MIRA : MULTIMODAL INTERLEAVED REASONING WITH ADAPTIVE VISUAL TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) have demonstrated remarkable semantic understanding. Yet, their reliance on discrete linguistic tokens fundamentally bottlenecks performance on vision-centric tasks that demand fine-grained perceptual understanding, such as spatial localization, depth estimation, and detailed visual recognition. Recent advances in latent reasoning attempt to mitigate this by operating in continuous spaces. However, current latent methods either lack meaningful visual representations or adopt fixed reasoning templates that generate a predetermined set of visual tokens regardless of the input task. To overcome these limitations, we propose MIRA (**M**ultimodal **I**nterleaved **R**easoning with **A**daptive Visual Tokens), a novel framework that augments the conventional text vocabulary with continuous visual latent tokens distilled from vision foundation models. Our generated sequences seamlessly alternate between text and visual tokens, forming an interpretable interleaved reasoning chain. We further introduce Interleaved Reasoning Policy Optimization (IRPO), which incentivizes the model to actively select task-relevant visual tokens while preserving perceptual fidelity. Extensive evaluations demonstrate that MIRA significantly improves spatial and functional reasoning capabilities across diverse vision-centric benchmarks.

1 INTRODUCTION

Vision-Language Models (VLMs) (Bai et al., 2025; Liu et al., 2024; 2023; 2025c; Team et al., 2024; Zhu et al., 2025) have achieved remarkable progress across a wide range of multimodal tasks by integrating visual representations with the reasoning capabilities of large language models. However, their reasoning processes predominantly unfold through discrete linguistic tokens, creating a bottleneck for tasks that require detailed visual understanding (Yang et al., 2025b). While language effectively conveys high-level concepts, it is less suited to expressing dense visual information, including geometric structure, precise spatial relationships, and fine-grained appearance details (Liu et al., 2025b; Pham & Ngo, 2025). Consequently, strong language-based reasoning does not necessarily translate into reliable performance on vision-centric tasks such as object counting, relative depth estimation, and fine-grained visual recognition (Chen et al., 2025d; Yang et al., 2026).

A natural approach to mitigate this limitation is to reintroduce visual information through external tools during reasoning. Several methods leverage auxiliary models to edit images (Wu et al., 2026), or crop and re-encode salient image regions within the VLM reasoning process (Su et al., 2026; Zheng et al., 2026). Although these tool-augmented approaches ground textual reasoning with fine-grained visual evidence, they come with notable drawbacks. First, because re-encoding image regions necessitates re-feeding and processing a large number of additional visual tokens at each reasoning step, these methods incur high inference costs and latency (Qin et al., 2025). Moreover, while these methods are effective in retrieving relevant visual details, reasoning purely in text limits their ability to carry rich visual semantic information across reasoning steps (Li et al., 2026a). These limitations motivate a shift from external tool-augmented reasoning toward approaches that internalize visual processing within the model itself.

Along this direction, recent works (Li et al., 2026a; Wang et al., 2026b; Yang et al., 2025b) have explored internalizing tool-use by replacing discrete tool-calls with continuous latent embeddings within the reasoning chain. While this avoids the high latency of external tools, the resulting representations remain abstract and cannot be decoded into interpretable visual evidence. To address

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

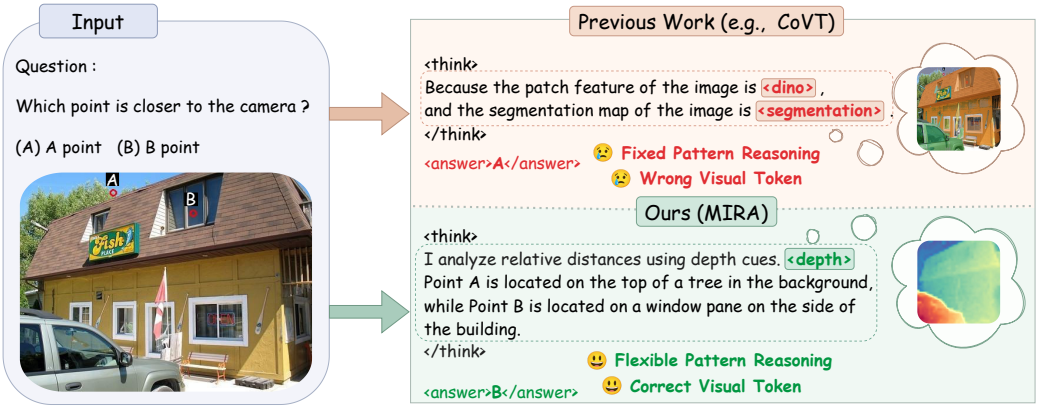


Figure 1: **Illustration of adaptive visual token selection and visually grounded reasoning in MIRA.** Given a spatially demanding task (relative depth estimation), MIRA (bottom) demonstrates context-driven adaptability by dynamically selecting task-relevant visual tokens (e.g., depth), whereas CoVT Qin et al. (2025) (top) relies on fixed reasoning patterns and retrieves task-irrelevant tokens.

this, several methods further introduce visual alignment, where latent tokens are supervised by vision foundation models to preserve grounded perceptual meaning (Chen et al., 2025d; Jeon et al., 2026; Qin et al., 2025). Although these methods successfully inject richer visual semantics into the reasoning process, they often adopt fixed reasoning templates, generating a predetermined set of visual tokens regardless of the input task rather than actively selecting task-relevant features. These limitations highlight a key gap, as current methods either lack explicit visual grounding or fail to adaptively select and integrate task-relevant visual representations into the reasoning process.

To bridge this gap, we propose **MIRA** (**M**ultimodal **I**nterleaved **R**easoning with **A**daptive Visual Tokens), a two-stage training framework that builds on the VFM-aligned token vocabulary of CoVT (Qin et al., 2025) and enables VLMs to adaptively invoke visual evidence during reasoning while maintaining semantically grounded latent representations. As illustrated in Figure 1, MIRA learns to select task-relevant visual modalities and interleave their corresponding tokens with textual reasoning. In the first stage, we curate 43K interleaved reasoning traces and perform Interleaved Supervised Fine-Tuning (SFT), combining language modeling with visual alignment objectives to establish grounded interleaved reasoning capabilities. We then introduce Interleaved Reasoning Policy Optimization (IRPO), which incentivizes the model to actively select task-relevant visual tokens with an F1-based token-selection reward, while a visual alignment loss preserves perceptual fidelity throughout optimization. Extensive experiments demonstrate that MIRA achieves higher average performance than the evaluated VLM reasoning and latent reasoning baselines across diverse vision-centric benchmarks.

Our main contributions are summarized as follows.

- We propose MIRA, a framework for adaptive interleaved visual reasoning built on VFM-aligned visual tokens. Using 43K curated reasoning traces, we train VLMs to integrate task-relevant visual modalities into textual reasoning beyond fixed templates.
- We introduce IRPO, which combines an F1-based token-selection reward with visual alignment regularization to improve task-relevant modality selection while maintaining grounded visual representations during policy optimization.
- We demonstrate strong performance across diverse vision-centric benchmarks, improving average accuracy by 7.4 percentage points over Qwen2.5-VL-7B. Human-annotated modality-selection evaluations, controlled comparisons, and component ablations provide evidence for the effectiveness of our approach.

