together with the evaluation prompt as a multimodal input. The evaluator processes each sample independently, and the resulting scores are accumulated across the dataset. Invalid responses are filtered out to prevent them from affecting the final statistics. The average accuracy and hallucination scores are then computed over all valid samples, providing a concise summary of model performance. This prompt design emphasizes clarity, explicit criteria, and structured output, which together help improve the consistency and reproducibility of automatic evaluation.

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}
```

G More Layer-wise Alignment Analysis

We conduct a comparative analysis of how visual representations evolve across network layers when integrated with two distinct LLM backbones. The results (Figures 7, 8, 9, 10, 11, 12, 13, 14) reveal that `Meta-MemGran` maintains more stable and consistent feature transitions throughout the network than competing encoders. This behavior suggests that our approach better preserves meaningful information across layers, leading to improved semantic interpretation and more reliable reasoning.

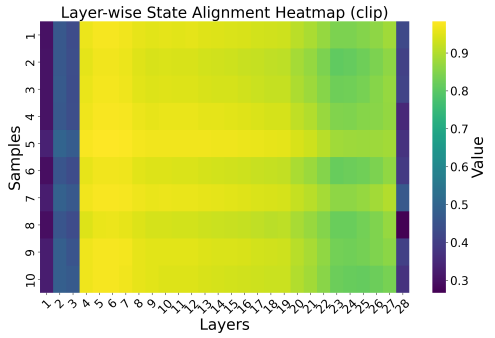


Figure 7: Cosine similarity between the reference state and the state of each corresponding layer for clip + qwen2.5

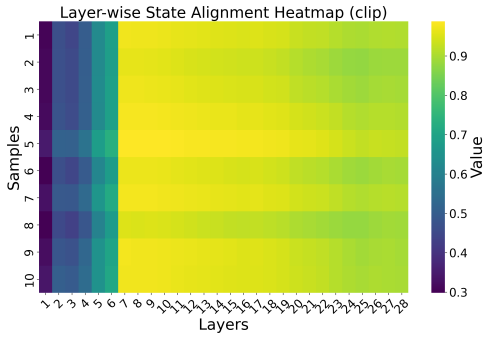


Figure 8: Cosine similarity between the reference state and the state of each corresponding layer for clip + qwen3

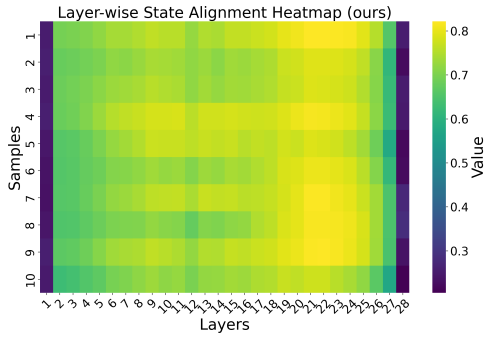


Figure 9: Cosine similarity between the reference state and the state of each corresponding layer for llama + ours

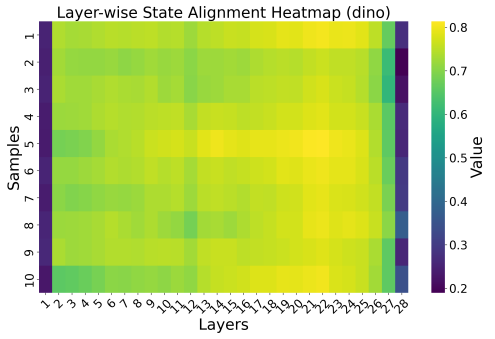


Figure 10: Cosine similarity between the reference state and the state of each corresponding layer for dinov3 + llama

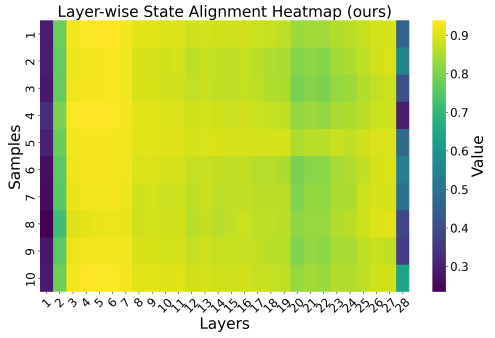


Figure 11: Cosine similarity between the reference state and the state of each corresponding layer for qwen2.5 + ours

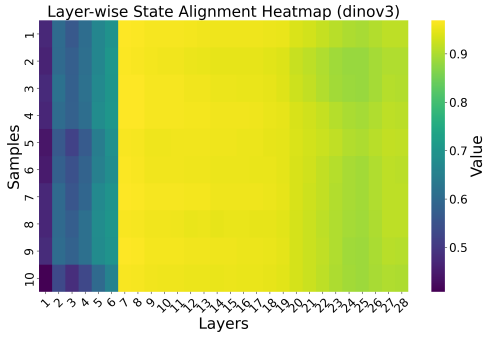


Figure 12: Cosine similarity between the reference state and the state of each corresponding layer for dinov3 + qwen3

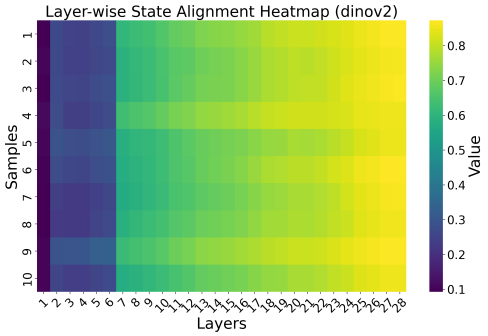


Figure 13: Cosine similarity between the reference state and the state of each corresponding layer for dinov2 + qwen3

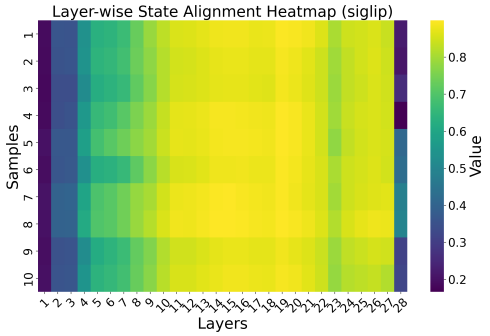


Figure 14: Cosine similarity between the reference state and the state of each corresponding layer for siglip + qwen2.5

H More Qualitative Case Study

