DISTILLING VISUAL REASONING INTO TEXT SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision–Language Models (LVLMs) have shown strong promise for multimodal reasoning, yet often struggle with tasks requiring concepts beyond what is directly observable in the input image. Existing methods generate intermediate images or latent visual tokens to guide reasoning, but these representations can introduce errors and increasingly interfere with textual reasoning as reasoning progresses. We propose Visual-to-Text Chain-of-Thought Distillation (V2T), a framework that enables LVLMs to internalize visual reasoning *without* generating intermediate visual representations at inference time. V2T first trains a teacher LVLM using interleaved visual and textual chains of thought, and then uses knowledge distillation to train a student LVLM using the teacher's logits and cross-entropy supervision from ground-truth textual reasoning. When reasoning images can be mapped to the original image, V2T can additionally distill the teacher's attention to corresponding regions, while ground-truth bounding boxes can further guide a subsequent reinforcement learning stage. Experiments across multiple multimodal reasoning benchmarks show that V2T consistently outperforms the teacher and existing baselines, improving average accuracy by 14.3% on a held-out set and 2.7% on the broader visual evaluation suite. Moreover, lightweight SFT and substantially reduced RL make V2T up to 42× faster to train than state-of-the-art baselines.

1 INTRODUCTION

Large vision–language models (LVLMs) have demonstrated strong performance on multimodal reasoning tasks involving an input image and textual question by generating textual chains of thought (CoTs). This is often achieved through supervised fine-tuning on long textual CoTs conditioned on the input image, followed by reinforcement learning to encourage step-by-step reasoning (Shao et al., 2024a; Xu et al., 2025a; Zhang et al., 2025a; Huang et al., 2025b; Wang et al., 2025c; Zhang et al., 2025b; Chen et al., 2025a). However, many problems cannot be solved from the raw image alone. For instance, as shown in Fig. 1, solving geometric problems often requires auxiliary constructions, such as additional lines or points, to reveal implicit relationships (e.g., perpendicular distances or tangency) that are not apparent in the original image. Humans naturally imagine such intermediate visual states, but LVLMs trained primarily in the text space often struggle with reasoning over such representations (Li et al., 2025b;c; Xu et al., 2025b).

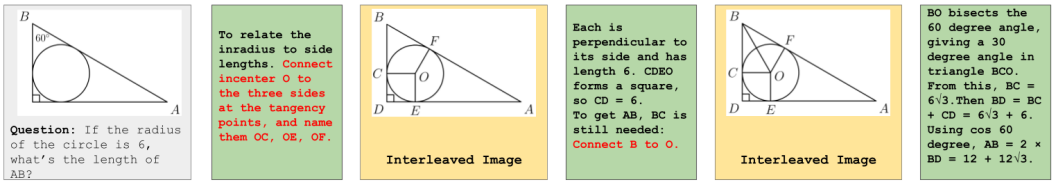
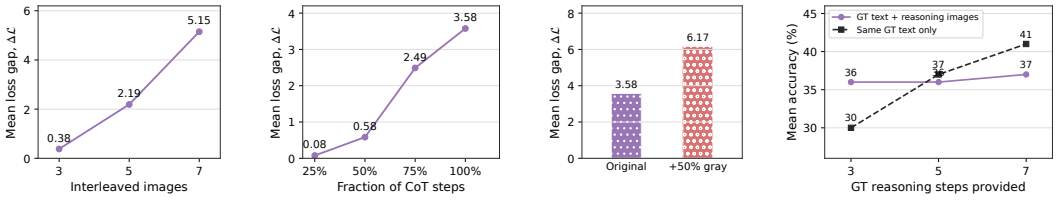


Figure 1: Example of multimodal CoT, where every step of thinking is paired with an image.

To provide such intermediate visual information, recent work interleaves visual information directly into the reasoning process. One line augments LVLMs with external visual tools (e.g., sketchpads or pixel-level operations), allowing models to manipulate images during reasoning via actions such as zooming, drawing auxiliary lines, or highlighting relevant regions (Hu et al., 2024a; Qi et al., 2024; Hu et al., 2024b; Wu et al., 2025; Su et al., 2025b; Huang et al., 2025c; Wu et al., 2023; Yang et al., 2023). Another line focuses on latent visual reasoning, training models on interleaved multimodal data where textual CoTs are paired with reasoning images, and learning visual tokens that act as

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



(*a*) Image count (*b*) Reasoning progress (*c*) Irrelevant visual content (*d*) Prefix-conditioned generation

Figure 2: **Interleaved visual inputs increasingly interfere with textual reasoning.** (a–c) The loss gap between reasoning with and without interleaved images widens as reasoning progresses and further increases with irrelevant visual content. (d) With longer reasoning prefixes, text-only continuation matches or outperforms continuation with interleaved images.

surrogates for intermediate images at inference time (Su et al., 2025a; Wang et al., 2025d; Li et al., 2025b; Tong et al., 2025; Dong et al., 2025; Yang et al., 2025b).

Despite their promise, these approaches face two key limitations. First, generated intermediate images or latent visual tokens can be inaccurate or incomplete, potentially misleading subsequent reasoning and propagating errors (Li et al., 2025d; Wang et al., 2025a; Liu et al., 2025). Moreover, as shown in Fig. 2, interleaved visual inputs can increasingly interfere with textual reasoning as the reasoning process unfolds: the loss gap between reasoning with and without interleaved images widens over successive steps, and even adding irrelevant visual content, such as a gray background, further increases this gap (see Appendix B for details). This suggests that interleaved visual representations can progressively disrupt the model's textual reasoning, even when they provide little additional useful information.

These observations motivate a central question: *can visual reasoning be internalized into text-based CoT, so that models can reason as if they had access to intermediate visual information without actually generating it?* We propose Visual-to-Text Chain-of-Thought Distillation (V2T), which first trains a teacher on multimodal reasoning trajectories with intermediate reasoning images, then distills its reasoning into a student that reasons entirely in text space. The student is trained using the teacher's logits and ground-truth textual CoT, with optional attention distillation when reasoning images can be mapped to the original image, e.g. crop or transformation. When ground-truth regions are available, a short subsequent reinforcement learning stage, e.g., GRPO, can further encourage attention to task-relevant regions. At inference, the student reasons without intermediate images, avoiding their interference and error propagation while retaining the teacher's visual reasoning capabilities. V2T requires lightweight SFT for interleaved teacher training and student distillation, while the strong initialization from distillation substantially reduces the amount of subsequent RL, resulting in up to 42$\times$ faster training than baselines, even with GRPO. Remarkably, the student consistently outperforms its teacher.

Through extensive evaluation on in-domain and broad multimodal reasoning benchmarks, we show that V2T consistently outperforms both external visual-tool and latent visual-reasoning methods. On a held-out set, V2T achieves 75.5% average accuracy, improving by 14.3% over its teacher and 14.6% over the strongest existing baseline. On general benchmarks, V2T achieves an overall score of 61.8%, outperforming the strongest baseline by 2.7%. Attention and error analyses further show that V2T attends more strongly to task-relevant regions and makes fewer reasoning errors than both text-only CoT and the teacher, indicating stronger visual grounding and more reliable reasoning without intermediate images. Notably, V2T's largest gains occur in spatial and global-context reasoning (-16.1%), highlighting the benefits of internalized visual grounding for scene understanding.

2 Related Work

Existing visual CoT approaches can be broadly grouped into the following three categories.

Text-Only Visual Chain of Thoughts The first category focuses on constructing datasets where images are paired with detailed textual reasoning steps (Qiao et al., 2024; Shao et al., 2024a; Xu et al., 2025a; Hu et al., 2022b). During training, the model learns to observe the image, verbalize its perception, and reason about it before arriving at a final decision. However, a key limitation is that reasoning can become dominated by long text sequences, overshadowing the underlying visual information.

