Sketch-in-Latents: ELICITING UNIFIED REASONING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) excel at text-based visual understanding but fall short on tasks requiring visual imagination. Existing approaches rely on predefined external tools or in-line image generation, whereas humans flexibly interleave visual and textual imagination within a unified mental space. Since MLLMs already encode visual and textual information in a shared feature space, we argue that visual tokens can be seamlessly inserted into text-token reasoning, allowing visual imagination to be carried entirely by latent features. To this end, we propose Sketch-in-Latents (SkiLa), a unified reasoning paradigm that extends MLLMs' auto-regressive capability to natively generate continuous visual embeddings, termed latent sketch tokens, as visual thoughts, dynamically alternating between a textual thinking mode and a visual sketching mode. SkiLa is trained in two stages: SFT *explicitly* grounds latent sketch tokens via a latent visual semantics reconstruction mechanism, while RL with masked GRPO on our curated SkiLa-RL-4K further *implicitly* refines them. Extensive experiments show SkiLa achieves superior performance on vision-centric tasks and generalizes well across diverse multimodal benchmarks. Code will be released.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Bai et al., 2025b; Liu et al., 2023; Li et al., 2023; Wang et al., 2025c) have made rapid progress by coupling visual encoders with large language models (LLMs) (Touvron et al., 2023; Chiang et al., 2023; Bai et al., 2023). Inspired by Chain-of-Thought (CoT) reasoning (Wei et al., 2022), recent multimodal methods (Zhang et al., 2023; Huang et al., 2025; Liu et al., 2025b) decompose complex image-referenced queries into step-by-step reasoning traces. However, most existing MLLMs still reason mainly in the linguistic domain (Wang et al., 2025a; Yu et al., 2025): even when images are involved, their intermediate reasoning is typically produced as sequential text tokens. This text-centric bottleneck forces models to verbalize problems that humans would solve intuitively through mental visualization (Farah, 1985; Kosslyn, 1996). For tasks demanding rich visual-spatial intuition (e.g., jigsaw, mazing) (Fu et al., 2024), this limitation is especially pronounced.

Recent work explores two main directions to move beyond purely textual reasoning. Tool-driven methods invoke external utilities or generate code (e.g., cropping, drawing auxiliary lines) to inspect and manipulate images (Wu & Xie; Shao et al., 2024a; Zheng et al., 2025; Zhang et al., 2025a; Liu et al., 2025a; Hu et al., 2024; Zhao et al., 2025), but their reasoning is bounded by predefined operations or fixed execution criteria with limited action spaces. Unified foundation models jointly generate text and images (Team, 2024; Chen et al., 2025; Deng et al., 2025), but pixel-level generation is computationally expensive, limited by generative fidelity, and has mainly been validated on specialized tasks such as navigation and maze solving (Li et al., 2025c; Xu et al., 2025).

Unlike these approaches, a key reason for humans' unlimited visual-text imagination is that this interaction occurs within an internal, unified mental space (Lake et al., 2017). A natural question arises: *Can MLLMs transcend their linguistic reasoning framework to generate visual thoughts natively as part of an internal, unified reasoning process?* Since MLLMs already encode images and text into the same feature space, we argue that **visual tokens can be seamlessly inserted into text-token reasoning**, forming an internal unified reasoning process. This brings two advantages: (1) arbitrary visual imaginations can be represented by token features, enabling unlimited visual-thinking

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.

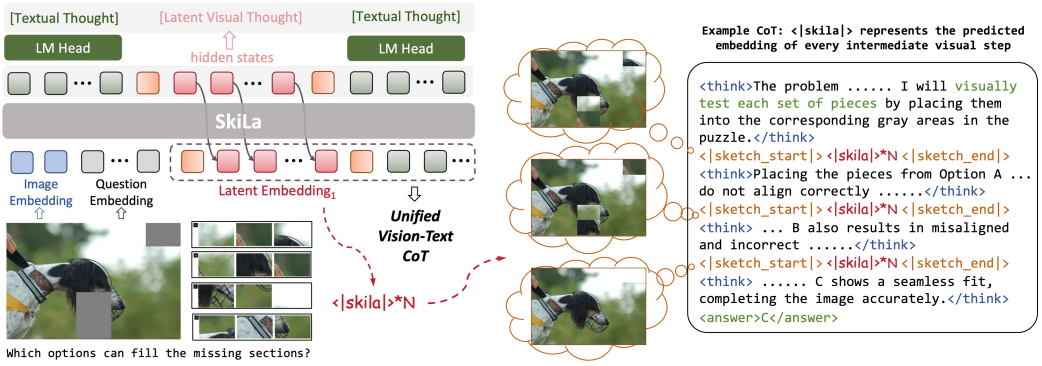


Figure 1: SkiLa enables MLLMs to reason in a unified manner via a hybrid auto-regressive process, producing multi-step, interleaved multi-modal reasoning traces without predefined visual-text actions. It dynamically alternates between a textual thinking mode that generates textual thoughts and a visual sketching mode that generates latent sketch tokens as visual thoughts.

manners; (2) the model can naturally produce interleaved visual-text reasoning traces, yielding more flexible, human-like thinking.

Based on these insights, we propose **Sketch-in-Latents (SkiLa)**, a unified multimodal reasoning paradigm that extends MLLMs' auto-regressive capability to natively generate continuous visual embeddings, termed *latent sketch tokens*, as visual thoughts. As shown in Fig. 1, SkiLa enables MLLMs to reason in a unified manner via a hybrid auto-regressive process producing multi-step, interleaved multimodal reasoning traces. It dynamically alternates between: (1) *textual thinking mode*, standard next-token prediction on discrete text tokens for textual thoughts, and (2) *visual sketching mode*, continuous latent sketch token generation for visual thoughts. To semantically ground the latent sketch tokens, we introduce a latent visual semantics reconstruction mechanism that ***explicitly aligns*** them with the visual semantics of intermediate "sketch images" from interleaved text–image reasoning traces, while textual thoughts are still learned via standard next-token prediction.

On top of this, a reinforcement learning stage further treats latent sketching as an internal reasoning action rather than an imitation target: instead of imposing direct supervision, it further ***implicitly refines*** the latent sketch tokens by rewarding the textual reasoning and final answers conditioned on them. Concretely, we apply a masked variant of GRPO (Shao et al., 2024b) that excludes the continuous sketch states from the importance ratio, and curate **SkiLa-RL-4K**, a 4K-sample dataset filtered to retain prompts with non-trivial reward variance, letting the model freely explore useful latent visual states beyond the SFT reconstruction targets.