To further illustrate the behavior of Meta-MemBank, we present a set of qualitative case studies as shown in Figure 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, . The results show that Meta-MemBank is able to maintain strong semantic consistency between the generated descriptions and the visual content. When errors do occur, they are typically restricted to subtle visual attributes, suggesting that the model may not fully capture certain fine-grained details. In comparison, methods built upon `CLIP` frequently demonstrate weaker alignment with the underlying image semantics and are more prone to generating hallucinated content. Notably, once an incorrect element is introduced, it often triggers additional spurious details that are semantically related but unsupported by the image, causing the overall description to progressively diverge from the ground truth.

Case Study 1: Reasoning Answer



CLIP: *Elegant couple in a traditional Chinese wedding attire, standing hand in hand, are captured in a serene and picturesque setting at the Grand Canyon, where the vast expanse of the canyon meets the azure waters of the Colorado River. The bride, adorned in a flowing red dress with intricate embroidery, and the groom, dressed in a classic black suit with a white shirt and a red tie, exude a sense of timeless romance. Their faces are filled with joy and anticipation, their hands intertwined as they gaze into each other's eyes. The background features a dramatic landscape of towering cliffs, lush greenery, and the shimmering river, creating a harmonious blend of natural beauty and human emotion. The composition is balanced and symmetrical, with the couple positioned centrally, drawing the viewer's attention to their intimate connection. The overall atmosphere is one of serenity and celebration, capturing the essence of a traditional Chinese wedding ceremony against the breathtaking backdrop of the Grand Canyon.*

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

Answer: A couple walks hand in hand along a serene shoreline, their figures captured in a romantic and intimate moment against the backdrop of a tranquil lake. The man, dressed in a crisp navy blue suit with a matching tie, stands slightly behind the woman, whose flowing pink and white floral dress contrasts beautifully with the natural surroundings. The woman's long, wavy hair cascades down her back, adding to the ethereal atmosphere. The composition centers on the couple, with the lake reflecting the sky and distant mountains, creating a harmonious blend of natural beauty and human connection. The scene is set in the scenic area of Lake Superior, where the lush greenery and majestic mountains frame the couple's romantic gesture. The overall mood is peaceful and timeless, emphasizing the enduring bond between the two individuals as they share this special moment together.