2 Related Works

2.1 Reasoning in Vision-Language Models

Inspired by the success of DeepSeek-R1 Guo et al. (2025), reinforcement learning (RL) has emerged as the predominant paradigm for eliciting chain-of-thought (CoT) reasoning in vision-language models (VLMs), superseding earlier SFT-based approaches Jiang et al. (2024); Shao et al. (2024); Xu et al. (2025). Subsequent efforts explore diverse training recipes Chen et al. (2025b;c); Wei et al. (2025) and reward designs Fan et al. (2025); Wang et al. (2026a;c); Zhang et al. (2025). Beyond outcome-based rewards, several works introduce perception-aware reward signals. VisionReasoner Liu et al. (2026) adopts bounding-box rewards, Visionary-R1 Xia et al. (2025) incorporates auxiliary caption rewards, and Vision-SR1 Li et al. (2026b) proposes a self-rewarding mechanism based on caption utility for answer prediction. Another line of work pursues tool-augmented reasoning, where VLMs invoke auxiliary visual operations during inference. Visual Sketchpad Hu et al. (2024) draws intermediate sketches via specialist vision models, VTool-R1 Wu et al. (2026) applies Python-based visual editing, and others Qi et al. (2024); Sarch et al. (2026); Su et al. (2026); Zheng et al. (2026) adopt cropping to reintroduce salient patches into the reasoning chain. Despite these advances, all such methods confine reasoning to the discrete text modality, either discarding fine-grained spatial and appearance details when projecting continuous visual signals into language tokens, or relying on external tools whose asynchronous multi-turn invocations re-feed a large number of additional visual tokens at each reasoning step, incurring substantial inference overhead.

2.2 Latent Visual Reasoning

Recent efforts seek to address the aforementioned limitations by augmenting the reasoning chain with continuous latent representations Chen et al. (2025a); Li et al. (2026a); Liu et al. (2025a); Qin et al. (2025); Tong et al. (2025); Wang et al. (2026b); Yang et al. (2025b); Zhang et al. (2026b). Mirage Yang et al. (2025b) introduces latent visual tokens as compact intermediate representations during reasoning, bypassing the computational overhead of pixel-level decoding. LVR Li et al. (2026a) generates a dedicated sequence of continuous visual embeddings before formulating the final textual answer, optimizing the latents through a direct application of GRPO. Monet Wang et al. (2026b) further proposes Visual Latent Policy Optimization (VLPO) to enable reinforcement learning in continuous spaces by approximating the generation likelihood of latent representations. However, the latent representations in these methods remain abstract and cannot be decoded into interpretable visual evidence, limiting their fidelity in conveying visual semantics during reasoning. CoVT Qin et al. (2025) takes a different route by imposing visual alignment through foundation models (*e.g.*, SAM Kirillov et al. (2023), DINOv2 Oquab et al. (2024), Depth Anything v2 Yang et al. (2024), and PidiNet Su et al. (2021)), but its reasoning pattern remains rigid and passive, generating a predetermined set of visual tokens regardless of the input task. To bridge these gaps, we aim to simultaneously endow latent representations with interpretable visual semantics and realize a truly adaptive interleaved latent reasoning paradigm.

3 Method

3.1 Problem Formulation

We first define the standard setting for Vision-Language Models (VLMs). Given an input image $\mathbf{I}$ and a textual query $\mathbf{Q}$, a model $\mathcal{M}$ parameterized by θ generates an output sequence $\mathbf{y} = (y_1, \ldots, y_T)$ by modeling the conditional probability:

$$p_\theta(\mathbf{y} \mid \mathbf{I}, \mathbf{Q}) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, \mathbf{I}, \mathbf{Q}). \tag{1}$$

To overcome the semantic bottleneck of text-only reasoning, we propose MIRA, a unified framework that enables adaptive multimodal reasoning by interleaving textual reasoning with visual latent tokens. Formally, the unified vocabulary is defined as $\mathcal{V}' = \mathcal{V}_{\text{txt}} \cup \mathcal{V}_{\text{vis}}$, where $\mathcal{V}_{\text{vis}}$ consists of modality-specific tokens (e.g., <|sam_pad|>, <|dino_pad|> and <|depth_pad|>), corresponding to segmentation, semantic, and depth representations, respectively. For each visual latent

token $y_t \in \mathcal{V}_{\text{vis}}$, the corresponding output hidden state h_{vis} is aligned with representations from pre-trained vision foundation models (Sec. 3.2). Crucially, these grounded hidden states can be optionally decoded by modality-specific decoders into interpretable visual outputs (e.g., segmentation masks, depth maps, or dense feature maps), making the visual information embedded in the reasoning chain explicitly interpretable. To optimize this framework, we adopt a two-stage training pipeline consisting of Interleaved Supervised Fine-Tuning (Sec. 3.2) and Interleaved Reasoning Policy Optimization (Sec. 3.3), and detail our inference strategy in Sec. 3.4.

3.2 Interleaved Supervised Fine-Tuning

Interleaved Dataset Synthesis. To construct the interleaved SFT dataset, we leverage Gemini 3 Flash Team et al. (2024) to synthesize reasoning traces, as illustrated in Fig. 2 (left). To improve the correctness and consistency of the synthesized traces, we provide the teacher model with the ground-truth answer, functional definitions of the visual latent tokens, and few-shot interleaved reasoning examples. We then apply rule-based post-processing to filter out incorrectly formatted traces and invalid visual-token sequences. Full prompts and filtering rules are provided in Appendix A.1.

Objective Optimization. To equip visual latent tokens with grounded perceptual meaning, we align their output hidden states with representations from pre-trained vision foundation models. Following CoVT Qin et al. (2025), we adopt modality-specific projection heads and alignment objectives tailored to each visual modality. Let $h_{\text{vis}}^{(m)}$ denote the hidden state of modality m, and ϕ_m its corresponding projection head, which maps the hidden state to $f_{\text{vis}}^{(m)} = \phi_m(h_{\text{vis}}^{(m)})$. For SAM Kirillov et al. (2023) and Depth Yang et al. (2024), modality-specific decoders map the projected features $f_{\text{vis}}^{(m)}$ to segmentation masks and depth maps. These outputs are supervised using Dice and binary cross-entropy (BCE) losses for SAM and an L_1 loss for Depth. For DINO Oquab et al. (2024), the projected features are directly aligned with dense semantic features using mean squared error (MSE). The total alignment loss combines these modality-specific objectives.

$$\mathcal{L}_{\text{align}} = \sum_{m \in \{\text{SAM,Depth,DINO}\}} \mathcal{L}_m. \tag{2}$$

The final SFT objective combines the standard cross-entropy loss with the visual alignment loss:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{align}}, \tag{3}$$

where λ balances language modeling and visual alignment.

3.3 Interleaved Reasoning Policy Optimization

While Stage 1 equips the model with grounded interleaved reasoning capabilities, standard outcome-based reinforcement learning primarily supervises final-answer correctness, providing limited guidance on the intermediate perceptual reasoning process. Consequently, correct answers may still be rewarded despite irrelevant modality choices or poorly grounded visual representations. To address this gap, we introduce Interleaved Reasoning Policy Optimization (IRPO), which augments GRPO with an F1-based token-selection reward to encourage task-relevant modality selection and a visual alignment regularizer to preserve the perceptual meaning of visual hidden states.