Extensive experiments confirm that SkiLa acquires an intrinsic unified reasoning capability: it can begin reasoning linguistically, seamlessly switch to visual sketching for in-latent simulation, then revert to text generation. This establishes a flexible paradigm for multi-step unified reasoning where visual cognition and linguistic thought are deeply integrated. Our contributions are as follows:

- We propose Sketch-in-Latents (SkiLa), a novel unified reasoning paradigm that enables a hybrid auto-regressive process, allowing MLLMs to flexibly and seamlessly interleave multi-step explicit textual thoughts and latent visual thoughts without pre-defined external actions.
- We design a two-stage training paradigm: an SFT stage that *explicitly* grounds sketch tokens in visual semantics via our latent reconstruction mechanism, and an RL stage with masked GRPO on our curated SkiLa-RL-4K *implicitly* refines sketch tokens by rewarding the textual reasoning and final answers conditioned on these tokens.
- Extensive experiments show that SkiLa achieves state-of-the-art results on vision-centric tasks and generalizes well across diverse general multimodal benchmarks.

2 Related Works

Multi-modal CoT. Chain-of-Thought (CoT) (Wei et al., 2022) improves LLM reasoning by decomposing complex problems into intermediate steps. Recent works (Wang et al., 2025a; Yu et al.,

2025; Yan et al., 2025; Hong et al., 2025a; Li et al., 2025d) extend this paradigm to multimodal settings, with methods such as Vision-R1 (Huang et al., 2025) and Visual-RFT (Liu et al., 2025b) training MLLMs for step-by-step reasoning. However, their reasoning verbalizes visual concepts in the linguistic domain, which hinders performance on vision-intensive tasks.

Existing "think with images" efforts fall into two categories. Tool-driven methods (Wu & Xie; Shao et al., 2024a; Zhao et al., 2025) equip MLLMs with predefined utilities for cropping, drawing, or marking regions. For instance, Visual Sketchpad (Hu et al., 2024) generates code for auxiliary visual aids, and DeepEyes (Zheng et al., 2025) uses zoom-in tools to crop images. Despite being effective, they rely on predefined toolkits with fixed action spaces, limiting flexibility and failing to cultivate an intrinsic reasoning faculty. The second direction pursues unified foundation models such as Janus-Pro (Chen et al., 2025) and Bagel (Deng et al., 2025), which interleave text and image generation within reasoning chains. However, this paradigm suffers from the high cost of pixel-level synthesis, and its understanding is bounded by generative fidelity, so models cannot reason about what they cannot draw. Their effectiveness has also been shown mainly in specific domains such as navigation and maze solving (Xu et al., 2025; Li et al., 2025c), raising concerns about generalizability. In contrast, SkiLa natively generates intrinsic visual thoughts, eliminating dependence on external tools and costly pixel-level generation.

Latent Reasoning. Several studies perform reasoning in a continuous embedding space rather than discrete token space (Hao et al., 2024; Cheng & Van Durme, 2024; Shen et al., 2025; Zhang et al., 2025b) to improve LLM reasoning efficiency. Recent work extends latent reasoning to MLLMs: Mirage (Yang et al., 2025) uses hidden states trained to approximate annotated auxiliary visual information, while LVR (Li et al., 2025b) reasons over reconstructed image crops in latent space. However, these methods have two limitations. First, they are tailored to specific scenarios: Mirage trains separate models on different datasets for distinct settings, and LVR focuses only on cropped regions, constraining generalization beyond predefined contexts. Second, their reasoning is typically confined to a single inferential step, limiting applicability to complex, multi-step tasks. In contrast, our paradigm enables a multi-step, interleaved reasoning trace while remaining generalizable across diverse multimodal tasks.

3 METHODS

As shown in Figure 2, we introduce Sketch-in-Latents (SkiLa), a novel reasoning paradigm that enables MLLMs to jointly reason over discrete text and continuous visual representations within a unified auto-regressive process.

3.1 SKILA: SKETCH-IN-LATENTS

Model Architecture. SkiLa builds on a standard MLLM framework with three components: a vision encoder E_v that maps input images $\mathbf{I}$ to visual tokens $\mathbf{V} = E_v(\mathbf{I})$; a modality projector $proj$ that aligns $\mathbf{V}$ to the LLM's embedding space, $\mathbf{V}_e = proj(\mathbf{V})$; and an LLM backbone θ that auto-regressively generates responses conditioned on $\mathbf{V}_e$ and textual embeddings $\mathbf{T}$. To enable latent sketching, SkiLa introduces a sketch module, used only during training. It consists of a sketch encoder E_s and projector $proj_s$, which encode intermediate "sketch images" (i.e., visual thoughts) $\mathbf{I}_s$ into visual semantic reconstruction targets $\mathbf{V}_s = proj_s(E_s(\mathbf{I}_s))$.

Hybrid Auto-Regressive Process. SkiLa's LLM backbone performs a hybrid auto-regressive process to achieve unified internal reasoning, using two special tokens, <|sketch_start|> and <|sketch_end|>, to switch reasoning modes. Upon generating <|sketch_start|>, the model enters "visual sketching mode" and auto-regressively produces a sequence of continuous latent sketch tokens $\mathbf{H}_{sketch} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$, trained to reconstruct the visual thought semantics $\mathbf{V}_s$. Upon generating <|sketch_end|>, it seamlessly resumes standard next-token prediction for textual thoughts. Formally, given an image-question pair $(\mathbf{I}, \mathbf{Q})$, SkiLa generates a unified reasoning trace $\mathcal{S} = \{s_1, \ldots, s_N\}$ via

$$s_i \sim \mathcal{P}_\theta(s|\mathbf{V}_e, \mathbf{T}, s_1, \cdots, s_{i-1}) \tag{1}$$

where $s_i \in \{t_i, h_i\}$ is either a text token t_i or a latent sketch token h_i.