Ground Truth: *Couple walking hand-in-hand along the grassy shore of Lake Dillon, their figures framed by the tranquil expanse of serene waters and the majestic backdrop of snow-capped mountains. The woman, dressed in a flowing floral dress with a delicate blush hue and red accents, her hair cascading down her back, leads slightly ahead, while the man, clad in a sharp navy blue suit, follows closely, their hands intertwined in a gesture of affection and unity. The lush green grass beneath them contrasts beautifully with the calm, reflective surface of the lake, which mirrors the dense forested hills and distant peaks. The overcast sky adds a soft, diffused light to the scene, enhancing the peaceful ambiance. This composition captures a moment of timeless connection, where the gentle rhythm of nature's embrace intertwines with the tender bond between two souls, evoking a sense of romance and serenity. The image, rendered in a photorealistic style, emphasizes the harmonious blend of human intimacy and natural beauty, making it a poignant depiction of love and the enduring allure of Lake Dillon.*

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

Case Study 2: Reasoning Answer



CLIP: *Model showcasing the Nike Air Max 270 React "Black" sneakers, a sleek and contemporary design that exudes both comfort and style. The shoes feature a predominantly black upper with subtle white accents, including a white Swoosh logo on the lateral side and a white heel tab. The midsole is prominently highlighted in a vibrant red, contrasting sharply with the black upper, while the outsole is a matte black with a distinctive waffle pattern. The model stands against a plain white background, emphasizing the clean lines and minimalist aesthetic. The composition is straightforward, focusing on the shoes themselves, with the model's hands resting naturally at their sides, highlighting the overall fit and design. The image captures the essence of modern athletic wear, blending functionality with a touch of boldness, making it a striking visual representation of the Nike Air Max 270 React "Black" model.*

OURS: *Question: Please analyze this image step by step. Explain your reasoning process. Describe this image and give as much information as possible*
Answer: Model posing confidently against a plain white background, showcasing the "Biker" jeans from the 2018 collection by H&M. The jeans feature a classic fit with a high waist and straight legs, complemented by a pair of black sneakers that add a modern touch. The model's hands are tucked into the pockets, highlighting the functional details such as the cargo pockets and belt loops. The overall composition is clean and minimalist, emphasizing the sleek design and quality of the jeans. The image is photorealistic, capturing the textures and details of the fabric and hardware, with a focus on the model's posture and the jeans' silhouette. The background is neutral, allowing the focus to remain on the product, making it ideal for a retail setting or promotional material.

Ground Truth: *Model wearing dark gray cargo pants with multiple functional pockets, including large front cargo pockets secured with silver buttons and additional side pockets, highlighting practicality and utility. The pants feature an elastic waistband with adjustable drawstrings, ensuring a comfortable and customizable fit. Paired with sleek black athletic shoes, the ensemble emphasizes both style and functionality, ideal for active wear. The model stands against a clean, white background, drawing attention to the garment's design and detailing. The composition focuses on the lower half of the body, showcasing the pants' fit and features from a slightly angled perspective, while the absence of visible branding maintains a neutral and versatile aesthetic. The overall presentation suggests suitability for outdoor activities or casual, everyday use, blending comfort with practicality.*

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

Case Study 3: Reasoning Answer



CLIP: *Cute and festive, a white teddy bear dons a vibrant red Santa Claus hat adorned with a fluffy white pom-pom, standing on a snowy landscape. The bear's expressive eyes and friendly smile add to its charming demeanor, while its white fur contrasts with the red hat. The scene is set against a background featuring twinkling stars and a gradient of blues and purples, evoking a magical winter atmosphere. The bear's paws are slightly raised, as if ready to greet the season, and the overall composition is centered. The image is rendered in a photorealistic style, The artist's signature, "M. M. M.," is placed at the bottom right corner.*

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

Answer: Cute and festive, a white bear dons a vibrant red Santa Claus hat adorned with a fluffy white pom-pom, standing proudly on a snowy landscape. The bear's round, expressive eyes and friendly smile add to its charming demeanor, while its fluffy white fur contrasts beautifully with the red hat. The scene is set against a soft, dreamy background featuring twinkling stars and a gentle gradient of blues and purples, evoking a magical winter wonderland. The bear's paws are slightly raised, as if ready to greet the season, and the overall composition is centered, drawing the viewer's attention to the bear's joyful expression. The image is rendered in a high-definition, photorealistic style, capturing every detail with clarity and warmth, making it a delightful representation of holiday cheer. The artist's signature, "M. M. M.," is subtly placed at the bottom right corner.