Token-Selection Reward. To explicitly supervise intermediate perceptual decisions, we introduce an F1-based token-selection reward (r_{F1}). For each image-question pair, we use Gemini 3 Flash Team et al. (2024) to generate a pseudo ground-truth set $\mathcal{G} \subseteq \{\texttt{<|sam_pad|>}, \texttt{<|dino_pad|>}, \texttt{<|depth_pad|>}\}$ specifying the visual modalities required to answer the question. We verify these pseudo labels against human annotations in Appendix B. Let $\hat{\mathcal{G}}$ denote the modalities invoked by the model in its reasoning trace. We compute the F1 score between $\hat{\mathcal{G}}$ and $\mathcal{G}$:

$$r_{\text{F1}} = \frac{2 \cdot |\hat{\mathcal{G}} \cap \mathcal{G}|}{|\hat{\mathcal{G}}| + |\mathcal{G}|}. \tag{4}$$

By jointly balancing precision and recall, r_{F1} discourages the inclusion of irrelevant modalities and the omission of relevant ones, guiding the model's modality-selection decisions beyond final-answer correctness.

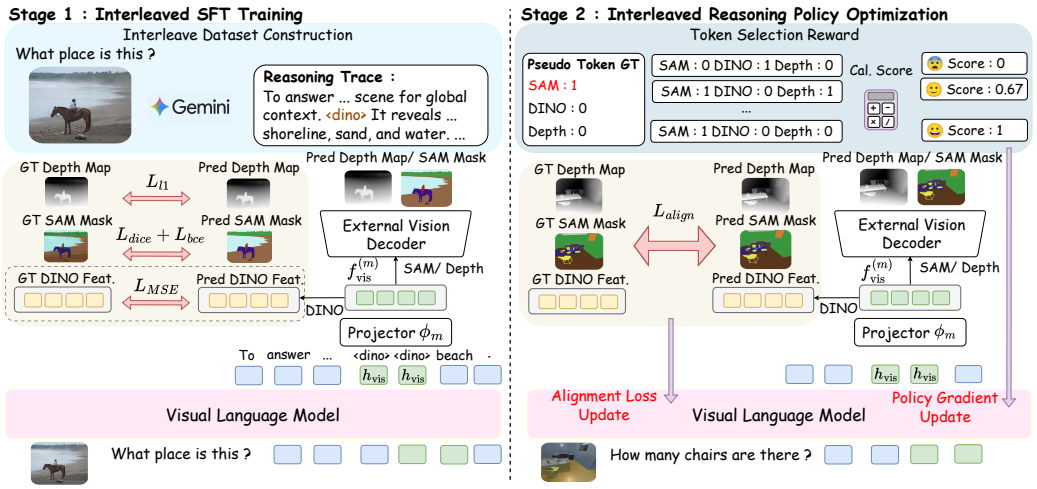


Figure 2: **Overview of the two-stage MIRA framework. Stage 1 (left): Interleaved SFT Training.** The VLM model learns to interleave textual reasoning and visual tokens seamlessly. **Stage 2 (right): Interleaved Reasoning Policy Optimization.** The model learns to adaptively utilize task-relevant visual tokens guided by a token-selection reward, while simultaneously preserving visual interpretability through alignment regularization.

The composite reward combines r_{F1} with a standard answer reward (r_{ans}) for final-answer correctness and a format reward (r_{fmt}) enforcing structured outputs (e.g., `<think>...</think>` and `<answer>...</answer>`):

$$R = r_{\text{ans}} + r_{\text{fmt}} + \omega_{\text{F1}} r_{\text{F1}}, \tag{5}$$

where ω_{F1} weights the contribution of the token-selection reward.

Policy Optimization with Alignment Regularization. We optimize the composite reward R via Group Relative Policy Optimization (GRPO) Guo et al. (2025). To ensure the visual hidden states remain decodable into interpretable visual outputs throughout RL updates, we further augment GRPO with the alignment regularizer $\mathcal{L}_{\text{align}}$ from Eq. 2. The core GRPO objective optimizes the reasoning policy while constraining updates via a KL penalty:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\left[\min\left(\rho_t(\theta)\hat{A}_t, \text{clip}(\rho_t(\theta), 1-\epsilon_l, 1+\epsilon_h)\hat{A}_t\right)\right] + \lambda_{\text{KL}}\mathbb{E}\left[\log \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\right] \tag{6}$$

where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}$ is the probability ratio, $\hat{A}_t$ is the advantage estimated from R, and ϵ_l, ϵ_h are the clipping bounds. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{GRPO}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \tag{7}$$

where λ_{align} scales the alignment regularization.

3.4 Inference

At inference, the model autoregressively generates a sequence $\mathbf{y}$ that adaptively interleaves text and visual latent tokens. Following the learned interleaved reasoning policy, the model decides which visual modalities to invoke and when, producing reasoning traces tailored to each task without external tool calls or auxiliary image inputs. The hidden states of generated visual latent tokens serve as intermediate perceptual context for subsequent reasoning. When interpretability is desired, these hidden states can be *optionally* decoded by modality-specific projection heads and decoders into structured visual outputs (e.g., segmentation or depth maps). Otherwise, the model reasons directly over the latent visual representations without decoding, avoiding additional decoding overhead.

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Implementation details. For Stage 1 (Interleaved SFT), we use a learning rate of 2×10^{-5} and set the alignment loss coefficient (λ in Eq. 3) to 0.75. The model is initialized from a Qwen2.5-VL-7B-based checkpoint with the VFM-aligned token vocabulary from CoVT Qin et al. (2025), covering SAM Kirillov et al. (2023), DINOv2 Oquab et al. (2024), and Depth Anything v2 Yang et al. (2024). For Stage 2 (IRPO), we build on EasyR1 Zheng et al. (2025) and train for 200 steps with a global batch size of 128 and 5 rollouts per prompt. We set $\lambda_{\text{align}} = 5 \times 10^{-3}$ (Eq. 7) and $\omega_{\text{F1}} = 0.5$ (Eq. 5). All experiments use 4 NVIDIA H200 GPUs.

Training Datasets. For interleaved Supervised Fine-Tuning (SFT), we construct our dataset by sampling examples from Visual CoT Shao et al. (2024) and SPAR-7M Zhang et al. (2026a), leveraging Gemini 3 Flash Team et al. (2024) to synthesize 43k high-quality interleaved reasoning traces. For the Interleaved Reasoning Policy Optimization (IRPO) stage, we source multi-modal QA pairs from Vision SR1-47k Li et al. (2026b) and the SAT Ray et al. (2025) dataset. In this RL phase, Gemini 3 Flash is utilized to automatically annotate the pseudo ground-truth sets of task-relevant visual modalities to supervise our F1-based token-selection reward.

Evaluation Benchmarks. To comprehensively assess the visual reasoning capabilities of our model, we conduct experiments across a diverse set of vision-centric benchmarks using the standardized VLMEvalKit Duan et al. (2024) framework, with `chatgpt-0125` serving as the judge model. We utilize V* Wu & Xie (2024) to evaluate the model's ability to localize and reason over fine-grained visual details in complex scenes. To assess geometric perception skills, including object counting, spatial relations, depth ordering, and distance estimation, we evaluate on CV-Bench Tong et al. (2024). Furthermore, we employ BLINK Fu et al. (2024) to test diverse visual perception and reasoning capabilities across a wide spectrum of visual tasks. HRBench Wang et al. (2025) is used to measure high-resolution image understanding at 4K and 8K resolutions.

4.2 QUANTITATIVE EVALUATION

Main Results. As shown in Table 1, we evaluate MIRA against closed-source models, VLM reasoning models Bai et al. (2025); Ming et al. (2025); Yang et al. (2025a), and latent reasoning models Li et al. (2026a); Qin et al. (2025); Wang et al. (2026b); Zhang et al. (2026b) across a diverse set of vision-centric benchmarks. Overall, MIRA outperforms the Qwen2.5-VL-7B baseline by 7.4 points on average. The largest improvements are observed on CV-Bench 3D, with notable gains on the Depth (+13.7) and Distance (+12.1) subtasks, resulting in an overall improvement of +12.8 points. Compared with VLM reasoning and latent reasoning models, MIRA achieves the highest overall performance, outperforming the second-best method, CrystaL, by 2.5 points.