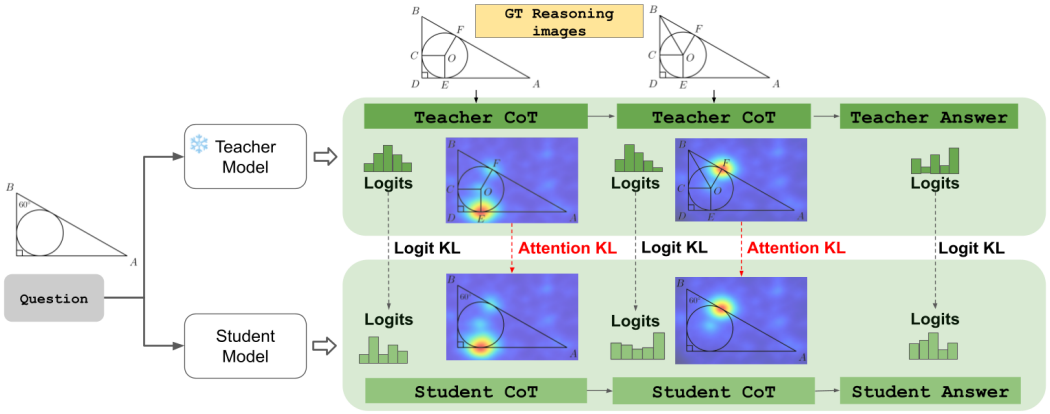


Figure 3: **V2T Training.** V2T first trains a teacher with interleaved images to ground its CoT in visual cues. The student model, which does not have access to interleaved images, is trained using KL divergence to match the teacher's textual CoT and logits, along with attention distillation to transfer the teacher's visual grounding, and cross-entropy loss to match the ground-truth textual CoT.

Tool-Augmented Visual Chain of Thoughts The second category extends beyond textual reasoning by incorporating visual tools, such as drawing bounding boxes, sketching shapes, cropping, or zooming into salient regions (Hu et al., 2024a; Qi et al., 2024; Hu et al., 2024b; Wu et al., 2025; Su et al., 2025b; Huang et al., 2025c; Wu et al., 2023; Yang et al., 2023). These tools generate intermediate images or annotations that are fed back into the model for further reasoning. While this can produce more precise and interpretable reasoning, it often requires strong models or reinforcement learning to reliably learn tool-use policies, limiting scalability. Moreover, reasoning is constrained by tool capabilities, and inaccuracies in intermediate images can propagate through the autoregressive process and degrade performance.

Interleaved Latent Visual Chain of Thoughts A third line of work integrates visual information directly into the model's reasoning process. MINT-CoT (Chen et al., 2025c) learns interleaved tokens corresponding to specific image regions, grounding each reasoning step in localized visual evidence. PixelReasoner (Su et al., 2025a) maintains a latent visual state that is iteratively updated to refine fine-grained details. Other approaches (Li et al., 2025b; Wang et al., 2025d) predict latent visual representations that can be decoded into approximate image regions, helping the model refocus on relevant areas. Finally, unified latent-space methods reason directly within shared visual representations, enabling manipulation of abstract visual features beyond the original input. However, these interleaved visual inputs can increasingly interfere with textual reasoning as the reasoning process unfolds, harming the performance (cf. Fig. 2).

We propose V2T, which distills visual reasoning from a teacher trained with interleaved images into a student that reasons entirely in text space. By avoiding intermediate visual representations during inference, V2T retains visual reasoning capabilities without the modality switching, image-generation errors, and error propagation of existing approaches. In doing so, it achieves superior performance.

3 Distilling Visual Chains of Thought into Text Space

In this section, we introduce V2T, which enables LVLMs to reason over the original problem image *as if* intermediate reasoning images were available, while generating a purely textual CoT at inference (Fig. 3; an inference example is shown in Fig. 4).

3.1 Reasoning with Large Vision-Language Models

An LVLM consists of a vision encoder and a large language model (LLM) that jointly process visual and textual inputs. Given a text prompt and an input image, the model first encodes the image and prompt into visual v and text x tokens, then performs autoregressive reasoning by modeling:

$$p_\theta(y_{1:T} \mid v, x) = \prod_{t=1}^{T} p_\theta(y_t \mid v, x, y_{<t}) \,, \tag{1}$$

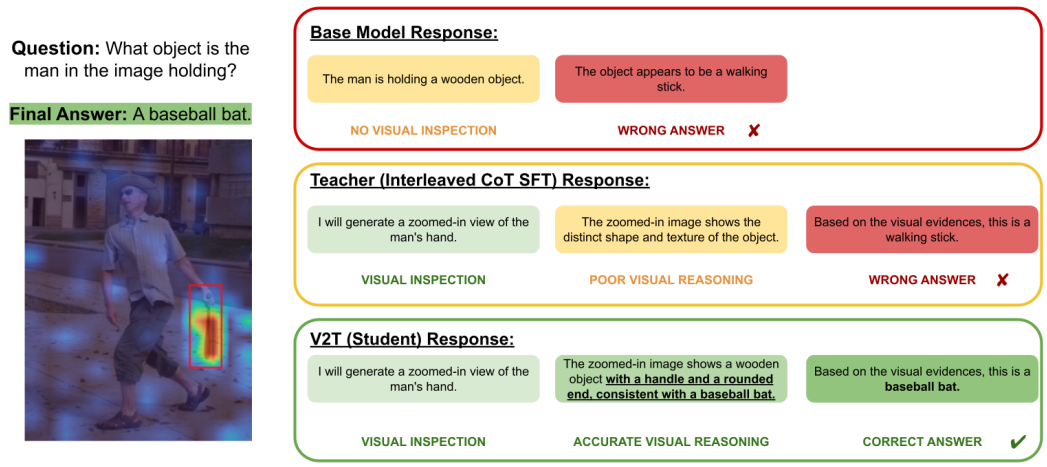


Figure 4: **V2T inference does not require intermediate images.** At inference, the base model, teacher (Interleaved CoT SFT), and V2T student (without attention distillation or GRPO) receive the question and original problem image and generate textual responses. The base model and teacher misidentify the object as a walking stick, while the student correctly identifies a baseball bat. Phrases such as "zoomed-in view" are inherited from the teacher's interleaved CoT format; no intermediate image is generated or provided, and V2T reasons only over the original image.

where $y_{<t}$ are intermediate text CoT tokens.

Limitations of LVLMs Reasoning. As shown in Fig. 1 and Fig. 9 in the Appendix, many challenging tasks benefit from pairing each CoT step with a corresponding *reasoning image*. However, standard VLMs operate autoregressively over discrete tokens and lack a reliable mechanism to generate faithful intermediate visual states. As illustrated in Fig. 2, intermediate images or visual tokens can increasingly interfere with textual reasoning as reasoning progresses. Without strong supervision, errors can propagate through autoregressive reasoning and degrade performance.

3.2 V2T: Distilling Visual Chains of Thought into Text Space

To address these limitations, our key idea is to enable LVLMs to reason in text space, *as if* grounded in rich, step-by-step visual information. V2T consists of two stages. In Stage 1, it trains a teacher on CoT steps paired with intermediate reasoning images, grounding its reasoning in visual information. In Stage 2, it uses knowledge distillation to transfer this visual reasoning to a student that reasons entirely in text space, using the teacher's logits and ground-truth textual CoT. When reasoning images can be mapped to the problem image, it additionally distills the teacher's attention to corresponding regions, followed by RL when ground-truth regions are available. V2T retains visual grounding while avoiding intermediate representations and error propagation at inference. Next, we describe V2T in detail.

3.2.1 Stage 1: Training the Teacher Model on Multimodal Reasoning Data

Let $\mathcal{D}$ denote a dataset of interleaved multimodal reasoning examples. Each example consists of a text prompt and an initial problem image, with corresponding textual and visual tokens x and v_{prob}. It further contains a K-step CoT, where each step $t \in [K]$ consists of ground-truth text y_t and may be paired with an intermediate reasoning image with visual tokens $v_{\text{reason},t}$. The complete target text sequence is $y = (y_1, \ldots, y_K)$, including the textual CoT and final answer.