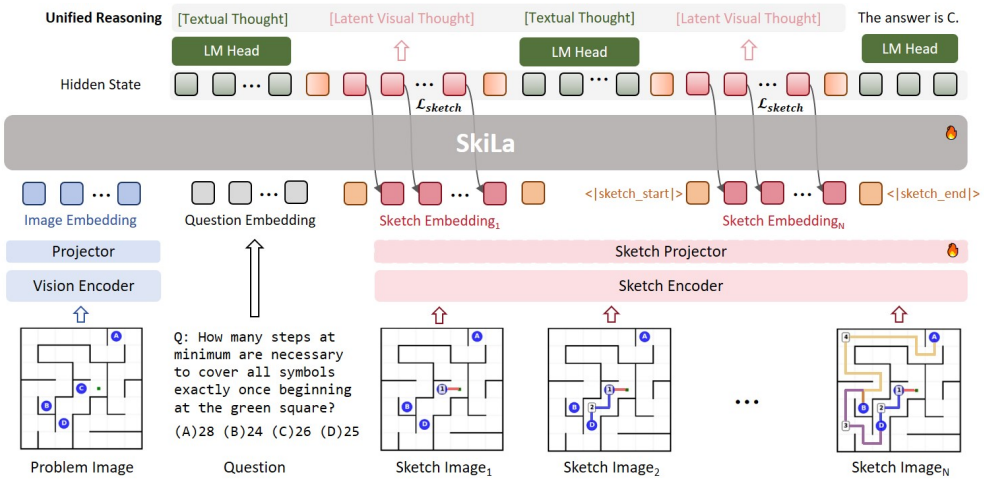


Figure 2: Training and inference of SkiLa. The model dynamically alternates between textual thinking mode (generating textual thoughts) and visual sketching mode (generating latent sketch tokens as visual thoughts). During training, a sketch module (encoder and projector) extracts visual embeddings from sketch images as reconstruction targets to ensure the latent sketch tokens are semantically grounded; this module is discarded at inference.

3.2 SFT: Explicitly Ground Sketch Tokens

We jointly optimize the model's generation of coherent text and meaningful latent visual semantics, supervising over an interleaved reasoning trace of textual thoughts $\mathbf{X}_t$ and sketch images $\mathbf{I}_s$.

For each textual thought $\mathbf{X}_t$, the target is the standard sequence of text tokens $\mathbf{T}_t$. For each sketch image $\mathbf{I}_s$, we formulate a concise visual target: E_s extracts M patch-level features (e.g., 729 for SigLIP2), $proj_s$ projects them into the LLM's embedding space, and we apply average pooling to compress them into K salient vectors (e.g., $K = 27$), balancing visual-thought length against textual thoughts and providing concise semantics for supervision. The final reconstruction target is $\mathbf{V}_s = \{\mathbf{v}_{s,1}, \ldots, \mathbf{v}_{s,K}\}$. We then train the LLM backbone θ with a two-part objective:

Latent Sketch Reconstruction Loss. In latent sketching mode (i.e., after <|sketch_start|>), the model auto-regressively generates K latent sketch tokens $\mathbf{H}_{sketch} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$ by propagating each step's last hidden state as the next step's input embedding. We enforce $\mathbf{H}_{sketch}$ to approximate the ground-truth $\mathbf{V}_s$ via mean squared error (MSE):

$$\mathcal{L}_{Sketch} = \frac{1}{K} \sum_{j=1}^{K} \|\mathbf{h}_j - \mathbf{v}_{s,j}\|_2^2, \tag{2}$$

where $\mathbf{h}_j$ and $\mathbf{v}_{s,j}$ are the j-th generated and target vectors. The sketch projector is jointly trained via this loss, requiring no separate training stage.

Next-Token Prediction (NTP) Loss. For all discrete tokens (including special tokens, textual thoughts $\mathbf{X}_t$, and the final answer $\mathbf{X}_A$), we apply a standard cross-entropy loss. A sketch mask is applied to ensure this loss is not computed on the latent sketch tokens that are enclosed by the special tokens <|sketch_start|> and <|sketch_end|>.

$$\mathcal{L}_{NTP} = -\sum_{t_i} \log p_\theta(t_i | \mathbf{V}_e, \mathbf{T}, s_{1:i-1}). \tag{3}$$

Joint Objective. The final training loss for SkiLa is a weighted sum balanced by λ_{Sketch} (default 0.1):

$$\mathcal{L}_{SkiLa} = \mathcal{L}_{NTP} + \lambda_{Sketch} \cdot \mathcal{L}_{Sketch}. \tag{4}$$

3.3 RL: Implicitly Refine Sketch Tokens

After supervised training, we apply GRPO to further refine SkiLa. Unlike SFT, which explicitly aligns each latent sketch token to a visual embedding target, RL imposes no constraint on intermediate sketches, allowing the model to freely explore useful latent visual states. This reframes latent sketching as an internal reasoning action rather than an imitation target: a sketch benefits training only if it improves the final answer.

Masked GRPO for Latent Sketching. For each prompt, SkiLa samples G responses, each containing textual reasoning, sketch delimiters, continuous latent sketch states, and a final answer. The reward is $R^g = R^g_{\text{acc}} + R^g_{\text{fmt}}$, where R^g_{acc} measures answer correctness and $R^g_{\text{fmt}} = 0.2$ rewards valid response formatting. The normalized advantage is

$$A^g = \frac{R^g - \text{mean}(\{R^j\}_{j=1}^G)}{\text{std}(\{R^j\}_{j=1}^G) + \epsilon}. \tag{5}$$

While this is standard, directly applying the update to all tokens is infeasible since latent sketch states are continuous embeddings without well-defined likelihoods. We therefore mask them out and apply the policy-gradient loss only to discrete tokens:

$$\mathcal{L}_{\text{GRPO}} = -\frac{\sum_{g,i} m_i^g \min\left(\rho_i^g A^g,\ \bar{\rho}_i^g A^g\right)}{\sum_{g,i} m_i^g}, \tag{6}$$

where ρ_i^g is the per-token importance ratio between current and old policies, $\bar{\rho}_i^g$ its clipped form with threshold δ, and $m_i^g = \mathbb{I}[i \notin \mathcal{S}^g_{\text{latent}}]$ masks latent sketch positions. Although excluded from the loss, latent sketches still condition subsequent tokens, so RL implicitly refines them toward correct answers. Following DAPO (Yu et al., 2026), we dynamically discard groups with zero reward variance during training, as they provide no relative learning signal.

RL Dataset Curation. We curate SkiLa-RL-4K, which contains 4,130 samples from Thyme (Zhang et al., 2025a), VisualProbe (Lai et al., 2025), and DeepEyesV2 (Hong et al., 2025b), spanning multimodal reasoning and fine-grained visual understanding. We retain only samples with pass@8 $\in (0, 1)$ to ensure non-trivial reward variance, and apply a multimodal judge to exclude prompts with corrupted images or semantic ambiguity. See Appendix A.2 for details.

4 Experiments

Training Details.