Modality Selection Accuracy. To evaluate whether MIRA learns adaptive visual modality selection, we measure selection accuracy on a manually annotated test set of approximately 500 image-question pairs. For each example, humans annotate the visual modalities required to answer the question, providing an evaluation set independent of the Gemini-generated pseudo labels used during RL training. We compare MIRA with CoVT at a temperature of 0.7.

As shown in Table 2, MIRA consistently outperforms CoVT across all token types, with overall F1 improving from 0.64 to 0.81. The largest gains appear in DINO and Depth selection, where F1 increases from 0.50 to 0.73 and from 0.55 to 0.88, respectively. These results demonstrate that MIRA selects visual modalities that better align with human judgments of task-relevant evidence.

We further examine this adaptivity across BLINK tasks that emphasize different perceptual cues. As shown in Fig. 3, MIRA preferentially invokes Segmentation tokens for Counting, DINO tokens for Artstyle, and Depth tokens for Relative Depth. In contrast, CoVT exhibits less variation in modality usage across tasks. These patterns further support MIRA's ability to adapt visual modality selection to the perceptual demands of each task.

Table 1: Comparison with closed-source models, VLM reasoning models, and latent reasoning models on vision-centric benchmarks. Avg. is computed over V*, BLINK, CV-Bench (2D, 3D), and HRBench (4K, 8K). Results for our model are reported as the mean and standard deviation over three random seeds. The best result in each column is in **bold** and the second best is underlined.

Model	V*	BLINK	CV-Bench						HRBench		Avg.
			Count	Rel.	2D	Depth	Dist.	3D	4K	8K	
Closed-Source Models											
Claude-4-Sonnet	15.2	39.6	–	–	–	–	–	–	32.3	22.7	–
GPT-4o	45.9	**63.0**	–	–	–	–	–	–	50.6	46.7	–
Open-Source Base Model											
Qwen2.5-VL-7B	76.4	55.7	67.5	81.6	74.5	77.6	71.3	74.5	68.6	64.9	69.1
VLM Reasoning Models											
R1-OneVision-7B	60.7	48.6	63.7	72.4	68.0	65.2	68.0	66.6	62.3	53.6	60.0
Ocean-R1	78.0	57.2	71.2	83.4	77.3	83.8	81.8	82.5	72.4	68.4	72.6
Latent Reasoning Models											
LVR (ICLR'26)	80.6	52.5	67.2	84.9	75.2	83.2	74.5	78.8	69.6	66.1	70.5
Monet (CVPR'26)	**83.3**	53.1	68.2	79.0	73.6	75.0	71.3	73.2	71.0	68.0	70.4
CoVT (ECCV'26)	78.0	56.0	68.1	82.5	75.3	86.7	82.8	84.8	71.0	69.4	72.4
CrystaL (arXiv'26)	82.7	55.7	69.3	83.5	76.4	87.8	81.3	84.6	73.3	71.1	74.0
Ours	**83.3**	58.6	**75.9**	**86.2**	**81.1**	**91.3**	**83.4**	**87.3**	**76.8**	**71.7**	**76.5**
± Std.	±0.6	±0.7	±0.4	±0.2	±0.2	±1.2	±1.2	±1.0	±0.8	±0.4	±0.2

Table 2: Modality selection accuracy across visual token types, where Prec., Rec., and F1 denote Precision, Recall, and F1-score. MIRA notably outperforms CoVT across modalities.

Token Type	CoVT			Ours		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Segmentation	0.65	**0.98**	0.78	**0.71**	**0.98**	**0.83**
DINO	0.57	0.45	0.50	**0.74**	**0.72**	**0.73**
Depth	0.41	0.85	0.55	**0.82**	**0.96**	**0.88**
Total	0.55	0.78	0.64	**0.75**	**0.89**	**0.81**

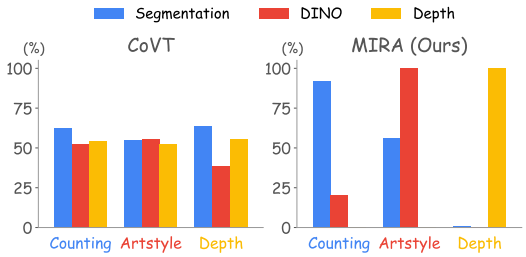


Figure 3: Visual token distribution across question types on BLINK. MIRA adaptively predicts task-relevant visual tokens, whereas CoVT exhibits a near-uniform distribution.

4.3 Qualitative Results

We evaluate MIRA's reasoning process through representative examples to illustrate its adaptive and flexible nature. Our analysis focuses on two advantages: adaptive latent token invocation and superior interpretability compared to template-based approaches. For visualization, we denote 8 consecutive <|sam_pad|> tokens, 4 <|depth_pad|> tokens, and 4 <|dino_pad|> tokens as <segmentation>, <depth>, and <dino>, respectively.

Fig. 4 shows how MIRA adaptively integrates latent visual tokens into its reasoning process. In the left example, MIRA invokes segmentation tokens to ground the skier's pose, mitigating visual hallucinations. The center example demonstrates spatial reasoning by activating depth tokens to locate points A and B and make a physically grounded proximity judgment rather than relying on ambiguous 2D cues. In the right example, MIRA first uses segmentation tokens to locate the subjects, then DINO tokens to extract fine-grained features for accurate object identification.

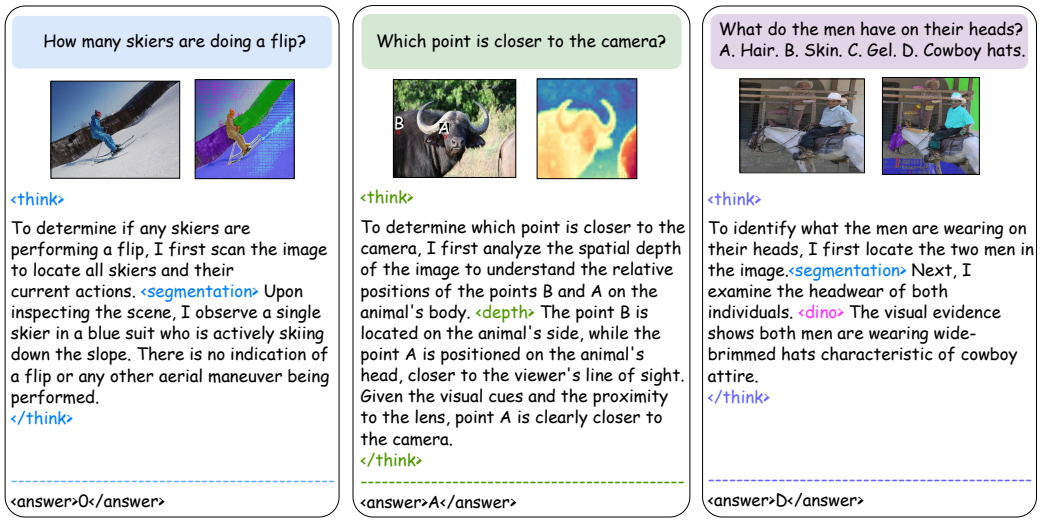


Figure 4: Qualitative examples of MIRA's adaptive latent token invocation across three tasks: object counting (left), depth estimation (center), and fine-grained object identification (right).