We train the teacher model f_θ using cross-entropy on textual tokens, conditioned on the preceding text-image pairs:

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{t=1}^{K} \log p_\theta \left(y_t \mid x, v_{\text{prob}}, (y_j, v_{\text{reason},j})_{j<t}, v_{\text{reason},t} \right). \tag{2}$$

Here, y_t denotes the ground-truth text at step t. Each reasoning image $v_{\text{reason},t}$ serves as a conditioning input and receives no prediction loss. We use step-level notation for simplicity; in practice, the loss is computed token by token under teacher forcing.

Table 1: Zebra-CoT held-out set. Best and second-best results across all methods are highlighted in **bold** and underline, respectively. As Bagel-Zebra is trained on the full Zebra-CoT dataset, we exclude it from the evaluation. V2T achieves the best overall performance of 75.5%, outperforming the interleaved teacher model by 14.3%.

Method	Graph Algorithm	RPM	Tetris	Visual Search	Avg
LVLMs with text-only visual CoT					
R1-VL-7B	44.5	10.6	55.0	60.5	42.6
VLAA-Thinker-7B	<u>58.0</u>	54.8	68.0	63.0	60.9
VL-Rethinker-7B	45.0	55.8	<u>69.0</u>	64.5	58.6
LVLMs with interleaved latent visual CoT					
LVR-7B	17.5	23.1	25.5	<u>66.5</u>	33.2
MONET-7B	39.0	7.5	27.5	59.0	33.3
Interleaved SFT and baseline					
Qwen2.5-VL-7B	54.5	45.7	<u>69.0</u>	61.0	57.6
Interleaved SFT	48.5	<u>65.8</u>	65.0	65.5	<u>61.2</u>
V2T (Ours)	**61.5**	**81.9**	**81.5**	**77.0**	**75.5**

Teacher limitation. While interleaved reasoning images provide critical information during training, they are unavailable at inference, creating a train–test mismatch that limits the teacher's performance.

3.2.2 Stage 2: Distilling Multimodal Reasoning into the Text Space

To address this limitation, in the second stage we train a student LVLM that reasons entirely in text space, without access to intermediate images, while retaining the visual reasoning learned by the teacher. We freeze the teacher f_θ and distill its next-token logits to train a student f_ϕ to generate textual reasoning from the prompt x and original problem image v_{prob}. The student receives no intermediate reasoning images during training or inference. The teacher additionally conditions on the current reasoning image and preceding interleaved text-image pairs. Both models use the same ground-truth textual prefix under teacher forcing:

$$q_\theta(y_t) = p_\theta\left(y_t \mid x, v_{\text{prob}}, v_{\text{reason},t}, (y_j, v_{\text{reason},j})_{j<t}\right), \qquad \text{Teacher,} \tag{3}$$

$$p_\phi(y_t) = p_\phi\left(y_t \mid x, v_{\text{prob}}, y_{<t}\right), \qquad \text{Student.} \tag{4}$$

Here, y_t denotes the ground-truth text at step t, and $y_{<t}$ denotes the preceding textual CoT steps.

The logits distillation loss is therefore:

$$\mathcal{L}_{\text{KD}}(\phi) = \sum_t \text{KL}\left(q_\theta(y_t) \,\|\, p_\phi(y_t)\right). \qquad \text{Teacher logit distillation} \tag{5}$$

Remark. During distillation, the teacher is conditioned on the ground-truth interleaved reasoning trajectory rather than generated intermediate images, preventing errors from generated visual states from propagating through the distillation process.

We also supervise the student on the ground-truth textual tokens:

$$\mathcal{L}_{\text{text}}(\phi) = -\sum_t \log p_\phi\left(y_t \mid y_{<t}, x, v_{\text{prob}}\right). \qquad \text{Ground-truth supervision} \tag{6}$$

The final V2T objective is as follows:

$$\mathcal{L}_{\text{total}}(\phi) = \mathcal{L}_{\text{text}}(\phi) + \lambda \mathcal{L}_{\text{KD}}(\phi). \tag{7}$$

where λ controls the strength of the distillation signal. We use $\lambda = 0.1$ in our experiments and ablate its effect in Appendix A.5.

3.2.3 Auxiliary: Distilling Visual Grounding Via Attention

When each intermediate reasoning image can be spatially mapped to a region of the original problem image, such as a crop or transformation, V2T can further distill the teacher's attention as an auxiliary

signal. For each reasoning step, we extract attention at the first ground-truth text token following the reasoning image, following prior work on probing visual attention at reasoning-step boundaries and the first generated token (Gao et al., 2025; Chen et al., 2025b; Li et al., 2026). At this position, the reasoning image is fully available without intervening text. We project the teacher's attention onto the corresponding region of the original image and align it with the student's attention using a KL loss. This encourages attention to task-relevant regions and improves visual grounding.

Formally, let $\mathbf{a}_{\theta,t}$ and $\mathbf{a}_{\phi,t}$ denote the teacher's and student's normalized attention weight distribution to the reasoning image and problem image, respectively, at reasoning step t. Let $S \subseteq [K]$ be the reasoning steps for which $v_{\text{reason},t}$ can be spatially mapped to a region of v_{prob}.

We minimize their forward KL divergence:

$$\mathcal{L}_{\text{attn}}(\phi) = \frac{1}{|S|} \sum_{t \in S} \text{KL}\left(\mathbf{a}_{\theta,t} \,\|\, \mathbf{a}_{\phi,t}\right). \tag{8}$$

The V2T training objective becomes:

$$\mathcal{L}^{\text{attn}}_{\text{total}}(\phi) = \mathcal{L}_{\text{text}}(\phi) + \lambda \mathcal{L}_{\text{KD}}(\phi) + \beta \mathcal{L}_{\text{attn}}(\phi), \tag{9}$$

where β controls the strength of attention supervision.

3.2.4 Auxiliary: Reinforcement Visual Grounding

After Stage 2, with or without attention distillation, the V2T student learns to reason without intermediate images while retaining the teacher's visual reasoning signals. Because distillation already provides a strong initialization with effective visual reasoning capabilities, only a short GRPO fine-tuning stage can be optionally applied to further improve the generalization performance. We use standard task rewards and, when ground-truth bounding boxes are available, an additional *visual grounding reward* that encourages the student to attend to task-relevant image regions.

Specifically, for each reasoning step, we extract the student's attention to the problem image to identify the attended region $\widehat{M}$ and compare it with the ground-truth region M. We use this alignment as a visual grounding reward:

$$r_{\text{ground}} = \frac{1}{2}\left(\frac{|\widehat{M} \cap M|}{|\widehat{M}|} + \frac{|\widehat{M} \cap M|}{|M|}\right). \tag{10}$$

This reward balances how much of the attended region overlaps the target region with how much of the target region is covered. The final GRPO reward is:

$$r = r_{\text{task}} + \eta r_{\text{ground}}, \tag{11}$$

where r_{task} combines a binary answer-correctness reward and a format reward weighted by 0.1. We set $\eta = 0.5$; the effect of the visual grounding reward is evaluated in Sec. 4.2.

4 Experiments

Training Datasets. We use the image–text interleaved CoT dataset **Zebra-CoT** (Li et al., 2025a), a curated mixture spanning scientific diagrams, 2D visual puzzles, and visually grounded logic and strategic games. Following Wang et al. (2025d), we apply a two-stage filtering process to remove trivial samples and noisy interleaved images. First, we use Qwen2.5-VL-7B with only the problem image and question, removing samples it answers correctly without interleaved images. Second, we use Qwen2.5-VL-72B with the interleaved images and remove samples it answers incorrectly, filtering out inaccurate images and overly difficult problems. This yields approximately 23K training samples, with 200 samples per category held out for testing.