We adopt Qwen3-VL-8B (Bai et al., 2025a) as the base model, with SigLIP2-So/14-384 (Tschannen et al., 2025) as the sketch encoder and a single-layer MLP as the sketch projector. Only the LLM and sketch projector are trained; all other parameters (vision encoder, sketch encoder) remain frozen. For SFT, we train for 1 epoch with batch size 128 and a cosine scheduler with initial learning rate 1e-5. For RL, we use batch size 64 with 8 rollouts per prompt. The learning rate is set to 1e-6, and the KL penalty coefficient is set to 0 to encourage extensive exploration.

SkiLa Training Sample

Interleaved Textual and Visual Thoughts

<think> [Textual Thought 1] </think>
<|sketch_start|> [Sketch Image 1] <|sketch_end|>
<think> [Textual Thought 2] </think>
<|sketch_start|> [Sketch Image 2] <|sketch_end|>
... ...
<think> [Textual Thought N] </think>
<|sketch_start|> [Sketch Image N] <|sketch_end|>
<think> [Final Textual Thought] </think>
<answer> [Final Answer] </answer>

Figure 3: The SkiLa training sample format structure.

Training Datasets. Our SFT data is curated from Zebra-CoT (Li et al., 2025a) (182K interleaved text-image reasoning traces). We filter this dataset to select samples suitable for SkiLa training, retaining 101K samples and reformatting them into interleaved textual thoughts and visual sketch targets following the structure in Figure 3. RL training uses SkiLa-RL-4K described in Sec. 3.3.

Methods	Visual Understanding			Multimodal Reasoning	
	V*Bench	HRBench 4K	HRBench 8K	MME-RW-Lite	MME-RW
Textual Reasoning Models					
GPT-4o	66.0	59.0	55.5	52.0	62.8
LLaVA-OneVision-7B	75.4	63.0	59.8	48.5	57.4
Qwen2.5-VL-7B	78.5	68.5	65.0	44.3	56.8
Qwen2.5-VL-32B	80.6	74.1	69.9	46.2	59.1
Qwen3-VL-8B*	83.8	76.5	71.3	48.7	61.9
InternVL3-8B	81.2	70.0	69.3	47.9	62.0
Latent Visual Reasoning Models					
LVR	81.7	69.6	66.1	49.1	62.5
Monet	83.3	71.0	68.0	46.9	-
CoVT	78.5	72.5	69.9	48.2	63.3
Multimodal Agentic Models					
Pixel Reasoner	84.3	72.6	66.1	49.7	64.4
DeepEyes	83.3	73.2	69.5	53.2	64.1
Thyme	82.2	77.0	72.0	55.2	64.8
DeepEyesV2	81.8	77.9	73.8	52.4	64.9
SkiLa	**88.0**	**81.5**	**77.5**	**59.4**	**69.8**

Table 1: **Performance on fine-grained perception and multimodal understanding benchmarks.** SkiLa consistently outperforms existing methods. Here, $*$ denotes the results tested by ourselves.

4.1 Evaluation Benchmarks and Baselines

Evaluation Benchmarks. We evaluate SkiLa from two complementary angles. For *fine-grained visual understanding and multimodal reasoning*, we adopt V*Bench (Wu & Xie, 2024), HRBench 4K/8K (Wang et al., 2025d), MME-RealWorld (Zhang et al., 2024) and its lightweight variant MME-RealWorld-Lite. These benchmarks emphasize high-resolution perception, real-world complexity, and detail-sensitive reasoning. For *general multimodal capability*, we further test SkiLa on MMStar (Chen et al., 2024), RealWorldQA (RWQA) (XAI, 2024), and MMBench v1.1 (Liu et al., 2024), which evaluate broad multimodal reasoning beyond vision-centric scenarios.

Baselines. To comprehensively evaluate our approach, we compare SkiLa against three suites of models. 1) Textual reasoning models that verbalize visual content for reasoning, including the proprietary GPT-4o (Hurst et al., 2024) and open-source LLaVA-OneVision (Li et al., 2024), Qwen2.5-VL-7B/32B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3-8B (Wang et al., 2025c). 2) Latent visual reasoning models that perform visual reasoning in continuous embedding space, including LVR (Li et al., 2025b), Monet (Wang et al., 2025b), and CoVT (Qin et al., 2025). 3) Multimodal agentic models that invoke predefined visual tools (e.g., cropping) during reasoning, including Pixel Reasoner (Su et al., 2025), DeepEyes (Zheng et al., 2025), Thyme (Zhang et al., 2025a), and DeepEyesV2 (Hong et al., 2025b). In Tables 1 and 2, $*$ denotes results reproduced by ourselves; the remaining numbers are taken from the corresponding original papers or the OpenVLM (Duan et al., 2024) leaderboard.

4.2 Main Results

4.2.1 Vision-Centric and Multimodal Tasks

Table 1 reports results on five visually intensive benchmarks. SkiLa consistently achieves the best performance across all benchmarks, yielding substantial gains over its Qwen3-VL-8B backbone. These results validate two findings. **(1) Latent visual thinking is more effective than verbalizing visual concepts.** Compared with textual reasoning baselines, SkiLa's latent sketch tokens directly capture fine-grained visual cues that linguistic descriptions struggle to express, leading to consistent improvements on high-resolution and detail-sensitive tasks (V*Bench, HRBench). **(2) Unified reasoning is more flexible than tool-based agents and prior latent-only methods.** Tool-based agents are bounded by predefined visual operations, while prior latent visual reasoning methods such as LVR and Monet are confined to a single inferential step. SkiLa unifies textual and visual thoughts in a single auto-regressive trace, which enables multi-step interleaved reasoning and yields the largest

gains on benchmarks that simultaneously demand fine-grained perception and complex reasoning, such as MME-RW-Lite and MME-RW.

4.2.2 General VQA Tasks

To assess generalization beyond vision-centric scenarios, we further evaluate SkiLa on three general-purpose multimodal benchmarks in Table 2. Despite being trained primarily for visually intensive reasoning, SkiLa still achieves the best performance on all three benchmarks, outperforming its Qwen3-VL-8B backbone. The improvement on MMStar (+6.6 over the backbone) is particularly notable, indicating that latent sketch tokens not only help fine-grained perception but also strengthen high-level multimodal reasoning. These results confirm that the unified reasoning paradigm of SkiLa generalizes well to broad VQA settings rather than overfitting to vision-centric tasks.