Ground Truth: *Whimsical and enchanting, a fluffy polar bear cub sits serenely on a bed of pristine snow, exuding pure joy and innocence. Adorned with a vibrant red Santa hat featuring white fur trim and a fluffy pom-pom, the cub's cheerful expression—complete with a closed eye and a wide, toothy grin—is the focal point of this charming digital illustration. The background, crafted by PAB, is a serene winter wonderland, with delicate snowflakes gently falling against a soft teal sky, enhancing the festive and peaceful ambiance. The cub's fur is rendered in soft, creamy tones, contrasting beautifully with the bright red hat and the cool blue hues of the surroundings. The overall style is cartoonish and highly detailed, blending whimsical charm with high-definition digital artistry, capturing the timeless spirit of holiday cheer and the warmth of shared celebration. The artist's signature, Äriv,ïs subtly placed in the bottom right corner, adding a personal touch to this enchanting scene.*

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

Case Study 1: Visual Question Answering (VQA)



CLIP: *To determine the season on the grassland where the elephants are grazing, we need to analyze the image and consider the typical characteristics of each season. Let's go through the choices step by step: 1. Rainy Season: This season is characterized by heavy rainfall, lush vegetation, and generally wet conditions. Elephants typically prefer areas with abundant water sources and vegetation, which is more common during the rainy season.*
2. Winter: Winter is associated with cold temperatures, snow, and less vegetation. Elephants generally prefer warmer climates and areas with more vegetation, which is less common during winter.
3. Dry Season: This season is marked by dry conditions and less vegetation, and often hot temperatures. Elephants prefer areas with more water and vegetation, which is less common during the dry season.
4. Fall: Fall is characterized by cooler temperatures, changing foliage, and less vegetation. Elephants generally prefer warmer climates and areas with more vegetation, which is less common during fall.
Given these characteristics, the most likely season for the grassland where the elephants are grazing is the rainy season. The image shows lush vegetation and water sources, which are more typical during the rainy season.
Final Answer: (A) rainy
OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\n Question: What season is it on the grassland where the elephants are grazing?\n Choices:\n (A) rainy\n (B) winter\n (C) dry\n (D) fall*
Answer: Based on the image provided, the season depicted in the grassland where the elephants are grazing is (C) dry. The reasoning is as follows:
1. Visual Cues: The landscape shows sparse vegetation and dry, brown grass, which are typical characteristics of a dry season. The lack of lush greenery and the presence of dry, withered plants indicate a period of reduced rainfall.
2. Elephant Behavior: Elephants are often seen in such environments during the dry season as they seek out water sources and food. The elephants in the image are grazing, which is consistent with their behavior during this time of year.
3. Background Elements: The background includes bare trees and sparse foliage, further supporting the idea of a dry season. In a rainy season, one would expect more vibrant and lush vegetation.
4. Color Palette: The overall color palette of the image is predominantly brown and beige, which is characteristic of a dry season in grassland ecosystems.