Implementation Details. For Zebra-CoT, we use Qwen2.5-VL-7B (Bai et al., 2025) as the base model and train both the teacher and the student with LoRA SFT (Hu et al., 2022a). The teacher uses the filtered Zebra-CoT set; the student uses its visual-search subset. We then run full-parameter GRPO (Shao et al., 2024b) on TreeVGR-RL-37K (Wang et al., 2025b), with $\lambda = 0.1$, $\beta = 0.5$, and $\eta = 0.5$. SFT and GRPO hyperparameters are listed in Tables 6 and 7 (Appendix A.2).

Table 2: Results on the broader visual evaluation suite. Best and second-best results are highlighted in **bold** and underline, respectively. V2T achieves an average of 61.8, outperforming the strongest non-V2T baseline by 2.7 points; V2T + GRPO reaches 63.4. Training costs marked with * are estimated from the reported training configurations. V2T trains in 20 GPU-hours (80 with GRPO) at $1\times$ inference, compared with $2\times$ for latent reasoning, $30\times$ for Bagel-Zebra, and $45\times$ for tool-based methods.

Method	SFT	GRPO	Train. (GPU-h)	HR4K	HR8K	V*	BLINK						POPE	Avg
							Func.	Visual	Obj. Loc.	Foren.	Jig.	Multi.		
LVLMs with text-only visual reasoning														*Inference:* $1\times$
R1-VL-7B	✓	✓	167*	52.9	45.2	43.5	22.3	52.3	48.4	44.7	65.3	42.1	90.1	50.6
VLAA-Thinker-7B	✗	✓	167*	50.5	41.4	48.7	22.3	50.0	48.4	47.7	60.0	42.1	86.1	49.7
VL-Rethinker-7B	✓	✓	128	61.0	48.9	64.4	23.9	55.2	46.7	48.5	56.0	47.4	86.4	53.8
Vision-R1-7B	✓	✓	3392	**75.3**	70.5	80.6	29.2	50.6	48.4	50.0	59.3	39.1	88.2	59.1
LVLMs with latent visual reasoning														*Inference:* $2\times$
LVR-7B	✓	✓	160	69.4	63.9	77.5	27.7	49.4	53.3	40.2	50.0	50.4	85.2	56.7
MONET-7B	✓	✓	330*	65.8	55.5	76.4	26.2	55.8	50.8	36.4	55.3	50.4	84.7	55.7
Unified interleaved model														*Inference:* $30\times$
Bagel-Zebra-7B	✓	✗	110*	62.3	51.4	64.4	33.9	**70.9**	**60.7**	26.5	**74.0**	46.6	87.0	57.8
Tool-based visual reasoning														*Inference:* $45\times$
Thyme	✓	✓	1424	66.6	63.1	71.2	32.3	60.5	47.5	**54.6**	64.0	44.4	75.3	58.0
DeepEyes	✗	✓	1536	72.8	69.2	**88.5**	28.5	39.5	47.5	43.2	57.3	52.6	91.8	59.1
Controlled baselines and ours														*Inference:* $1\times$
Qwen2.5-VL-7B	✗	✗	–	68.9	64.1	75.9	24.6	56.4	45.9	47.2	64.7	43.6	86.5	57.8
V2T (Ours)	✓	✗	20	69.3	64.5	80.1	28.5	66.3	51.6	47.7	66.0	52.6	91.8	61.8
V2T + GRPO (Ours)	✓	✓	80	73.5	**71.3**	82.2	**36.9**	59.9	45.1	**54.6**	62.7	**55.6**	**92.0**	**63.4**

Efficiency. V2T requires approximately 20 NVIDIA H100 GPU-hours in total for teacher SFT and student distillation. The optional GRPO stage adds approximately 60 GPU-hours, bringing the total training cost to 80 GPU-hours. As shown in Table 2, even V2T + GRPO uses $17.8\times$, $19.2\times$, and $42.4\times$ fewer GPU-hours than Thyme, DeepEyes, and Vision-R1-7B, respectively. At inference, V2T matches text-only reasoning at $1\times$, compared with $2\times$ for latent visual reasoning, $30\times$ for Bagel-Zebra and $45\times$ for tool-based methods.

Baselines. We compare V2T with text-based reasoning LVLMs (R1-VL-7B (Zhang et al., 2025a), VLAA-Thinker-7B (Chen et al., 2025a), VL-Rethinker-7B (Wang et al., 2025c), Vision-R1-7B (Huang et al., 2025b)), latent visual reasoning (LVR-7B (Li et al., 2025b), MONET-7B (Wang et al., 2025d)), the interleaved model Bagel-Zebra-7B (Li et al., 2025a), and tool-based methods (Thyme (Zhang et al., 2025c), DeepEyes (Zheng et al., 2025)). The controlled baseline trained with our data in Table 2 is Qwen2.5-VL-7B. Text-Only SFT and Interleaved SFT, the teacher, are reported in Table 3. We also report V2T + GRPO. Details are provided in Appendix A.3.

Evaluation Datasets. We evaluate on two complementary suites: a held-out Zebra-CoT set covering Graph Algorithm, RPM, Tetris, and Visual Search, and a broader visual evaluation suite including HR-Bench 4K/8K (Wang et al., 2024), V*Bench (Wu & Xie, 2024), BLINK (Fu et al., 2024), and POPE (Li et al., 2023). We use VLMEvalKit (Duan et al., 2024) with Qwen3-235B-A22B-Instruct-2507-FP8 (Yang et al., 2025a) as the judge model. Evaluation details are provided in Appendix A.3. We additionally conduct a judge-consistency check on selected benchmarks in Appendix D. We set the maximum generation length to 2048 tokens.

4.1 Main Results

Zebra-CoT held-out tasks. Table 1 reports results on the Zebra-CoT held-out subset, covering Graph Algorithm, RPM, Tetris, and Visual Search. V2T achieves the highest average accuracy of 75.5%, outperforming the Interleaved SFT teacher by 14.3% and the strongest existing baseline,

Table 3: Ablation of attention supervision: attention distillation during SFT and the visual grounding reward during GRPO, compared with Text-Only SFT and Interleaved SFT, the teacher. Best and second-best results are highlighted in **bold** and underline, respectively. Evaluation categories follow Table 2.

Method	HR4K	HR8K	V*	BLINK						POPE	Avg
				Func.	Visual	Obj. Loc.	Foren.	Jig.	Multi.		
Text-Only SFT	60.1	47.3	66.0	25.4	60.5	50.0	52.3	63.3	45.1	91.7	56.2
Interleaved SFT	56.8	47.1	62.8	26.9	61.6	48.4	44.7	60.0	48.9	91.6	54.9
V2T (Ours)	69.3	64.5	80.1	28.5	**66.3**	**51.6**	47.7	**66.0**	52.6	91.8	61.8
V2T w/o Attention Distillation	59.0	49.6	63.4	31.5	60.5	46.7	48.5	64.7	**55.6**	91.5	57.1
V2T + GRPO (Ours)	**73.5**	71.3	82.2	**36.9**	59.9	45.1	54.6	62.7	**55.6**	**92.0**	**63.4**
V2T + GRPO w/o Grounding Reward	73.1	**71.4**	**84.3**	20.8	55.2	50.8	**59.1**	58.0	**55.6**	**92.0**	62.0

VLAA-Thinker-7B, by 14.6%. V2T performs best on all four held-out task categories, confirming the effectiveness of our visual distillation into text space.

Broader visual evaluation. Table 2 reports results across high-resolution perception (HR-Bench 4K/8K), visual search (V*Bench), fine-grained perception (BLINK), and hallucination (POPE). Without GRPO, V2T achieves an average score of 61.8%, outperforming the strongest existing visual reasoning models, Vision-R1-7B and DeepEyes, by 2.7%. V2T also exceeds Qwen2.5-VL-7B by 4.0%. Compared with Qwen2.5-VL-7B, V2T improves on every reported task, with the largest gains on BLINK Visual Correspondence (+9.9%), Multi-view Reasoning (+9.0%), Object Localization (+5.7%), and POPE (+5.3%). Comparisons with Text-Only SFT and Interleaved SFT are in Table 3. The results confirm that V2T distillation improves both visual grounding and overall generalization without requiring intermediate visual states at inference.