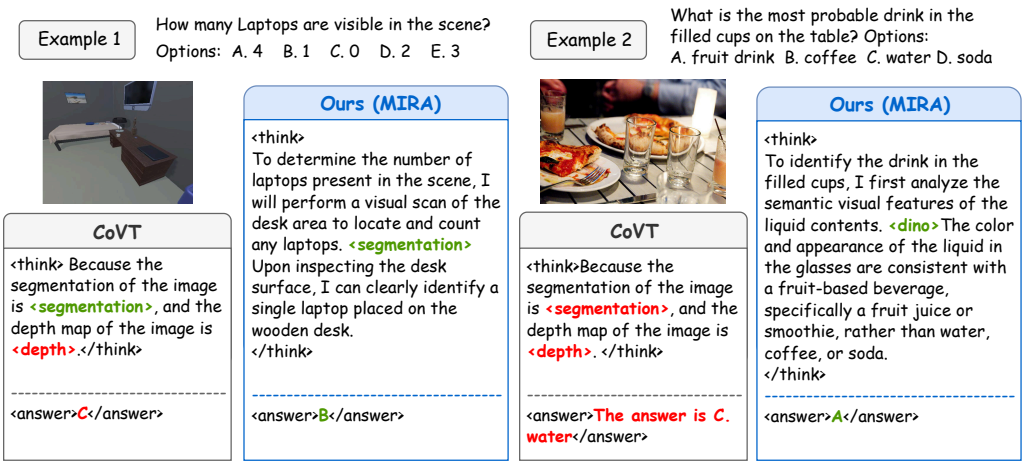


Figure 5: Comparison between CoVT and our proposed MIRA on counting task and semantic recognition task, highlighting MIRA's ability to adaptively select the most relevant visual tokens for different tasks instead of relying on rigid reasoning template.

Fig. 5 illustrates the structural advantage of MIRA over CoVT. Rather than forcing a fixed sequence of token invocations regardless of context, MIRA dynamically activates only the necessary latent tokens on demand. As shown in Fig. 5 (left), MIRA accurately counts the laptops by invoking only segmentation tokens, while CoVT's rigid inclusion of depth tokens leads to a counting failure. In the right example, MIRA correctly identifies the drink via DINO tokens for semantic analysis, whereas CoVT's fixed pipeline defaults to an incorrect answer. These results highlight the importance of adaptive token selection for robust multimodal reasoning.

4.4 Ablation Study

Component Ablation. In Table 3, we ablate the token-selection reward and alignment regularization in IRPO. All RL variants outperform the Stage-1 SFT baseline (74.6), supporting the benefit of policy optimization. Starting from standard GRPO (75.6), adding alignment regularization improves the average score to 76.0, while adding the token-selection reward yields 76.1. Combining

Table 3: Results for Stage-2 variants are reported as mean ± standard deviation over three random seeds, all initialized from the same Stage-1 checkpoint. The Stage-1 SFT baseline is reported from a single run.

	Ablation				CV-Bench		HRBench		
Model	Rew.	Align.	V*	Blink	2D	3D	4k	8k	**Avg.**
Stage 1: Interleaved SFT									
Interleaved SFT Only	–	–	82.7	55.8	78.7	84.6	75.5	70.4	74.6
Stage 2: IRPO									
RL Variants	✗	✗	82.9 ± 1.2	57.6 ± 0.1	80.9 ± 0.1	87.1 ± 0.9	74.8 ± 0.8	70.5 ± 0.5	75.6 ± 0.1
	✗	✓	83.1 ± 0.4	57.9 ± 0.8	**81.1 ± 0.3**	87.1 ± 0.6	74.9 ± 0.5	**72.1 ± 0.2**	76.0 ± 0.1
	✓	✗	82.2 ± 0.9	58.0 ± 0.8	**81.1 ± 0.1**	**87.3 ± 0.4**	76.3 ± 0.9	72.0 ± 0.4	76.1 ± 0.2
Our Method	✓	✓	**83.3 ± 0.6**	**58.6 ± 0.7**	**81.1 ± 0.2**	**87.3 ± 1.0**	**76.8 ± 0.8**	71.7 ± 0.4	**76.5 ± 0.2**

Table 4: Effect of the token-selection reward on modality selection. Precision, recall, and F1 scores are measured against human annotations.

Token Type	w/o Token Reward			w/ Token Reward		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Segmentation (SAM)	**0.71**	0.88	0.79	**0.71**	**0.98**	**0.83**
DINO	**0.94**	0.33	0.49	0.74	**0.72**	**0.73**
Depth	0.54	0.85	0.66	**0.82**	**0.96**	**0.88**
Total (Micro)	0.67	0.70	0.69	**0.75**	**0.89**	**0.81**

both components achieves the highest average score of 76.5, indicating that modality-selection supervision and visual alignment regularization provide complementary benefits.

Effect of Token-Selection Reward on Modality Selection. Standard outcome-based GRPO primarily supervises final-answer correctness and provides limited guidance on intermediate modality-selection decisions. Our token-selection reward explicitly supervises this intermediate perceptual process, encouraging the model to select appropriate visual modalities. To evaluate its effect, we compare modality-selection performance with and without the reward against human annotations, keeping alignment regularization enabled in both variants.

As shown in Table 4, introducing the token-selection reward improves the overall micro-F1 against human annotations from 0.69 to 0.81. The largest gains are observed for DINO (0.49→0.73) and Depth (0.66→0.88), demonstrating more accurate selection of visual modalities relevant to the question, as assessed against human annotations. Together with the downstream improvement in Table 3, these results suggest that explicitly supervising intermediate perceptual decisions benefits both modality-selection accuracy and visual reasoning performance. This supports the value of guiding the model's intermediate modality choices in addition to rewarding final-answer correctness. Qualitative examples visualizing the effect of the token-selection reward are provided in Appendix C.3.

5 Conclusion

We present MIRA, a framework for adaptive interleaved reasoning that augments vision-language models with grounded visual latent tokens. Through Interleaved Supervised Fine-Tuning followed by Interleaved Reasoning Policy Optimization (IRPO), MIRA learns to adaptively invoke task-relevant visual modalities. IRPO combines an F1-based token-selection reward with visual alignment regularization to guide modality selection while preserving perceptual fidelity. Extensive experiments show that MIRA outperforms strong baselines on diverse vision-centric benchmarks, achieving an average improvement of 7.4 percentage points over Qwen2.5-VL-7B while producing interpretable reasoning traces that integrate text with visually grounded latent representations.

AI USE STATEMENT

We used Gemini 3 Flash to synthesize interleaved reasoning traces for supervised fine-tuning and to generate pseudo labels for visual modality selection during reinforcement learning, as described in Section 3. We also used an LLM-based judge through VLMEvalKit for benchmark evaluation, as described in the experimental setup. Claude assisted with implementing parts of the method designed by the authors, and the generated code was manually reviewed and tested by the authors. Additionally, we used ChatGPT for language editing. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work uses publicly available vision-language benchmarks and does not involve the collection of sensitive personal data or experiments with human subjects. We use generative models to synthesize interleaved reasoning traces and pseudo labels for visual modality selection, which may inherit biases or errors from the underlying models. To mitigate this issue, we apply filtering to the generated data and evaluate the quality of the pseudo labels against human annotations. Our method is intended for research on multimodal reasoning, and we do not make claims regarding its suitability for high-stakes or safety-critical applications.