Therefore, the correct answer is (C) dry,
Ground Truth: *Reasoning: The grass is not getting enough water. it is turning brown. The season is dry. The grass is brown and dying.\nFinal answer: dry*

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

Case Study 2: Visual Question Answering (VQA)



CLIP: *<think>\nOkay, let's try to figure out why the person is wearing a white jacket based on the image and the choices given. The options are nurse, doctor, cold, or chef. \n\nFirst, I need to visualize the image. The user mentioned the image is about a person wearing a white jacket. Since I can't see the image, I'll have to rely on common scenarios where white jackets are worn. \n\nWhite jackets are commonly associated with certain professions. For example, doctors and nurses often wear white coats as part of their uniforms. Chefs also wear white jackets, usually in the kitchen. However, the term "cold" here might refer to the weather, like wearing a jacket because it's cold outside. But the question is asking why the person is wearing the jacket, so it's more about the reason related to their role or activity.\n\nLooking at the choices, (A) nurse and (B) doctor are both healthcare professionals who typically wear white coats. (D) chef is another profession that uses white jackets. (C) cold is a condition, but it's more about the weather or health, not a reason for wearing the jacket. \n\nWait, but the question is phrased as "why is the person wearing a white jacket.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\nQuestion: Why is the person wearing a white jacket? \nChoices: \n(A) nurse \n(B) doctor \n(C) cold \n(D) chef*
Answer: <think>\nOkay, let's try to figure out why the person is wearing a white jacket based on the image description provided. The question gives four choices: nurse, doctor, cold, or chef. \n\nFirst, I need to analyze the image details. The description mentions elements like a white jacket, a person, and some items like a stove, baking dishes, and possibly ingredients. The presence of a stove and baking items suggests a kitchen setting. The person is likely involved in cooking or food preparation.\n\nLooking at the choices, c̈hef(D) seems plausible because chefs often wear white jackets, also known as chef coats. The other options: nurse (A) and doctor (B) typically wear different uniforms, like scrubs or white coats, but not specifically white jackets. C̈old(C) is a condition, so that doesn't explain the clothing choice. \n\nThe image also includes elements like a stove, baking dishes, and possibly pastries or bread, which are common in a kitchen. The person's attire matches a chef's uniform. The white jacket is a key part of a chef's outfit, so the answer is likely (D) chef.

Ground Truth: *Reasoning: He is a chef. The person wants to cook. Chefs usually wear all white or all black chef's jackets.\nFinal answer: chef.*

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

Case Study 3: Visual Question Answering (VQA)



CLIP: *The image captures a serene scene of people walking along a beach, with the sun shining down on them. The beach is lined with a series of wooden jetties, each with a small wooden platform at its end, where people are standing and walking. The shoreline stretches out to the horizon, with a few boats visible in the distance. The sky above is clear, with a few wispy clouds scattered across it. The overall atmosphere is peaceful, with the sound of waves gently lapping against the shore creating a soothing background noise.\n\n To answer the question, the time of day can be inferred from the image. The sun is high in the sky, casting a warm glow over the scene. This suggests that it is late morning or early afternoon, as the sun is not yet at its peak but is still high enough to cast a strong shadow on the beach. The absence of any clouds or shade also implies that it is a sunny day, which further supports the conclusion that it is late morning or early afternoon.*

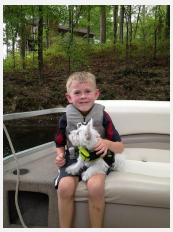
OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\n Question: What time of the day are people exploring the beach?\n Choices:\n (A) mixed tide\n (B) spring tide\n (C) high tide\n (D) low tide\n* <image>

Answer: The image depicts a serene beach scene with a calm, shallow body of water, likely a lagoon or estuary, where people are leisurely exploring the shoreline. The water is clear and reflective, with a few individuals walking along the edge, some holding surfboards, suggesting they are engaged in recreational activities such as walking or surfing. The beach is relatively empty, with only a few people visible, creating a peaceful and tranquil atmosphere. The sky is overcast, adding to the subdued and calm mood of the scene. The composition focuses on the beach and the water, with the distant shoreline and trees providing a sense of depth and context. The overall aesthetic is naturalistic, capturing the beauty of a quiet, unspoiled beach environment. Based on the scene, it appears to be during a ***low tide*** *(D), as the water is shallow and recedes significantly from the shore, allowing people to walk on the sand and explore the area comfortably.*