Methods	MMStar	RWQA	MMBv1.1
GPT-4o	61.6	68.6	83.1
LLaVA-OV-7B	61.9	69.9	80.9
Qwen2.5-VL-7B	60.3	67.4	82.5
Qwen2.5-VL-32B	70.3	70.2	-
Qwen3-VL-8B*	64.7	71.8	83.2
InternVL3-8B	68.2	70.8	81.7
LVR*	61.3	67.7	74.6
CoVT	-	71.8	-
Thyme	65.9	70.2	-
SkiLa	**71.3**	**74.3**	**84.6**

Table 2: **Performance on general VQA tasks.**

4.3 Ablation Studies

Models	V*	HR4K	HR8K	MMStar
Baseline	83.8	76.5	71.3	64.7
Vision Only SFT	84.3	77.9	71.9	65.3
Text Only SFT	83.8	76.4	71.2	66.1
SkiLa-SFT	85.3	78.0	72.3	66.9
SkiLa-RL	88.0	81.5	77.5	71.3

Table 3: Ablation on SkiLa's training paradigm.

SFT Stage	Sketch Token Num	MMStar	HR4K	V*
SkiLa-QwenViT	28	66.4	77.0	83.8
SkiLa-CLIP	24	65.1	77.5	84.3
SkiLa-SigLIP2	27	**66.9**	**78.0**	**85.3**

Table 4: Impact of sketch encoders used as the SFT reconstruction target for latent semantics.

Effect of Each Component. Table 3 dissects SkiLa's training paradigm. *Vision Only SFT*, which keeps only latent sketch supervision and removes textual thoughts, brings consistent but moderate gains on visually intensive benchmarks (e.g., +1.4 on HR4K). *Text Only SFT*, which keeps only textual CoT, mainly benefits MMStar (+1.4) but is essentially flat on V*Bench, HR4K, and HR8K. Combining the two into our *SkiLa-SFT* consistently outperforms either single-modality variant on every benchmark, indicating that latent visual thinking and textual reasoning are complementary rather than redundant. Adding RL on top (*SkiLa-RL*) further yields the largest jump, confirming that RL effectively refines latent sketches by allowing the model to freely explore useful latent visual states.

Choice of Sketch Encoder. Table 4 compares three sketch encoders for the latent reconstruction target, namely the native QwenViT, CLIP, and SigLIP2, under comparable token budgets. The key question is whether architectural consistency with the backbone outweighs the richer semantics of independently pre-trained encoders. The results decisively favor the latter: SigLIP2 achieves the best score on all three benchmarks, indicating that feature generality from a powerful, contrastively pre-trained encoder matters more than architectural consistency.

Num	1	9	27	81
V* Bench (SFT Result)	83.2	84.3	**85.3**	84.8

Table 5: Impact of the number of sketch tokens.

λ_{Sketch}	0.1	0.2	0.5	1.0
V* Bench (SFT Result)	**85.3**	84.3	83.8	82.7

Table 6: Impact of the latent sketch loss weight.

Impact of Sketch Token Number. We study how the number of reconstructed sketch visual tokens affects model performance. As illustrated in Table 5, performance improves steadily with more visual tokens, peaking at 27, and then slightly drops at 81. Empirically, 27 tokens yield the best performance, suggesting that more sketch visual tokens do not always bring further gains. Too many visual tokens

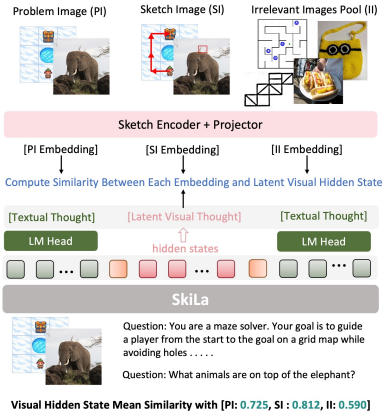


Figure 4: Quantitative interpretability of latent sketch tokens.

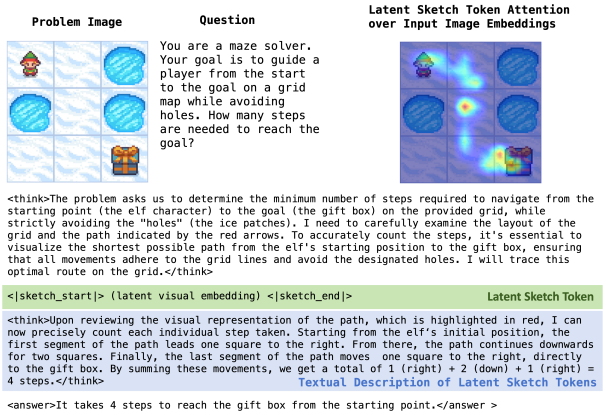


Figure 5: Qualitative interpretability of latent visual sketch tokens.

may reduce the relative contribution of textual reasoning, highlighting the need to balance visual and textual thoughts.

Impact of the Reconstruction Loss Weight. Table 6 studies the latent sketch reconstruction loss weight on V*Bench. Performance is highest at λ_{Sketch}=0.1 and degrades monotonically as the weight grows. This indicates that an overly large reconstruction weight overshadows the next-token prediction objective, harming the model's textual reasoning ability and ultimately hurting accuracy. We therefore set 0.1 as the default.

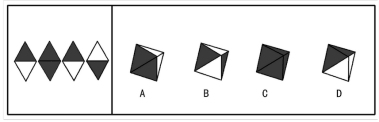
4.4 Interpretability of Sketch Visual Tokens

Quantitative Semantic Similarity. We encode the problem image (PI), the intermediate sketch target (SI), and a pool of irrelevant images (II) with the same sketch encoder and projector, and compute the **mean similarity** between their embeddings and SkiLa's latent sketch tokens (averaged over all evaluation samples). As shown in Figure 4, these hidden states are most similar to SI (**0.812**), then PI (0.725), and least to II (0.590), confirming that the tokens semantically encode the intermediate visual thinking process rather than redundant input features or random noise.

Qualitative Behavior Analysis. We further inspect what the latent sketch tokens attend to and what they convey, as illustrated in Figure 5. *(1) Attention Map Visualization.* Attention from the latent sketch tokens to input image embeddings is strongly concentrated along the correct navigation path rather than diffusely distributed, providing direct evidence that the tokens attend to task-relevant visual regions and encode purposeful visual reasoning. *(2) Conditioned Textual Captioning.* After generating the latent sketch tokens, the model accurately describes the intended reasoning trajectory, i.e., the planned path from start to goal, indicating that the tokens encode coherent and human-interpretable visual semantics that guide subsequent textual reasoning.