4.2 Ablation Studies

Attention distillation and visual grounding reward. As shown in Table 3, V2T without attention distillation averages 57.1, above Text-Only SFT (56.2) and Interleaved SFT (54.9). Adding attention distillation improves the average from 57.1% to 61.8%, a gain of 4.7%. The largest improvements occur on HR-Bench 4K (+10.3%), HR-Bench 8K (+14.9%), and V*Bench (+16.7%), indicating that attention distillation is particularly important for high-resolution perception and visual search. During GRPO, the visual grounding reward increases the average from 62.0% to 63.4%. The improvement is especially pronounced on BLINK Functional Correspondence (+16.1%), while performance varies across the other BLINK categories, suggesting that attention grounding and task-reward optimization provide complementary but task-dependent signals.

Hyperparameter Ablations. We study the sensitivity to the logit distillation weight λ, attention supervision weight β, and visual grounding reward weight η. We use $\lambda = 0.1$, $\beta = 0.5$, and $\eta = 0.5$ in our final configuration. Full results are provided in Appendix A.5.

5 Error Analysis

Beyond aggregate accuracy, we analyze whether V2T genuinely internalizes visual grounding and which types of errors it reduces. We study this from two perspectives: attention alignment with ground-truth visual regions on test set, and error distributions across different failure categories. In this section we analyze V2T without the subsequent GRPO training.

Attention Analysis. We compare attention patterns across V2T, the interleaved SFT model evaluated with and without reasoning images, and the text-only model. Following prior analyses showing that cross-modal visual information is primarily processed in the middle Transformer layers (Kaduri et al., 2025; Jiang et al., 2025), we average attention over the middle 8 layers to obtain a more stable estimate of visual grounding. We take the maximum across heads and measure attention IoU between the top-k attended pixels and the ground-truth bounding box, where k matches the bounding-box area. Since attention to visual tokens is strongest in the first few generated tokens and rapidly decreases



(a) Text-Only SFT (b) Interleaved SFT (teacher) (c) V2T (student)

Figure 5: Qualitative comparison of attention maps among Text-Only SFT, Interleaved SFT, and V2T on the Visual Search split of Zebra-CoT. V2T attends more strongly to task-relevant visual regions, suggesting improved visual grounding compared to text-only and Interleaved SFT (teacher) models.

afterward (Ramazzina et al., 2025), we focus on the first text token following each reasoning-image insertion point, consistent with the attention distillation in Section 3.2.3. We evaluate on 100 random examples from Zebra-Visual Search.

Table 4: Attention IoU analysis. (Inf) and (Tra) denote the interleaved model evaluated without and with reasoning images respectively. V2T pays highest attention to the bounding boxes.

Model	Base Model	Text-Only SFT	Interleaved (Inf)	Interleaved (Tra)	V2T (Ours)
IoU	0.191	0.215	0.223	0.168	**0.269**

As shown in Table 4 and Fig. 5, Text-Only SFT only slightly improves over the base model (0.215 vs. 0.191), suggesting that textual CoT supervision alone provides limited visual grounding. The interleaved model evaluated without reasoning images (Inf) reaches 0.223, but when reasoning images are provided (Tra), its attention shifts to the reasoning image (IoU 0.234) and its problem-image IoU drops to 0.168, below the base model, indicating that the reasoning image can act as a shortcut. In contrast, V2T achieves the highest problem-image IoU (0.269), showing that V2T transfers genuine visual grounding into text space rather than collapsing into linguistic pattern matching.

Error Analysis. To understand where V2T improves, we use GPT-4o to classify incorrect predictions into four categories: (1) *Perception and Grounding*, (2) *Spatial and Global Context*, (3) *Reasoning and Knowledge*, and (4) *Task and Answer Consistency*.

Table 5: Error distribution. V2T has the largest relative reduction in Spatial & Global Context errors.

Category	Text-Only SFT	Interleaved SFT	V2T (Ours)
Perception & Grounding	366	382	**364**
Spatial & Global Context	241	230	**193**
Reasoning & Knowledge	941	916	**872**
Task & Answer Consistency	199	209	**179**
Total	1747	1737	**1608**

As shown in Table 5, V2T reduces total errors by 7.4% relative to Interleaved SFT, with the largest reduction in Spatial and Global Context errors (-16.1%) and Task and Answer Consistency errors (-14.4%). This is consistent with the attention analysis: stronger grounding to the original problem image particularly improves scene-level spatial reasoning. Perception and Grounding errors remain similar across methods, while Reasoning and Knowledge errors still dominate, suggesting that V2T primarily improves visually grounded reasoning rather than basic perception or domain knowledge.

6 Conclusion

We propose V2T, a training framework that uses interleaved visual information as training-time supervision and distills its visual grounding into text-based chain-of-thought reasoning. On held-out and broader visual benchmarks, V2T outperforms text-only reasoning models, latent visual reasoning approaches, and interleaved models, including its teacher. These results show that visual reasoning can be internalized through distillation: at inference, V2T reasons entirely in text at the same $1\times$ cost as text-only models, without intermediate visual representations or their error propagation.

AI USE STATEMENT

We used generative AI tools solely to improve the manuscript's writing, including grammar, phrasing, and clarity. All AI-assisted text was carefully reviewed and edited by the authors. Generative AI was not used to develop research ideas, formulate theoretical claims, or produce experimental results.

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.

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. *arXiv preprint arXiv:2504.11468*, 2025a.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025b.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025c.

Shuai Dong, Siyuan Wang, Xingyu Liu, and Zhongyu Wei. Interleaved latent visual reasoning with selective perceptual modeling. *arXiv preprint arXiv:2512.05665*, 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.

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.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19520–19529. IEEE, 2025.

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

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022b.

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, 2024a.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024b.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Weiming Zhang, and Nenghai Yu. Deciphering cross-modal alignment in large vision-language models via modality integration rate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 218–227, 2025a.

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*, 2025b.

Zeyi Huang, Yuyang Ji, Anirudh Sundara Rajan, Zefan Cai, Wen Xiao, Haohan Wang, Junjie Hu, and Yong Jae Lee. Visualtoolagent (vista): A reinforcement learning framework for visual tool selection. *arXiv preprint arXiv:2505.20289*, 2025c.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14558, 2025.

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.

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.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026.

Xuchen Li, Xuzhao Li, Jiahui Gao, Renjie Pi, Shiyu Hu, and Wentao Zhang. Look less, reason more: Rollout-guided adaptive pixel-space reasoning. *arXiv preprint arXiv:2510.01681*, 2025d.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023.

Zujing Liu, Junwen Pan, Qi She, Yuan Gao, and Guisong Xia. On the faithfulness of visual thinking: Measurement and enhancement. *arXiv preprint arXiv:2510.23482*, 2025.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*, 2024.

Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *Advances in Neural Information Processing Systems*, 37:111863–111898, 2024.

Andrea Ramazzina, Tobias Haab, David Fitzek, Stefano Gasperini, Jonas Uhrig, and Mario Bijelic. Reasoning with fewer eyes: Efficient visual token withdrawal for multimodal reasoning. In *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 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.

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*, 2025a.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, et al. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*, 2025b.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. MM-MATH: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1358–1375, 2024.

Zhenwei Tang, Difan Jiao, Blair Yang, and Ashton Anderson. SEAM: Semantically equivalent across modalities benchmark for vision-language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=lI4LgGv4sX.

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.

Changpeng Wang, Haozhe Wang, Xi Chen, Junhan Liu, Taofeng Xue, Chong Peng, Donglian Qi, Fangzhen Lin, and Yunfeng Yan. From illusion to intention: Visual rationale learning for vision-language reasoning. *arXiv preprint arXiv:2511.23031*, 2025a.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025b.