Ground Truth: *Reasoning: The tide is far out and you can see a lot of sand. The water has receded and you can see puddles of water that have been left as the tide went out. The ocean is not very high. it is easy to see the sand.\nFinal answer: low tide*

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

Case Study 4: Visual Question Answering (VQA)



CLIP: *Question: What is the name of the dog breed in the image?\n \n Choices:\n (A) Poodles\n (B) Retriever\n (C) Bulldog\n (D) Pomeranian\n \n The correct answer is (B) Retriever. The image depicts a young girl wearing a bright orange life jacket and a pair of water shoes, standing on a dock overlooking a serene lake surrounded by lush greenery and towering trees. The girl's hair is styled in a ponytail, and she has a cheerful expression, suggesting she is excited for an adventure. The dock, adorned with a wooden railing and a rope ladder, extends into the water, hinting at a fun and active outing. The overall scene captures a tranquil yet vibrant atmosphere, with the girl's enthusiasm and the natural beauty of the surroundings creating a sense of joy and exploration. The description of the girl's outfit, including a life jacket and water shoes, suggests a water-based activity, which aligns with the characteristics of a retriever, a breed known for its love of water and retrieving abilities. This makes option (B) Retriever the most fitting choice.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\nQuestion: What is the name of this dog breed?\nChoices:\n(A) poodles\n(B) retriever\n(C) bulldog\n(D) pomeranian\n<image>*

Answer: *<think>\nOkay, let's see. The user provided an image and a question asking for the name of the dog breed. The choices are poodles, retriever, bulldog, and pomeranian. The image description mentions a child sitting on a boat's deck, wearing a yellow life jacket, and a white dog with a fluffy coat. The dog is sitting next to the child, and the background is a forest with a lake.\n\nFirst, I need to focus on the dog's characteristics. The description says the dog has a fluffy coat. Poodles are known for their curly, fluffy coats, especially the standard poodle. Retriever dogs, like golden retrievers, have a medium to long coat but not as fluffy as poodles. Bulldogs have a short, wrinkled coat, so that's not matching. Pomeranians have a fluffy, silky coat, but they are smaller. The image might show the dog's size, but since the child is sitting next to it, maybe the dog is medium-sized. However, the key detail here is the fluffy coat. Poodles are often associated with that fluffy look, especially the standard variety. The other options don't fit as well. The answer is likely poodles.*

Ground Truth: *Reasoning: The breed is a poodle. The dog is very fluffy. It looks like a terrier but has the size and color of some poodles.\nFinal answer: poodles*

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

Case Study 5: Visual Question Answering (VQA)



CLIP: *The scene depicts a busy street scene during a rainy day, with individuals walking in the foreground. A woman in a dark-colored raincoat and holding a black umbrella stands near a parked car, her gaze directed toward the umbrella held by a man in the background. The umbrella casts a shadow on the wet pavement, creating a sense of movement and interaction between the two figures. The rain falls heavily, creating a misty veil that surrounds the scene, with water pooling on the ground and reflecting off the wet surfaces. The background features a mix of pedestrians, including a man with a backpack and a woman in a yellow raincoat, all moving in different directions. The street is lined with trees, their branches swaying gently in the wind, and there are utility poles and signs visible in the background. The overall atmosphere is one of dynamic movement and interaction, with the rain adding a sense of urgency and energy to the scene. The composition highlights the contrast between the figures and the wet, reflective surfaces, creating a visually striking and dynamic image that captures the essence of a rainy day in the city.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\n Question: What type of rain is this called?\n Choices:\n (A) average\n (B) drizzle\n (C) sprinkle\n (D) downpour\n <image>*

Answer: *A woman walks briskly along a wet sidewalk, her black boots sinking slightly into the muddy ground, as she carries a black umbrella over her head. The scene is set in a rainy urban environment, with a car parked nearby and a streetlight casting a soft glow through the misty air. The woman's attire includes a dark jacket and a scarf, adding to the somber tone of the rainy day. In the background, another figure can be seen walking towards the camera, also equipped with an umbrella, suggesting a shared experience of the weather. The overall composition captures the essence of a typical rainy day in an urban setting, where the rain has turned the streets into a reflective, almost surreal landscape. The choice of d̈ownpour(D) best describes the type of rain in this image.*