REPRODUCIBILITY STATEMENT

We provide the prompts used for data generation in Appendix A.1, as well as the training configurations, hyperparameters, and implementation details for reproducing our experiments in Section 4.1. We also plan to publicly release the source code upon acceptance to further support reproducibility.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Chao Chen, Zhixin Ma, Yongqi Li, Yupeng Hu, Yinwei Wei, Wenjie Li, and Liqiang Nie. Reasoning in the dark: Interleaved vision-text reasoning in latent space. *arXiv preprint arXiv:2510.12603*, 2025a.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or RL? an early investigation into training r1-like reasoning large vision-language models. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL https://openreview.net/forum?id=wZI5qkQeDF.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025c.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025d.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Eric Wang. GRIT: Teaching MLLMs to think with images. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*, 2025. URL https://openreview.net/forum?id=Wb2jI9PSg2.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=GNSMl1P5VR.

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *arXiv preprint arXiv:2602.04476*, 2026.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=skLtdUVaJa. Featured Certification, Outstanding Certification.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=j84WR5ORsC.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Lee Boyd-Graber, Haitao Mi, and Dong Yu. Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=C1M4ETatgM.

Chengzhi Liu, Yuzhe Yang, Yue Fan, Qingyue Wei, Sheng Liu, and Xin Eric Wang. Reasoning within the mind: Dynamic multimodal interleaving in latent space. *arXiv preprint arXiv:2512.12623*, 2025a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods. *arXiv preprint arXiv:2511.15722*, 2025b.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4122–4134, 2025c.

Lingfeng Ming, Yadong Li, Song Chen, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*, 2025.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. 2024.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

Arijit Ray, Jiafei Duan, Ellis L Brown II, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=DW8U8ZWa1U.

Gabriel Herbert Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=1amnhVRQ3l.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=aXeiCbMFFJ.

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*, 2026. URL https://openreview.net/forum?id=VeZkY3JjWV.

Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5117–5127, 2021.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*, 2025.

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Vi8AepAXGy.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani zheng, Sule Bai, Zijian Kang, Jiashi Feng, Wang Zhuochen, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=zwMpimSZEq.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language. In *CVPR*, 2026b.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru WANG, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=izbBqTL8vb.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, ChengXiang Zhai, and Klara Nahrstedt. VTool-r1: VLMs learn to think with images via reinforcement learning on multimodal tool use. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Idst6X6gmy.

Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*, 2025.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cFTi3gLJ1X.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2376–2385, 2025a.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=L2W4wQsNkY.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025b.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026a. URL https://openreview.net/forum?id=GzgPleFl8f.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

Yang Zhang, Danyang Li, Yuxuan Li, Xin Zhang, Tianyu Xie, Mingming Cheng, and Xiang Li. Crystal: Spontaneous emergence of visual latents in mllms. *arXiv preprint arXiv:2602.20980*, 2026b.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xUyMXkI958.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A INTERLEAVED SFT DATASET CONSTRUCTION

A.1 PROMPT DESIGN FOR INTERLEAVED SFT DATASET

We present the prompt templates used for the reasoning trace collection of the Visual CoT and SPAR datasets below. Note that for readability within the prompt examples, the placeholders `[segmentation]`, `[depth]`, and `[dino]` are used to denote the full sequences of 8 <|sam_pad|>, 4 <|depth_pad|>, and 4 <|dino_pad|> tokens, respectively.

Visual CoT Prompt The following prompt (`VISUAL_COT_PROMPT`) was used for the Visual CoT dataset to elicit latent visual reasoning.

VISUAL_COT_PROMPT

```
You are given an image-based or diagram-based question.

Your task is to generate:
1. Textual reasoning that integrates visual perception.
2. Global visual perception tokens (latent visual reasoning).
3. A short final answer.

Rules:
- Use <think> for the reasoning.
- **MINIMAL VISUAL TOKEN USAGE:** Use visual tokens SPARINGLY.
- **ADAPTIVE TOKEN USAGE:**
  - **Simple Tasks:** Use EXACTLY ONE type of visual token (e.g., Counting -> SAM).
  - **Complex Tasks:** If the question involves Multi-step Reasoning
    (e.g., "Shape of the object behind X"), you may use MULTIPLE distinct token
    blocks (e.g., Depth + SAM).
- **GLOBAL SCOPE:** Visual tokens represent the **WHOLE IMAGE** features.
- **NO REDUNDANCY:** Do not stack tokens that provide the same information. Only
combine tokens if they offer orthogonal information (e.g., SAM + Depth).
- **Fluency:** Integrate tokens naturally into the sentence structure.
- **MANDATORY TAGS:** Every response MUST start with <think> and conclude the final
answer inside <answer>. These tags are strictly required for structural parsing and
must not be omitted.
- **STRICT SPACING:** Between </think> and <answer>, there must be exactly two newlines
 (\\n\\n) and NO other characters, symbols, text, or horizontal rules.
- **TEXT-ONLY / TABLE / CHART EXCLUSION:** If the input is a text-based table, chart or
 purely textual information, do NOT use any visual tokens (SAM, Depth, or Dino). Visual
 tokens are only for image/diagram analysis.
Visual Token Selection Guide (Choose the Best Match):

1. **<|sam_pad|> (8 tokens) - [Counting / Locating objects]**
   - **Use Case:** Counting, Locating objects.

2. **<|depth_pad|> (4 tokens) - [Depth Only / Distance / Relative Position (front/
behind)]**
   - **Use Case:** Questions PURELY about distance, relative position (front/behind).

3. **<|dino_pad|> (4 tokens) - [Semantics / Category / Material Identification /
Reasoning]**
   - **Use Case:** Questions PURELY about category/material identification or reasoning
.

Example 1 (Simple Counting -> SAM):
<think>
To count the vehicles, I scan the scene to locate all transport items. [segmentation]
I see a yellow dirt bike and a grey motorcycle. Total 2.
</think>

<answer>2</answer>

Example 2 (Simple Distance -> Depth):
<think>
The question asks which of the two marked points is closer to the camera. I consult the
 global depth map. [depth] The depth cues show point A is in the foreground, while
point B is in the background.
</think>

<answer>A</answer>

Example 3 (Simple Category -> Dino):
<think>
```

```
To identify the material of the large block, I analyze the semantic texture features. [
dino] The features indicate a metallic surface, not rubber.
</think>

<answer>Metal</answer>

Example 4 (Object Distance Comparison -> SAM + Depth):
<think>
The question asks which object is closer, the red sphere or the blue cube. I locate
both objects. [segmentation] The red sphere is on the left and the blue cube is on the
right. I check their depths. [depth] The red sphere is in the foreground, while the
blue cube is in the background.
</think>

<answer>Red Sphere</answer>

Question: {question}
Answer Hint: {answer}
```

SPAR Prompt The following prompt (SPAR_PROMPT) was used for the SPAR dataset to elicit latent visual reasoning.