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*, 2025c.

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*, 2025d.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint arXiv:2408.15556*, 2024.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

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. *arXiv preprint arXiv:2505.19255*, 2025.

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.

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, 2025a.

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*, 2025b.

Yige Xu, Yongjie Wang, Zizhuo Wu, Kaisong Song, Jun Lin, and Zhiqi Shen. Do vision-language models truly perform vision reasoning? a rigorous study of the modality gap. *arXiv preprint arXiv:2604.16256*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

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*, 2025a.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, 2025b.

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*, 2025c.

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.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, 2025.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none that we feel must be specifically highlighted here.

Limitations. While V2T achieves strong gains, several limitations remain. First, it requires interleaved multimodal reasoning data with paired intermediate images, which is expensive to collect and limited in scale. Second, the student inherits any biases or errors in the teacher's visually-grounded reasoning. Third, our error analysis shows that text-space distillation primarily improves spatial reasoning, with smaller gains on perception-level and domain-knowledge errors, indicating it cannot substitute for stronger underlying visual perception. Finally, due to computational constraints, all experiments use a single training run; multi-seed evaluation and larger model scales are left to future work.

A Experimental Details

A.1 SFT Hyperparameters

We list the SFT hyperparameters in Table 6. We use LoRA-based fine-tuning with rank $r = 32$ and $\alpha = 64$ on four NVIDIA H100 80GB GPUs. The overall batch size is 64 with four gradient-accumulation steps. To avoid out-of-memory errors on samples containing many interleaved images, we constrain each training sequence to between $512 \times 28 \times 28$ and $1280 \times 28 \times 28$ image pixels.

Table 6: Hyperparameters for SFT.

Hyperparameter	**Value**
LoRA rank	32
LoRA alpha	64
LoRA target modules	All
Learning rate	1.0×10^{-4}
Per-device batch size	4
Gradient accumulation steps	4
Training epochs	2
Learning-rate scheduler	Cosine
Warmup ratio	0.1
Maximum image pixels	$1280 \times 28 \times 28$
Minimum image pixels	$512 \times 28 \times 28$
Distillation coefficient λ	0.1
Attention supervision coefficient β	0.5

A.2 GRPO Hyperparameters

The optional GRPO stage starts from the distilled V2T student and performs full-parameter training on TreeVGR-RL-37K (Wang et al., 2025b). For each prompt, we sample five rollouts. The reward is $r = r_{\text{task}} + \eta r_{\text{ground}}$ (Section 3.2.4), where r_{task} combines a binary answer-correctness reward with a format reward weighted by 0.1, and $\eta = 0.5$. We train for 160 steps with EasyR1 on four NVIDIA H100 GPUs, which takes approximately 60 GPU-hours. Table 7 summarizes the settings.

Table 7: Hyperparameters for GRPO.

Hyperparameter	Value
Initialization	V2T student
Training data	TreeVGR-RL-37K
Fine-tuning	Full-parameter
Learning rate	1.0×10^{-6}
Learning-rate scheduler	Constant
KL coefficient (low-variance KL)	0.01
Rollout batch size (prompts)	512
Global batch size	128
PPO epochs	1
Training steps	160
Rollouts per prompt	5
Rollout temperature	1.0
Maximum prompt / response length	2048 / 2048
Format reward weight	0.1
Visual grounding reward weight η	0.5

A.3 Evaluation and Baseline Details

Evaluation Datasets. We use two complementary evaluation suites. First, we evaluate on a held-out subset of Zebra-CoT comprising 200 randomly sampled examples from each of Graph Algorithm, RPM, Tetris, and Visual Search (800 in total). Second, we evaluate broader visual generalization on HR-Bench 4K and 8K (Wang et al., 2024), V*Bench (Wu & Xie, 2024), BLINK (Fu et al., 2024), and POPE (Li et al., 2023). HR-Bench evaluates fine-grained perception at high image resolutions, while V*Bench measures visual search and spatial grounding. From BLINK, we report Functional Correspondence, Visual Correspondence, Object Localization, Forensic Detection, Jigsaw, and Multi-view Reasoning. POPE evaluates object hallucination. Together, these benchmarks cover in-domain multimodal reasoning as well as out-of-domain high-resolution perception, visual correspondence, localization, compositional and multi-view reasoning, and hallucination.

We adopt VLMEvalKit (Duan et al., 2024) for evaluation and use Qwen3-235B-A22B-Instruct-2507-FP8 (Yang et al., 2025a) as the judge model.

Baselines. All baselines are 7B-scale models. We group them by how they incorporate visual information during reasoning. (1) *Text-only visual reasoning*: R1-VL-7B (Zhang et al., 2025a), VLAA-Thinker-7B (Chen et al., 2025a), VL-Rethinker-7B (Wang et al., 2025c), and Vision-R1-7B (Huang et al., 2025b), which generate textual CoTs over the original image after SFT and/or RL post-training. (2) *Latent visual reasoning*: LVR-7B (Li et al., 2025b) and MONET-7B (Wang et al., 2025d), which interleave learned latent visual tokens with textual reasoning. (3) *Unified interleaved generation*: Bagel-Zebra-7B (Li et al., 2025a), which generates explicit intermediate images and is trained on the full Zebra-CoT dataset; we therefore exclude it from the Zebra-CoT held-out evaluation. (4) *Tool-based visual reasoning*: Thyme (Zhang et al., 2025c) and DeepEyes (Zheng et al., 2025), which operate on the image through tool calls (e.g., cropping) during reasoning. To match our single-pass inference setting, we reduce Thyme's default five answering attempts to a single attempt. Qwen2.5-VL-7B (Bai et al., 2025) serves as the controlled base model for Text-Only SFT, Interleaved SFT (the teacher), and V2T. For all baselines, we use the official checkpoints and the VLMEvalKit evaluation pipeline. Bagel-Zebra-7B is served with vLLM, which supports only its image-understanding pathway, so it is evaluated with text-only reasoning. Note that R1-VL-7B is built on Qwen2-VL rather than Qwen2.5-VL.

Training and inference cost. Training costs in Table 2 are taken from the GPU-hours reported by each method; entries marked with * are estimated from the reported training configurations and public implementations. Tool-based methods also use more expensive RL: for example, DeepEyes (1536 GPU-hours) samples 16 rollouts per prompt with repeated image cropping and re-encoding, compared with five text-only rollouts in our GRPO stage. Thyme and DeepEyes also train on V* data, which shares its distribution with V*Bench, whereas V2T is trained on Zebra-CoT without comparable V* examples. Inference cost is normalized to standard text-only autoregressive reasoning.

A.4 Filtered Zebra-CoT Composition

The filtered training set contains 23,260 examples. Table 8 reports its composition by source subset.

Table 8: Composition of the filtered Zebra-CoT training set.

Subset	Percentage
Visual logic and strategic games – ARC-AGI-New	0.056%
Scientific reasoning – Geometry	1.045%
Scientific reasoning – Competitive Programming	1.040%
Visual logic and strategic games – RPM	2.266%
Visual logic and strategic games – ARC-AGI	0.056%
3D visual reasoning – Embodied CoT	0.172%
3D visual reasoning – Multi-hop Object Counting	5.864%
Scientific reasoning – Chemistry	12.455%
2D visual reasoning – Visual Search	33.780%
Visual logic and strategic games – Maze	16.062%
Scientific reasoning – Physics	0.052%
3D visual reasoning – Robot Planning	0.254%
Visual logic and strategic games – Tetris	2.171%
Scientific reasoning – Graph Algorithms	8.060%
Visual logic and strategic games – Ciphers	0.080%
Visual logic and strategic games – Chess	15.900%
Visual logic and strategic games – Connect Four	0.688%

A.5 Hyperparameter Ablations

We study the effect of three loss weights: the logit distillation weight λ, the attention supervision weight β, and the visual grounding reward weight η.