Ground Truth: *Reasoning: The way the woman uses the umbrella. The rain is really heavy and uncomfortable to be in. You can see how hard it is raining and how much it is.\nFinal answer: downpour*

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

Case Study 6: Visual Question Answering (VQA)



CLIP: *The image captures a serene scene of a large field where rows of sheep graze peacefully under the warm sunlight, their coats glistening with dew. In the foreground, a bright red tractor with a large, open cab drives down the center of the field, its tires leaving tracks in the dry earth. The tractor is identified as a Ëndt 900 Vario,ä well-known model of agricultural machinery, suggesting its use in farming or livestock management. The sheep, ranging in color from white to black, are arranged in neat rows, their heads bent towards the ground as they feed on the lush green grass. The background is a vast, open expanse of the field, with the tractor and sheep forming a striking contrast against the natural landscape. The composition emphasizes the harmony between the machinery and the animals, highlighting the efficiency and productivity of the farming operation. The overall scene exudes a sense of rural tranquility and productivity, capturing the essence of agricultural work in a photorealistic manner.*

OURS: *Question: Please answer the question based on the image and explain your reasoning step by step.\nQuestion: What is the occupation of the person driving?\n Choices:\n (A) waiter\n (B) farmer\n (C) cashier\n (D) musician\n <image>*

Answer: *A tractor, with its distinctive red color and prominent front-mounted bucket, navigates through a vast field filled with sheep, creating a serene yet industrious scene. The tractor, likely used for farming purposes, moves slowly across the open landscape, its wheels leaving faint tracks in the soft earth. The sheep, uniformly white and grazing peacefully, are scattered across the field, their bodies forming a natural pattern that contrasts with the mechanical presence of the tractor. The image captures a moment of rural tranquility, where the farmer's daily work is evident in the organized movement of the livestock. The composition emphasizes the harmony between human activity and nature, with the tractor serving as a central element that connects the viewer to the broader context of agricultural life. The overall atmosphere is one of calm productivity, highlighting the essential role of farming in sustaining rural communities. Therefore, the occupation of the person driving is most likely (B) farmer.*

Ground Truth: *Reasoning: The place is full of sheep that shows the person is a farmer. Farmer is the obvious profession as the picture shows. With the tractor he is in and the livestock shown it is easy to surmise his profession.\nFinal answer: farmer*

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

Case Study 7: Visual Question Answering (VQA)



CLIP: *Question: In which country is this bus located? \n\nChoices: \n(A) uk\n(B) cuba \n(C) usa \n(D) mexico \n\nAnswer: (C) usa* ***OURS:*** *Question: Please answer the question based on the image and explain your reasoning step by step.\nQuestion: In which country is this bus located?\nChoices:\n(A) uk\n(B) cuba\n(C) usa\n(D) mexico\n<image>*
Answer: *A vibrant red double-decker bus, adorned with the distinctive L̈ondonsïgn, stands prominently on a street in London, UK, as a group of people gather around it. The bus, with its classic design and bright red color, serves as a focal point, drawing attention to its historical significance and the bustling urban environment. The scene captures a moment of curiosity and interaction, with individuals dressed in casual attire—some wearing jackets and others carrying backpacks—standing close to the bus, their faces partially obscured by the bus's windows. The background reveals a typical London street scene, with buildings and other vehicles in the distance, adding to the sense of place. The bus's presence, along with the people's reactions, suggests a moment of shared interest or perhaps a special event, highlighting the cultural and social dynamics of the city. The overall composition emphasizes the bus as a central element, blending seamlessly with the urban landscape while capturing the essence of everyday life in London\n\nAnswer: (A) uk.*
Ground Truth: *Reasoning: Most double decker buses designed like this are traditionally found in the uk. This is a double decker bus used in london England is known for its double-decker buses.\nFinal answer: uk*

Figure 24: Qualitative case study comparing CLIP and our method.