4.5 Discussion

Beyond quantitative metrics, our evaluation also reveals several emerging capabilities, which we analyze through case studies below. **Spatial Imagination.** Purely text-based reasoning methods often struggle with spatial tasks, since textual descriptions are inherently ambiguous when representing complex 3D geometric properties. In contrast, SkiLa overcomes this limitation by generating visual thoughts that mentally simulate the 3D object folded from a 2D pattern, and then matching them against candidate options to identify the correct fold, as illustrated in Figure 6. **Hallucination Mitigation.** Leveraging our unified reasoning, SkiLa can implicitly zoom in on key image regions to acquire fine-grained information, shown in Figure 7. This process effectively mitigates hallucinations on visual details, such as correctly identifying all four white pillows, while circumventing the latency associated with complex and explicit tool invocations.

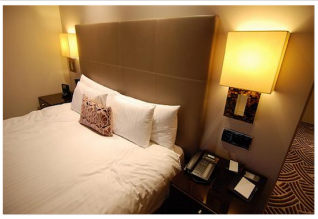


Question: Which one picture can be folded from the outer surface displayed on the left?
(A) Picture A (B) Picture B (C) Picture C (D) Picture D
Answer: (B)

Vision-R1: <think> ..... The key here is to figure out which of the options (A, B, C, D) can be folded from the outer surface. The image displays that Picture A is the correct answer because it matches the structure of the diamonds on the left ..... </think><answer> Final Answer: A </answer> ✗

SkiLa: <think>The problem asks to identify which of the provided options (A, B, C, or D) can be formed by folding the pattern shown in the left image. To determine the correct fold, it is helpful to visualize the three-dimensional shape that the two-dimensional pattern represents. I will generate an image showing the complete 3D object.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>The generated image shows the complete 3D object. By comparing this visual representation with the options provided in the original problem, we can see which option matches the shape and orientation of the folded object. Option B aligns perfectly with the visualized 3D object.</think><answer>D</answer> ✓

Figure 6: Spatial imagination example. SkiLa generates visual thoughts to imagine the 3D object from a 2D pattern.



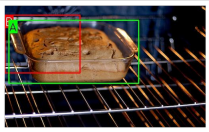
Question: How many white pillows are in this image? (A) 5 (B) 3 (C) 4 (D) 6
Answer: (A)

Vision-R1: <think> The image displays there are three white pillows at the head of the bed. Then there's a patterned pillow in between them. So the white ones are the three mentioned. The question is asking how many white pillows there are. The options are A to D. The answer should be B. </think><answer> Final Answer: A </answer> ✗

SkiLa: <think>The task is to count the number of white pillows visible on the bed in the provided image. To accurately count the pillows, I will focus on the area around the headboard where they are located.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>The zoomed-in view clearly shows four distinct white pillows arranged at the head of the bed. Based on this clear visual count, the total number of white pillows is four.</think><answer>D</answer> ✓

Figure 7: An example of hallucination mitigation.

Adaptive Unified Reasoning. Figure 8 shows that SkiLa supports adaptive unified reasoning. The task requires localizing the target and selecting the optimal bounding box from two candidates. In the upper case, SkiLa identifies the target in a single reasoning step. In the harder lower case, where the two candidates heavily overlap with comparable confidence, SkiLa adaptively performs two sequential steps to resolve subtle differences and pick the correct box, demonstrating its ability to adjust reasoning depth to task difficulty.



Question: Which bounding box more accurately localizes and encloses the dish? (A) Box A (B) Box B
Answer: (A)

SkiLa: <think>The problem asks to identify which of the two bounding boxes, A or B, provides a better fit for the plate containing the hot dog. To determine this, I will visually assess how well each bounding box aligns with the edges of the plate in the image.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>Observing the visual thought, it is clear that Box B extends beyond the plate's edge, indicating it does not fully contain it. In contrast, Box A precisely outlines the plate, showing no overlap or gaps. Based on this visual comparison, Box A is the correct choice as it perfectly encloses the plate.</think><answer>D</answer> ✓



Question: Which bounding box more accurately localizes and encloses the carrot? (A) Box A (B) Box B
Answer: (A)

SkiLa: <think>The problem asks to identify which bounding box best outlines the carrots in the image. I need to examine both options, A and B, to see how well they fit around the carrots. To clearly assess the fit of each bounding box, I will visualize the area covered by Box A.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>Observing the visual thought, Box A appears to be too narrow and does not fully encompass all the carrots. Now, let's visualize the area covered by Box B.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>The visual thought shows that Box B perfectly surrounds all the carrots, indicating it is the correct bounding box. Comparing the two options, Box B provides a more precise and complete boundary for the carrots than Box A.</think>
<answer>B</answer> ✓

Figure 8: An example of adaptive unified reasoning.

5 Conclusion

We present Sketch-in-Latents (SkiLa), a unified reasoning paradigm that enables MLLMs to natively generate latent sketch tokens as visual thoughts, dynamically interleaving textual thinking and visual sketching within a single auto-regressive trace. SkiLa is trained in two stages: SFT *explicitly* grounds latent sketch tokens via a latent visual semantics reconstruction mechanism, while RL with masked GRPO on our curated SkiLa-RL-4K further *implicitly* refines them. Extensive experiments confirm SkiLa's strong performance on vision-centric tasks and its solid generalization across diverse multimodal benchmarks.

AI use statement

In this work, we did not use generative AI tools for any task with required disclosure. Specifically, generative AI was not used to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or assist in writing proofs, propose or refine hypotheses, design or provide feedback on our research methodology or experiments, implement methods, perform translation, clean or reformat datasets, support qualitative or thematic analysis, or interpret results; these uses are not applicable to this work. Among tasks with recommended disclosure, we used

generative AI tools solely to aid and polish the writing, i.e., correcting grammar and spelling and improving the readability and phrasing of author-written text. We did not use generative AI for other recommended-disclosure tasks such as creating scientific figures, writing or editing code, searching for or identifying literature, formatting references, or brainstorming research ideas. We have reviewed all AI-assisted edits to ensure they are correct and faithfully preserve the authors' intended meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3 (5), 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

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.

Martha J Farah. Psychophysical evidence for a shared representational medium for mental images and percepts. *Journal of Experimental Psychology: General*, 114(1):91, 1985.

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.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025a.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025b.

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. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Stephen M Kosslyn. *Image and brain: The resolution of the imagery debate*. MIT press, 1996.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025a.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*, 2025b.

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*, 2024.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025c.