For V2T, we jointly vary λ and β over $\{0.1, 0.5\}$. As shown in Figure 6, the best performance is obtained with $\lambda = 0.1$ and $\beta = 0.5$, achieving an average accuracy of 61.8%. The other configurations, $(\lambda, \beta) = (0.1, 0.1)$, $(0.5, 0.1)$, and $(0.5, 0.5)$, achieve 56.2%, 56.7%, and 56.5%, respectively.

For V2T+GRPO, we further vary the visual grounding reward weight η over $\{0.1, 0.5\}$. $\eta = 0.5$ achieves the best average accuracy of 63.4%, compared with 63.0% for $\eta = 0.1$; we use $\eta = 0.5$ in our final configuration.

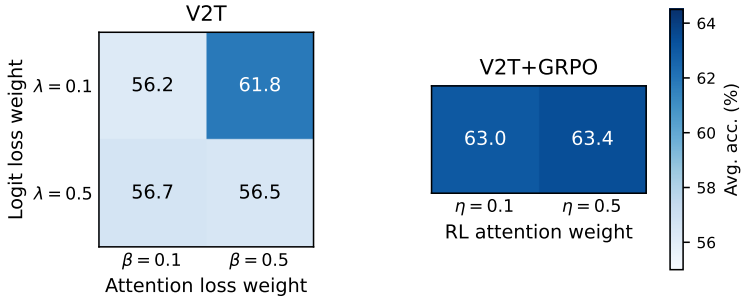


Figure 6: Hyperparameter ablations for V2T and V2T+GRPO. Left: average accuracy of V2T when varying the logit distillation weight λ and the attention supervision weight β. Right: average accuracy of V2T+GRPO when varying the visual grounding reward weight η.

B Controlled Analysis of Interleaved Visual Reasoning

The motivation for V2T is that intermediate visual states can provide useful supervision during training, while repeatedly conditioning the autoregressive textual reasoning process on additional

visual representations can also introduce interference. We therefore separate two effects that are closely related but conceptually distinct. First, as more intermediate reasoning images are inserted into a chain of thought, the loss of subsequent textual reasoning can increase. Second, visual tokens need not be equally relevant to the current reasoning step: background regions and other unmatched visual content can further interfere with textual reasoning even when the useful visual content is unchanged. The latter effect is consistent with prior observations that imperfect cross-modal alignment remains a challenge for vision–language models (Tang et al., 2025; Xu et al., 2026; Huang et al., 2025a).

Modality Switching and Number of Interleaved Images. We first compare the per-token loss of textual reasoning under two conditions: (1) the textual CoT is conditioned on its interleaved reasoning images, and (2) the same textual CoT is evaluated without the interleaved reasoning images. We conduct this analysis on held-out Zebra-CoT examples from Checkers, Connect Four, and Object Counting. Across all three categories, the loss gap becomes substantially larger near the end of the reasoning trajectory. At 25%, 50%, 75%, and 100% of the CoT, the loss gaps are 0.02, 1.09, 2.68, and 4.53 for Checkers; 0.11, 0.48, 1.77, and 2.29 for Connect Four; and 0.10, 0.18, 3.02, and 3.91 for Object Counting. We observe the same trend when varying the number of interleaved reasoning images directly: the mean loss gap increases from $+0.38$ with three images to $+2.19$, $+5.15$, and $+5.99$ with five, seven, and nine images, respectively. These results provide direct empirical evidence that repeatedly introducing intermediate visual states can increasingly disrupt subsequent textual continuation even when the images contain reasoning-relevant information.

Irrelevant Visual Content. We next isolate a second source of interference: visual content that is not useful for the corresponding textual reasoning. For each reasoning image, we place the original image on a larger gray canvas so that the useful reasoning content is unchanged while the amount of irrelevant visual content is increased by 50%. We then compare the textual-CoT loss against the same text-only condition used above. The loss gap increases for all three categories: from 4.53 to 5.12 for Checkers, from 3.91 to 9.90 for Object Counting, and from 2.29 to 3.50 for Connect Four. This shows that visually unmatched content can further interfere with textual reasoning even when the useful visual information is preserved.

Generation-Level Prefix Analysis. To verify that these effects are reflected in generation quality rather than only token-level loss, we conduct an additional experiment in which the model is given the first 3, 5, or 7 steps of a ground-truth multimodal reasoning trajectory and then generates the remaining textual reasoning and final answer. We compare three conditions: retaining the original reasoning images, removing the reasoning images while keeping exactly the same textual CoT context, and adding 50% gray background to the original reasoning images. For Object Counting, the accuracies for With Image / No Image / +50% Gray are 64/56/46 with three images, 60/62/56 with five images, and 60/66/58 with seven images. For Connect Four, the corresponding results are 8/4/4, 12/12/10, and 14/16/10.

The generation results reveal two complementary trends. When only a few reasoning images are present, the visual states can still provide useful information. However, as the number of interleaved images grows, retaining the same textual reasoning while removing the intermediate images matches or outperforms the interleaved condition. Moreover, increasing visually unmatched content consistently reduces accuracy. Together with the loss analysis, these experiments motivate V2T: intermediate images are valuable as training-time supervision, but their useful reasoning signal can be distilled into text space so that inference does not require repeated visual conditioning or carry the additional interference introduced by irrelevant visual tokens.

C Baseline Sanity Check

C.1 Math-Focused Reasoning Models

We additionally examine VLAA-Thinker-7B on the mathematical reasoning benchmarks emphasized by its training setup. As shown in Table 9, VLAA-Thinker-7B consistently exceeds the Qwen2.5-VL-7B base model on MathVerse (Zhang et al., 2024), MathVista (Lu et al., 2023), MM-Math (Sun et al., 2024), and DynaMath (Zou et al., 2025). This confirms that the lower performance observed for VLAA-Thinker-7B on our broader visual-reasoning suite is not due to a generally weak checkpoint.

Instead, the results suggest that its math-focused training improves mathematical reasoning while transferring less strongly to the more general visual tasks considered in our evaluation.

Table 9: Performance of Qwen2.5-VL-7B and VLAA-Thinker-7B on math-focused reasoning benchmarks.

Model	MathVerse	MathVista	MM-Math	DynaMath
Qwen2.5-VL-7B	43.9	69.1	41.1	56.5
VLAA-Thinker-7B	**45.8**	**71.4**	**43.1**	**57.3**

D Judge Consistency Analysis

To verify that our evaluation is not sensitive to the choice of judge model, we additionally re-evaluate the same outputs on selected benchmarks using GPT-4o (OpenAI, 2024) and compare the results with our Qwen3-235B-A22B-Instruct-2507-FP8 judge. Across HR-Bench 4K, HR-Bench 8K, and V*Bench, the two judges produce nearly identical scores, with a maximum difference of 0.5 percentage points.

Table 10: Judge-consistency comparison on selected benchmarks. GPT-4o is used only for this cross-check; Qwen3-235B-A22B-Instruct-2507-FP8 remains the judge used in our main evaluation.

Model	Judge	HR-Bench 4K	HR-Bench 8K	V*Bench
Qwen2.5-VL-7B	Qwen3-235B	68.9	64.1	75.9
Qwen2.5-VL-7B	GPT-4o	68.9	64.0	75.9
V2T + GRPO	Qwen3-235B	73.5	71.3	82.2
V2T + GRPO	GPT-4o	73.1	71.2	81.7

These results show that the reported performance is stable across the two judge models on the evaluated subset.