SPAR_PROMPT

```
You are given a spatial perception and reasoning question based on an indoor scene
image.

Your task is to generate:
1. Textual reasoning that integrates visual perception for spatial understanding.
2. Global visual perception tokens (latent visual reasoning) - use **TWO or MORE**
types for spatial tasks.
3. A short final answer.

Rules:
- Use <think> for the reasoning.
- **ADAPTIVE TOKEN USAGE:**
    - **Simple spatial tasks**(e.g.,  a straightforward depth or distance query about a
 single
    object): You MAY use **only ONE** visual token type if sufficient.
    - **Complex spatial tasks** (e.g., relative position between objects, BEV
coordinates,
    multi-object distance comparison): You SHOULD combine **at least TWO** visual token
 types
    (e.g., SAM + Depth, or Dino + Depth).
- **GLOBAL SCOPE:** Visual tokens represent the **WHOLE IMAGE** features.
- **Fluency:** Integrate tokens naturally into the sentence structure.
- **MANDATORY TAGS:** Every response MUST start with <think> and conclude the final
answer inside <answer>. These tags are strictly required for structural parsing and
must not be omitted.
- **STRICT SPACING:** Between </think> and <answer>, there must be exactly two newlines
 (\\n\\n) and NO other characters, symbols, text, or horizontal rules.

Visual Token Selection Guide (Combine for Spatial Reasoning):

1. **<|sam_pad|> (8 tokens) - [Locating / Segmenting]**
   - Locate objects and identify object boundaries and positions in 2D.

2. **<|depth_pad|> (4 tokens) - [Depth / Distance / 3D Spatial Position]**
   - Estimate distances, relative depth (front/behind/closer/farther), 3D spatial
relationships.

3. **<|dino_pad|> (4 tokens) - [Object Recognition / Semantic Identification /
Reasoning]**
   - Identify object categories, recognize semantic features of objects, or
   general visual reasoning/understanding.

Example 1 (Simple Depth Query -> Depth only):
<think>
The question asks for the dimensions of a single object.
I analyze the depth information to estimate the object's size in the scene. [depth]
Based on the depth cues, the pillow measures approximately 120 cm in length, 60 cm in
width, and 30 cm in height.
</think>
```

```
<answer>The dimension set for pillow is 120 cm (length), 60 cm (width), and 30 cm (
height).</answer>

Example 2 (Relative Position -> SAM + Depth):
<think>
The question asks about the relative position between two objects from the observer's
perspective.
I first locate both objects in the scene. [segmentation]
The red bbox object is on the left side and the blue bbox object is on the right side
of the image.
Now I analyze depth to determine their front/behind relationship. [depth]
The depth cues indicate the red object is closer to the observer at about 1.5 meters,
while the blue object is farther at 2.3 meters.
</think>

<answer>The red object is to the left and closer to the observer than the blue object.
It is perceived as closer.</answer>

Example 3 (Viewpoint Change Analysis -> SAM + Dino + Depth):
<think>
The question asks how the spatial relationship of an object changes when the observer
moves to a new position.I first locate all marked objects and their spatial layout. [
segmentation] The red bbox (treadmill) is on the left side, the green bbox (target
position) is center, and blue bbox (facing direction) is on the right.I identify each
object to understand the scene semantics. [dino] The features confirm: red=treadmill,
green=chair, blue=tv in an exercise room. Now I analyze depth to reason about the
spatial transformation after moving. [depth] From the original viewpoint, the treadmill
 is to the left and in front. After moving to the chair and facing the tv, the
treadmill would shift to the right and remain in front.
</think>

<answer>Before moving, treadmill is to the left of the observer and in front. After
moving to chair and facing tv, treadmill becomes to the right and remains in front.</
answer>

Question: {question}
Answer Hint: {answer}
```

A.2 Data Post-processing

To ensure the structural integrity of our generated dataset, we implemented a strict rule-based post-processing mechanism. This automated stage discarded responses with improper formatting. Specifically, it filtered out instances that missed the mandatory <think> or <answer> tags, contained incorrect spacing, generated an incorrect number of latent visual tokens (e.g., failing to output exactly 8 tokens for <|sam_pad|>, or 4 tokens for <|dino_pad|> and <|depth_pad|>), or violated other basic parsing constraints.

B Gemini Pseudo-label Quality Verification

Table 5: Agreement with human annotations for modality selection. We compare a simple keyword-based classifier with the Gemini-generated pseudo labels used to supervise the token-selection reward.

	Keyword Baseline			Gemini Pseudo Labels		
Token Type	Prec.	Rec.	F1	Prec.	Rec.	F1
Segmentation (SAM)	0.81	0.67	0.73	**0.88**	**0.98**	**0.93**
DINO	0.57	0.50	0.53	**0.83**	**0.99**	**0.90**
Depth	0.65	0.74	0.69	**0.82**	**0.98**	**0.89**
Total (Micro)	0.69	0.63	0.66	**0.85**	**0.99**	**0.91**

Evaluation Setup. To assess the reliability of the Gemini-generated pseudo labels used to supervise the token-selection reward, we manually annotate the task-relevant visual modalities for approximately 500 image-question pairs. These human annotations provide an independent reference for evaluating the quality of the generated modality labels. We additionally compare Gemini against a simple keyword-based classifier, which predicts the required visual modalities based on lexical cues in the question. Table 5 reports the precision, recall, and F1 scores with respect to the human annotations.

Results. The results show that the pseudo labels generated by Gemini align closely with human annotations of task-relevant visual modalities, achieving a micro-F1 score of 0.91, compared with 0.66 for the keyword-based classifier. This gap indicates that identifying relevant visual modalities requires more than simply matching lexical cues in the question. The difference is especially clear for DINO, with the keyword baseline reaching an F1 score of only 0.53, compared with 0.90 for Gemini. This suggests that the relevance of DINO is often implicit in the task rather than signaled by explicit keywords. Overall, these results indicate that the pseudo labels capture task-dependent information that is missed by simple keyword-based routing, making them a suitable source of supervision for adaptive modality selection in our framework.

C Additional Experiment Results

C.1 Controlled Comparison with CoVT

Table 6: Controlled comparison between CoVT and MIRA after Stage-1 SFT. Both methods use the same SFT data, training budget, and evaluation protocol. Avg. is computed over all six benchmarks.

Method	V*	BLINK	CV-Bench		HRBench		Avg.
			2D	3D	4K	8K	
CoVT	78.5	**56.0**	76.1	81.6	72.5	**71.1**	72.6
MIRA (Stage 1)	**82.7**	55.8	**78.7**	**84.6**	**75.5**	70.4	**74.6**

Evaluation Setup. We compare MIRA with CoVT under the same Stage-1 training setting. Both methods use the same image-question pairs and ground-truth answers, with the same training budget and evaluation protocol. CoVT follows its fixed visual-token pipeline, whereas MIRA performs adaptive interleaved visual reasoning. Table 6 reports the performance of both methods after Stage-1 SFT.

Results. As shown in Table 6, MIRA improves the average score from 72.6 to 74.6 under the matched Stage-1 setting, with notable gains on V*, CV-Bench, and HRBench 4K. This suggests that adaptive interleaved visual reasoning provides a more effective reasoning paradigm than CoVT's fixed visual-token pipeline under the same SFT conditions.

C.2 Robustness to Pseudo-Label Noise

Although the Gemini-generated modality labels show strong agreement with human annotations, they may still contain errors. To evaluate the sensitivity of IRPO to noisy supervision, we randomly flip binary modality labels to obtain approximately 0.75 and 0.50 F1 agreement with the original Gemini labels. Each setting is evaluated using a single training run with the same random seed and otherwise identical experimental settings. As shown in Table 7, the average performance decreases slightly from 76.4 to 76.2 under moderate corruption and to 75.4 under stronger corruption. These results suggest that MIRA tolerates moderate pseudo-label noise in this controlled setting, while more severe corruption leads to a larger performance decline.

C.3 Visualization of Token Reward Effect.

To illustrate how the F1-based token-selection reward affects modality invocation, we compare reasoning traces with and without the reward in Fig. 6 and Fig. 7. Without it, the model relies on

Table 7: Robustness to pseudo-label noise under the same-seed setting. We randomly corrupt the Gemini-generated modality labels to simulate different levels of supervision noise. Label F1 denotes the agreement between the corrupted labels and the original Gemini labels.