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, 2023.

Pengxiang Li, Shilin Yan, Joey Tsai, Renrui Zhang, Ruichuan An, Ziyu Guo, and Xiaowei Gao. Adaptive classifier-free guidance via dynamic low-confidence masking. *arXiv preprint arXiv:2505.20199*, 2025d.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025a.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025b.

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.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *CoRR*, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025a.

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. *arXiv preprint arXiv:2511.21395*, 2025b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

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, 2025d.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. *URL https://arxiv. org/abs/2312.14135*, 5.

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.

XAI. Grok-1.5 vision preview, 2024. URL `https://x.ai/news/grok-1.5v`.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

Shilin Yan, Jiaming Han, Joey Tsai, Hongwei Xue, Rongyao Fang, Lingyi Hong, Ziyu Guo, and Ray Zhang. Crosslmm: Decoupling long video sequences from lmms via dual cross-attention mechanisms. *arXiv preprint arXiv:2505.17020*, 2025.

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*, 2025.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025a.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025b.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. Pyvision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

A Training Data Details

A.1 SFT Dataset.

Our SFT training data is curated from the Zebra-CoT, which is a diverse large-scale dataset with 182K samples, containing logically coherent interleaved text-image reasoning traces. We filter this dataset to select samples suitable for SkiLa training, excluding 3D data and instances where the intermediate sketch images are too complex for the MLLM to effectively learn to reconstruct. Consequently, our training dataset contains 101K samples. Our detailed statistics of training dataset is shown in Table 7.

A.2 RL Dataset

We construct **SkiLa-RL-4K** from prompts in Thyme (Zhang et al., 2025a), VisualProbe (Lai et al., 2025), and DeepEyesV2 (Zheng et al., 2025). To obtain a reward landscape that is both well-posed and informative, we apply two complementary filters. (i) An automatic reviewer instantiated as Qwen3-VL-235B removes samples with corrupted or low-resolution images, under-specified or ambiguous questions, and image-text inconsistencies, so that every retained prompt admits a unique, verifiable answer. (ii) We run 8 rollouts per prompt with the backbone and retain only those with $\text{pass@8} \in (0, 1)$: prompts solved by all or none yield zero reward variance and contribute no learning signal under GRPO.

The resulting training set of 4,130 samples contains three categories, with their proportions visualized in Figure 9. *Visual Perception* (47.4%) targets fine-grained recognition and chart understanding with short, verifiable answers. *Visual Reasoning* (42.6%) covers multi-step mathematical and geometric problems grounded in images, scored by symbolic equivalence. *Capability Enhancement* (10.0%) provides long-form generation tasks, including detailed captioning, OCR transcription, and chart description, which preserve the model's general instruction-following ability during RL fine-tuning.

Category	Count	Percentage (%)
Visual Jigsaw	21,899	21.6
Visual Search	30,000	29.6
Geometry	1,058	1.1
Checkers	2,753	2.7
Chess	20,483	20.2
Connect Four	2,029	2.0
Maze	20,000	19.8
RPM	3,000	3.0
Total	101,222	100.0

Table 7: Detailed statistics of training data.

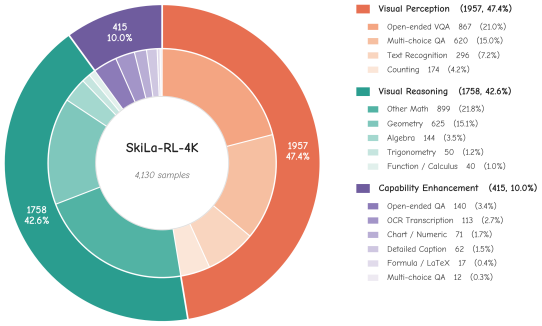
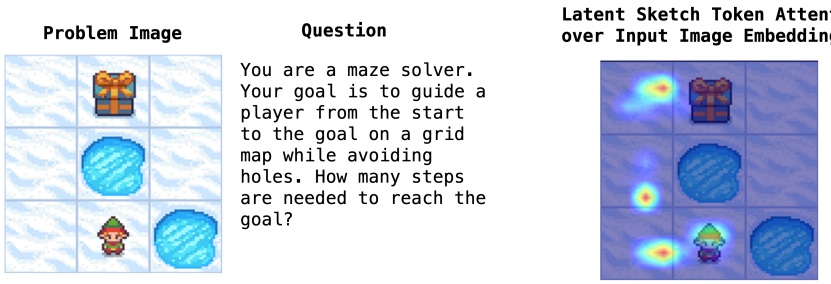


Figure 9: Category distribution of **SkiLa-RL-4K** (4,130 samples) across Visual Perception, Visual Reasoning, and Capability Enhancement.

B More Attention Map Cases

To complement the attention analysis in the main text, we present additional cases in Figure 10 that span a broader range of task types. Across these examples, attention from the latent sketch tokens to input image embeddings consistently concentrates on task-relevant regions rather than being diffusely distributed, and the model can faithfully verbalize the attended content as a coherent reasoning trajectory. These results further confirm that the latent sketch tokens encode purposeful, human-interpretable visual semantics that guide subsequent textual reasoning, rather than redundant or arbitrary features.



<think>The problem asks us to determine the minimum number of steps required to navigate from the starting position of the elf in the bottom-left corner to the gift box located at the top-left corner of the grid, as depicted in the provided image. The critical constraint is to avoid the ice holes, which are represented by the blue ice blocks. To accurately count the steps, it's essential to visualize the shortest possible path on the grid, ensuring that all movements adhere to the maze's rules and avoid the obstacles. I will trace the most direct route from the elf's starting point to the gift box, carefully navigating around the ice holes. This visual representation will help in precisely counting each individual step taken.</think>

<|sketch_start|> (latent visual embedding) <|sketch_end|> **Latent Sketch Token**

<think>Upon examining the path highlighted in the image, I can now systematically count each segment of the red line. Starting from the elf's initial position, the first segment goes left, then two segments go upward, and finally one segment goes right. Each of these movements counts as one step. Counting these distinct movements shows that the total steps needed to reach the goal are 4.</think> **Textual Description of Latent Sketch Tokens**

<answer>It takes 4 steps to reach the gift box from the elf's starting position.</answer>

Figure 10: Qualitative interpretability of latent sketch tokens: attention map visualization and conditioned textual captioning.

C More Cases

In the main text, we showed a subset of examples demonstrating SkiLa's effectiveness. Here, we include additional and broader cases across more task types.