E Qualitative Comparisons

E.1 Traffic-Light Recognition



Question: What color is the nearest set of traffic lights?
Options: A. Red, B. Green, C. Yellow
Answer (GT): A
Base Model Problem Category: object_presence_type_or_motion_misread
Interleaved Model Problem Category: object_presence_type_or_motion_misread

Base model generation

B. Green

Interleaved SFT generation

THOUGHT 0: To determine the color of the traffic lights, I need to locate them within the provided image.
THOUGHT 1: I will highlight the relevant area containing the traffic lights to focus my analysis.
THOUGHT 2: The highlighted region clearly shows the traffic lights and their illuminated state.
THOUGHT 3: Observing the illuminated lights in the highlighted area reveals their color.

The final answer is: B. Green

V2T generation

THOUGHT 0: To determine the color of the traffic lights, I need to locate them within the provided image.
THOUGHT 1: The traffic lights are small and distant in the original image. To clearly identify their color, I will generate a zoomed-in view of this specific area.
THOUGHT 2: The zoomed-in image clearly shows the traffic lights. Upon closer inspection, they appear to be illuminated in red.
THOUGHT 3: Based on the visual evidence from the magnified section, the traffic lights are red.

The final answer is: A. Red

Figure 7: **Qualitative comparison on a traffic-light recognition example.** The top panel shows the original question image. The base model predicts *Green*. The interleaved SFT model also predicts *Green* after performing step-wise reasoning over the scene. In contrast, V2T focuses on the distant traffic lights and correctly predicts *Red*, matching the ground-truth answer. Phrases such as "zoomed-in view" are inherited from the teacher's interleaved CoT format; no intermediate image is generated or provided, and V2T reasons only over the original image.

E.2 Flower Counting



Question: How many white flowers are in this image?
Options:
A. There are no white flowers in this image.
B. There is 1 white flower in this image.
C. There are 2 white flowers in this image.
Answer (GT): C
Base Model Problem Category: `counting_or_quantity_error`
Interleaved Model Problem Category: `counting_or_quantity_error`

Base model generation

B. There is 1 white flower in this image.

Interleaved SFT generation

THOUGHT 0: I need to carefully examine the provided image to identify and count all visible white flowers.
THOUGHT 1: To accurately determine the number of white flowers, I will focus on the central floral arrangement in the image.
THOUGHT 2: Upon close inspection of the main flower cluster, it is clear that there is only one prominent white flower present.
THOUGHT 3: Based on this observation, the total count of white flowers in the image is one.

The final answer is: B

V2T generation

THOUGHT 0: I need to examine the provided image to identify all visible white flowers.
THOUGHT 1: To clearly discern the number of white flowers, I will focus on the central floral arrangement.
THOUGHT 2: The zoomed-in view reveals two distinct white flowers, both appearing to be chrysanthemums.
THOUGHT 3: Based on this clear observation, I can now determine the total count of white flowers.

The final answer is: C. There are 2 white flowers in this image.

Figure 8: **Qualitative comparison on a counting example.** The question image contains two visible white flowers in the central arrangement. The base model and the interleaved SFT model both under-count the flowers and predict only one white flower. In contrast, V2T focuses on the central floral arrangement, identifies two distinct white flowers, and correctly predicts the ground-truth answer. Phrases such as "zoomed-in view" are inherited from the teacher's interleaved CoT format; no intermediate image is generated or provided, and V2T reasons only over the original image.

F Interleaved CoT Examples

Figures 9 and 10 show complete image–text interleaved reasoning examples from the chemistry and chess subsets of Zebra-CoT.

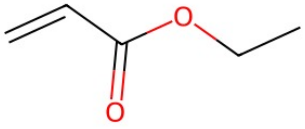
Zebra-CoT Example 1:

Question: What is the main organic product when `ethyl prop-2-enoate (C=CC(=O)OCC)` and `4-bromo-2-chloro-1-(trifluoromethyl)benzene (FC(F)(F)c1ccc(Br)cc1Cl)` are combined?

Image-text Interleaved CoT:

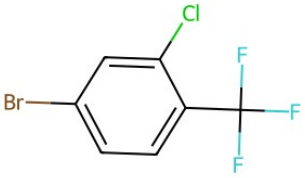
THOUGHT 0: The problem asks for the main organic product from the reaction of ethyl prop-2-enoate (an alkene derivative) and 4-bromo-2-chloro-1-(trifluoromethyl)benzene (an aryl halide). This combination suggests a palladium-catalyzed Heck reaction.

THOUGHT 1: To understand the reaction, first visualize the structure of ethyl prop-2-enoate based on its SMILES `C=CC(=O)OCC`.



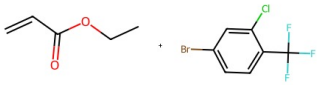
THOUGHT 2: The image shows ethyl prop-2-enoate, an α, β-unsaturated ester (ethyl acrylate). The key reactive site is the terminal C=C bond.

THOUGHT 3: Next, visualize the structure of 4-bromo-2-chloro-1-(trifluoromethyl)benzene from `FC(F)(F)c1ccc(Br)cc1Cl`.



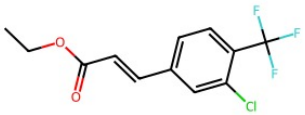
THOUGHT 4: The second reactant is an aryl halide containing both Br and Cl, plus a trifluoromethyl group.

THOUGHT 5: In a Heck coupling, oxidative addition typically occurs at C–Br rather than C–Cl. Thus, coupling happens at the bromine position. The next image shows both reactants.



THOUGHT 6: The aryl group (2-chloro-4-(trifluoromethyl)phenyl) couples to ethyl acrylate. Regioselectivity places the aryl at the terminal carbon of the alkene.

THOUGHT 7: After β-hydride elimination, the product is the substituted alkene, typically with trans (E) stereochemistry. Visualize the final product.



THOUGHT 8: The final image depicts the coupled product with a new C=C bond in the acrylate side chain, typically (E). The final answer is: **ethyl (E)-3-(2-chloro-4-(trifluoromethyl)phenyl)prop-2-enoate**.

Figure 9: Generating intermediate visual states and interleaving grounded reasoning for chemistry.

Zebra-CoT Example 2:

Question: What is the best move for Black to play?
A: Bf5+ B: h6 C: Bb4 D: Be6

Question Image:

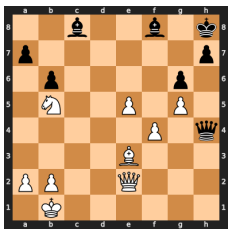
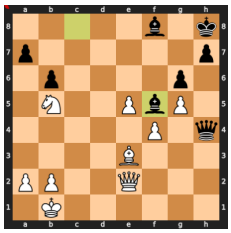
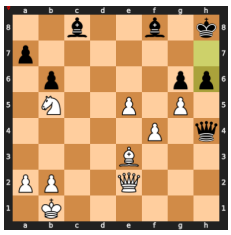


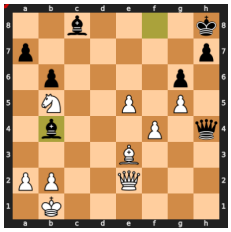
Image-text Interleaved CoT:
THOUGHT 0: Analyze the initial position shown in the diagram. Black needs to find the best move among the given options.
THOUGHT 1: Consider option A, Bf5+. Visualize the position after this move.



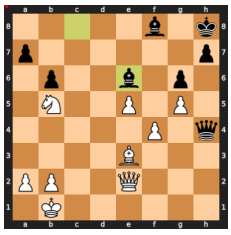
THOUGHT 2: The diagram shows the position after Black plays Bf5.
THOUGHT 3: Consider option B, h6. Visualize the position after this move.



THOUGHT 4: The diagram shows the position after Black plays h6.
THOUGHT 5: Consider option C, Bb4. Visualize the position after this move.



THOUGHT 6: The diagram shows the position after Black plays Bb4.
THOUGHT 7: Consider option D, Be6. Visualize the position after this move.



THOUGHT 8: The diagram shows the position after Black plays Be6.
THOUGHT 9: Based on the evaluation of these options, the best move for Black is Bf5+ (Option A).
The final answer is: A

Figure 10: Generating intermediate visual states and interleaving grounded reasoning for chess.