Pseudo-Label Setting	**Avg.**	**Δ vs. Original**
Original Gemini labels	**76.4**	–
Corrupted labels (F1 $\approx$ 0.75)	76.2	-0.2
Corrupted labels (F1 $\approx$ 0.50)	75.4	-1.0

segmentation tokens, detecting bicycles but missing fine-grained semantics and producing incorrect predictions. With the reward, it selects DINO tokens to capture semantic details (e.g., pedicabs), yielding correct answers. The reward also discourages unnecessary visual token usage. As shown in Fig. 7 in the Appendix, the model relies solely on depth tokens when sufficient for spatial reasoning. Without the reward, it tends to invoke redundant visual tokens, making the reasoning trace less interpretable.

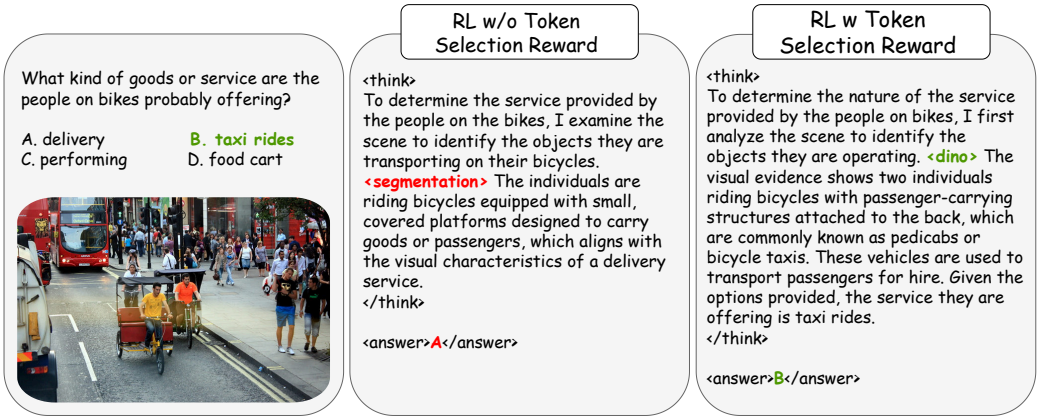


Figure 6: Without the token-selection reward, the model misapplies segmentation tokens to a fine-grained recognition task, leading to an incorrect answer. With the token reward, the model correctly selects DINO tokens, producing accurate reasoning.

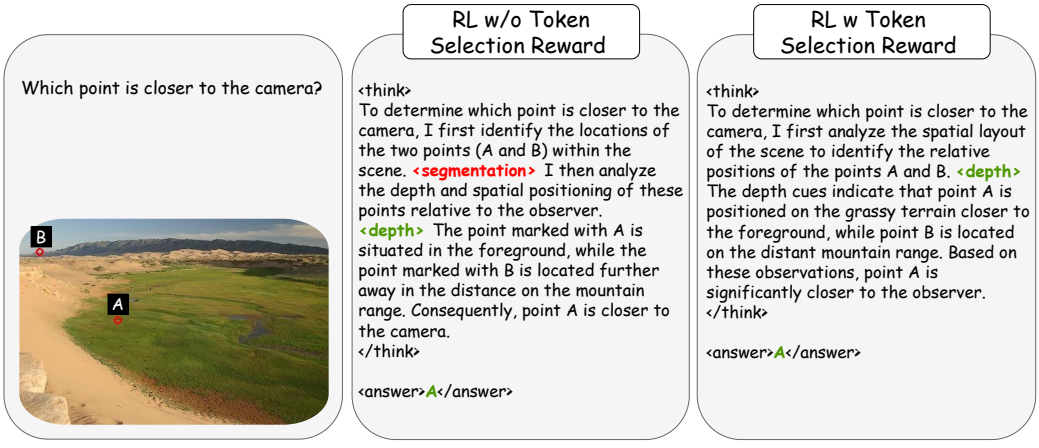


Figure 7: The F1-based token-selection reward encourages the adaptive use of visual tokens. With the reward, the model learns to adaptively select depth tokens for depth estimation tasks, rather than redundantly invoking segmentation tokens.

C.4 Inference Efficiency

Table 8: Average end-to-end inference time on 16 randomly sampled V* examples under the same experimental setup.

Method	Avg. Inference Time (s)
CoVT	**1.75**
PixelReasoner	4.10
MIRA	<u>2.96</u>

Evaluation Setup. We further compare the inference efficiency of MIRA with CoVT and PixelReasoner. We randomly sample 16 questions from the V* benchmark and measure the end-to-end inference time of each method under the same experimental setup.

Results. As shown in Table 8, MIRA averages 2.96 seconds per question, compared with 1.75 seconds for CoVT and 4.10 seconds for PixelReasoner. MIRA achieves lower latency than PixelReasoner, while offering stronger overall benchmark performance than CoVT at the cost of additional inference time. These results highlight a trade-off between inference efficiency and reasoning performance.

D Limitations and Future Work

Limitations. Our framework operates with a predefined set of visual latent tokens (e.g., segmentation, DINO, and depth). While the design naturally extends to additional vision foundation models for new perceptual modalities (e.g., optical flow or surface normals), incorporating each requires retraining the VLM with the alignment objective. Scaling to a broader set of modalities may also benefit from larger and more diverse training data to sufficiently cover the expanded perceptual space.

Future Work. Beyond expanding perceptual modalities, an interesting direction is to incorporate world models as an additional latent reasoning module, enabling the model to reason not only about static perceptual cues but also about possible future states and dynamics.

E FURTHER RESULTS

E.1 FURTHER QUALITATIVE EVALUATION

For better visualization, we replace each sequence of 8 consecutive <|sam_pad|> tokens, 4 <|depth_pad|> tokens, and 4 <|dino_pad|> tokens with the shorthand <segmentation>, <depth>, and <dino>, respectively.

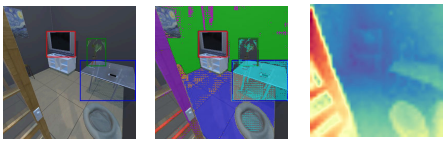
How many people are wearing suit?



<think>
To determine how many people are wearing a suit, I first identify the individuals in the image. <segmentation> Next, I analyze the clothing of each person to categorize their attire. <dino> The man on the left is wearing a formal suit jacket, while the woman on the right is wearing a ski suit. Therefore, only one person is wearing a suit.
</think>

<answer>1</answer>

Considering the relative positions, where is blue modern dining table (highlighted by a blue box) with respect to Television (highlighted by a red box)? Options: A. below B. above



<think>
To determine the relative position of the blue modern dining table (blue box) with respect to the television (red box), I first analyze their spatial arrangement in the scene. <segmentation> The television is positioned on a white stand against the back wall. The dining table is situated to the right of the television, extending into the foreground. By assessing the depth and vertical alignment, I can see that the dining table surface is lower than the television stand level. <depth> Given the visual evidence, the dining table is positioned below the television.
</think>

<answer>A. below</answer>

How many bicycles are visible in the image?



<think>
To determine the number of bicycles in the image, I will scan the scene to locate and count the relevant objects. <segmentation> The visual analysis confirms there is only one bicycle present in the scene.
</think>

<answer>1</answer>

Transparent umbrella is used only from protecting? Options: A. rain B. uv C. sun D. wind



<think>
To determine the primary function of the transparent umbrella, I analyze the scene context and the environmental conditions. <dino> The umbrella is being used in a rainy urban environment, with water droplets visible on the transparent surface, confirming its purpose is to protect against rain.
</think>

<answer>A</answer>

Figure 8: More Qualitative Results of Our Model (MIRA)