Fine-Grained Visual Grounding. As shown in Figure 11, SkiLa's unified latent reasoning enables it to internally focus on key regions of an image and recover fine-grained visual details without relying on external tools. This internal refinement allows SkiLa to correctly identify subtle attributes, such as recognizing that all three glasses contain water, while avoiding the latency introduced by explicit region-based operations. Conventional models often overlook such details or produce hallucinated predictions due to insufficient visual grounding.



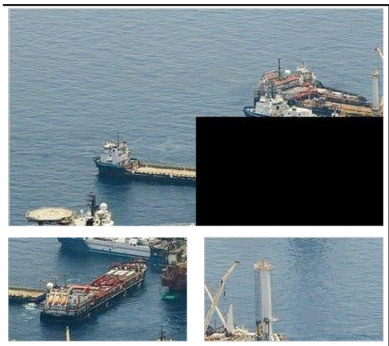
Question: How many glasses are filled with water? (A) 5 (B) 3 (C) 4 (D) 2?
Answer: (B)

Vision-R1: <think> The image displays there are three glasses on the table. Two of them have clear liquid, which is water. The third one is empty. The question is asking how many glasses are filled with water. So, the answer should be two. The options are A to D, with D being 2.</think><answer> Final Answer: D </answer> ✗

SkiLa: <think>The problem asks to count the number of glasses containing water in the provided image. I need to carefully examine each glass present on the table. To accurately identify and count the glasses with water, I will focus on the area around the pizza and the table setting.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>The zoomed-in view clearly shows three glasses on the table: one next to the pizza, one near the top right, and another at the bottom left. All three glasses appear to contain clear liquid, which is consistent with water. Based on this detailed observation, there are three glasses filled with water.</think><answer>B</answer> ✓

Figure 11: A case study demonstrates fine-grained visual grounding capability of SkiLa.

Latent Scene Completion. As illustrated in Figure 12, SkiLa can reconstruct the missing content of a partially occluded image within its latent sketch space and determine which candidate completion is semantically consistent with the underlying scene. This ability to integrate fragmented visual information enables SkiLa to recover the correct jigsaw piece. In contrast, Vision-R1 incorrectly identifies the missing part because it struggles to form a coherent global scene from the incomplete visual input.



Question: Which image is the missing part in the first image?
(A) the second image (B) the third image
Answer: (A)

Vision-R1: <think> The user is asking which image is the missing part in the first image. The options are A (the second image) or B (the third image). From the image provided, the first image is incomplete and lacks the vertical structure with cranes that's shown in the third image. The second image shows the ship from a different angle but doesn't include the cranes. The third image specifically displays the vertical structure with cranes, which is missing from the first image. So the missing part in the first image must be the third image. Therefore, the answer should be B. </think><answer> Final Answer: B </answer> ✗

SkiLa: <think>The problem asks to identify which image completes the scene shown in the first image. I need to compare the missing section of the first image with the provided options. To determine the missing part, I will visualize the complete scene without the black box covering it.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>The reconstructed image shows the full scene. Comparing this with the original image and the options, the missing area corresponds to the second image. The second image contains the parts that were obscured by the black box in the first image.</think><answer>A</answer> ✓

Figure 12: A case study demonstrates latent scene completion capability of SkiLa.

Cross-Image Visual Alignment. As shown in Figure 13, SkiLa conducts unified reasoning directly over fine-grained visual semantics. In this example, SkiLa accurately aligns the "REF" marker with the cat's front paw by reconstructing its precise spatial location across images. In contrast, Vision-R1 matches only at a coarse body-level region and therefore predicts incorrectly. This demonstrates that SkiLa not only captures subtle cross-image correspondences but also integrates them seamlessly into its subsequent textual reasoning.



Question: Which point is corresponding to the reference point? (A) Point A (B) Point B (C) Point C (D) Point D
Answer: (D)

Vision-R1: <think> The question is asking which point corresponds to the reference point. The image displays that in the first image, there's a reference point labeled \"REF\" near the cat's body. Then, in the second image, there are points A, B, C, D marked on the cat. The key detail here is that Point C is located in the same area as the reference point \"REF\" in the first image. So, the reference point from the first image is at Point C in the second image. The answer should be C.</think><answer> Final Answer: C </answer> ✗

SkiLa: <think>The problem asks to identify which point in the second image corresponds to the \"REF\" marker visible in the first image. To accurately match the reference point, I need to locate its position relative to the cat and the furniture in the second image. </think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think>The visual thought clearly shows the red circle labeled \"REF\" on the table surface in the first image. In the second image, this corresponds to the area where the cat's paws are resting on the black surface. Comparing the location of the \"REF\" marker in the first image to the points marked on the cat in the second image, it aligns with the cat's front right paw (point D).</think>
<answer>D</answer> ✓

Figure 13: A case study demonstrates cross-image visual correspondence alignment capability of SkiLa.

Multi-Round Visual Refinement. SkiLa demonstrates the ability to perform detailed, multi-step visual reasoning by iteratively examining and comparing regions of interest. As shown in Figure 14, it inspects the two points across multiple rounds, progressively focusing on their local visual properties. By integrating observations step by step, SkiLa accurately determines that the points have similar colors. In contrast, Vision-R1 relies on a coarse global judgment and predicts incorrectly. This case illustrates how SkiLa can conduct stepwise, focused analysis to capture subtle visual features that conventional models often miss.



Question: Consider the surface color (color without shading) of the two points in the image. Which one is darker, or the color is about the same? (A) A is darker (B) B is darker (C) About the same
Answer: (C)

Vision-R1: <think> ..... The image displays that point A is on a blue pot and point B is on a red pot. The blue pot is on the lower shelf, and the red pot is on the upper shelf. The key detail here is that the blue pot (point A) appears darker in color compared to the red pot (point B) ..... </think> <answer> Final Answer: A </answer> ✗

SkiLa: <think> The problem asks to compare the colors of two points, A and B, shown in the image. To accurately determine which point has a darker color, I need to visually inspect the area around each letter. I will focus on the region containing the letter 'A'.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think> Observing the zoomed-in view of point A, it appears to be a light gray or off-white color. Now, I will examine the region around point B.</think>
<|sketch_start|><|skila|>*27<|sketch_end|>
<think> The zoomed-in view of point B also shows a similar light gray or off-white color. Both points A and B exhibit very similar shades of color. Based on this visual comparison, the colors at both points appear to be about the same.</think>
<answer>C</answer> ✓

Figure 14: A case study demonstrates multi-round visual refinement capability of SkiLa.