CLUEJIGSAW: LEARNING TO REASON WITH VISUAL CLUES THROUGH JIGSAW RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual reasoning is essential for advancing vision-language models (VLMs) toward a deeper understanding of the visual world. However, perceiving and interpreting visual inputs remains challenging when familiar spatial structures are disrupted, as exemplified by jigsaw reconstruction. We find that successful reconstruction benefits from the joint use of global semantics and local structural clues. Motivated by this finding, **ClueJigsaw** is proposed as a post-training framework for clue-aware jigsaw reconstruction. We introduce a clue-guided trajectory construction method and use SFT cold-start training to develop models' ability to independently infer semantic clues and structural clues from shuffled inputs. Clue-aware reinforcement learning is then introduced through an easy-to-hard curriculum of reconstruction subtasks to progressively improve clue identification and utilization. We design a two-part reward that combines answer correctness with VLM-based assessment of clue use, additionally rewarding reasoning grounded in appropriate visual clues. Experiments show substantial improvements on both 2×2 and 3×3 jigsaw tasks, with 2×2 reconstruction accuracy rising from 10.12% to 97.38%, alongside gains on general visual reasoning benchmarks. These results highlight the value of visual clue supervision and verification in VLM post-training, helping models connect visual understanding with clue-grounded reasoning.

1 INTRODUCTION

Vision-language models (VLMs) have made substantial progress across a broad range of visual understanding and reasoning tasks (Liu et al., 2023; Bai et al., 2025). Building on these advances, deeper visual understanding requires complex spatial reasoning about how visual elements relate to one another within a coherent scene. Jigsaw reconstruction tests this ability by requiring models to recover an image's original layout from shuffled fragments. When familiar spatial relationships are disrupted, VLMs' reasoning performance declines substantially (Lyu et al., 2025; Zeng et al., 2026).

Recent work has explored jigsaw reconstruction as both an evaluation benchmark and a post-training task. Jigsaw-Puzzles introduces a benchmark for assessing spatial perception, structural understanding, and reasoning (Lyu et al., 2025), while AGILE enhances reconstruction through interaction trajectories, executable actions, and visual feedback (Zeng et al., 2026). Despite these advances, the role of visual clues in supporting successful reconstruction reasoning remains insufficiently understood. We investigate this role through a controlled experiment with Qwen3-VL-8b. The model achieves only 10.12% accuracy on 2×2 puzzles. The example in Figure 1 illustrates how its verbal reasoning can fail to support correct fragment placement, motivating us to examine whether visual clues improve reconstruction.

We distinguish two complementary types of visual clues based on the information they provide: *semantic clues* and *structural clues*. Semantic clues describe the overall scene, recognizable objects, and their contextual roles, helping identify what individual fragments depict and how they fit into the image as a whole. Structural clues capture spatial relationships and visual continuity across fragments, including relative positions, connected object contours, text continuity, and consistent colors and textures along patch boundaries. These clues help distinguish plausible neighboring patches and constrain their relative placement. Together, semantic clues provide scene-level context, while structural clues supply local spatial constraints for coherent reconstruction. We provide externally

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.

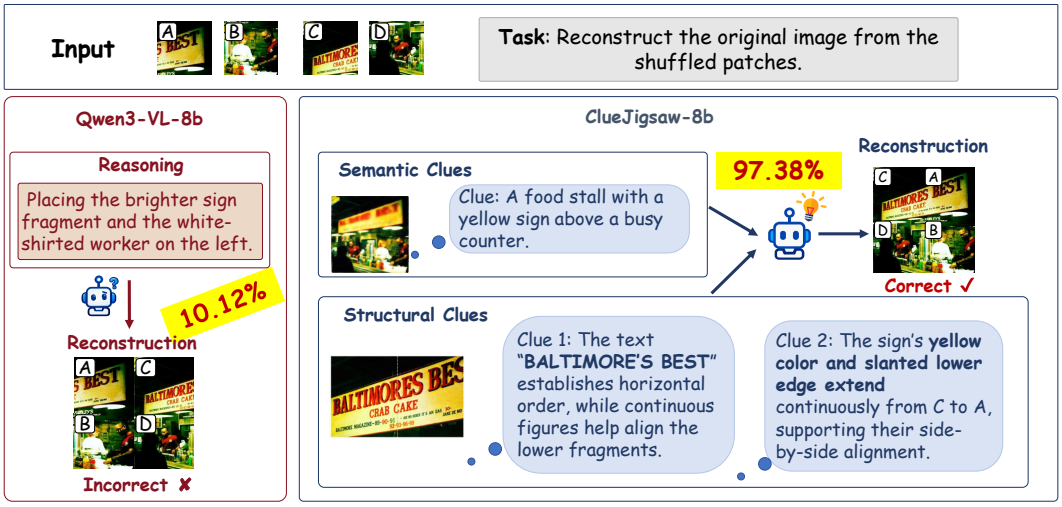


Figure 1: Qwen3-VL-8b produces an incorrect reconstruction through flawed reasoning. ClueJigsaw-8b combines semantic clues about the food stall with structural clues from text, color, and edge continuity to correctly reconstruct the image.

generated textual clues of both types as additional input to the same base model. This guidance raises reconstruction accuracy to 41.86%, a gain of over 30 percentage points, highlighting the value of explicit visual clues.

Motivated by this observation, we introduce **ClueJigsaw**, a post-training framework that teaches models to infer and use semantic clues and structural clues for jigsaw reconstruction. Supervised fine-tuning (SFT) cold-start training, followed by clue-aware reinforcement learning (RL), develops this capability without requiring intact images or externally supplied clues at inference time. The resulting **ClueJigsaw-8B** achieves 97.38% accuracy on 2×2 puzzles. Figure 1 illustrates how the model combines scene semantics with text, color, and edge continuity to recover the correct layout. We evaluate reconstruction accuracy, the effectiveness of clue utilization, and transfer to general visual reasoning tasks. Our contributions are threefold:

- **Clue-guided data construction.** We develop a pipeline that uses semantic clues and structural clues derived from intact images to generate reconstruction reasoning tailored to shuffled inputs. Answer verification and filtering by a VLM-based reasoning judge help ensure that reconstruction decisions are grounded in visual clues.
- **Clue-aware reinforcement learning.** We introduce an easy-to-hard curriculum of reconstruction subtasks with a two-part reward combining verifiable answer correctness and VLM-based assessment of clue use. Together, the curriculum and reward encourage models to select and integrate appropriate clues for decisions ranging from local relations to complete reconstruction.
- **Visual guidance as a post-training approach.** We propose a post-training approach that uses jigsaw reconstruction to guide model reasoning with visual clues, yielding improvements on general visual reasoning tasks beyond reconstruction.

2 METHOD

As shown in Figure 2, ClueJigsaw consists of two stages. In the first stage (a), we introduce a clue-guided trajectory construction pipeline that generates reconstruction demonstrations from semantic clues and structural clues, verifies both answers and reasoning, and uses the retained trajectories for supervised fine-tuning. In the second stage (b), we introduce clue-aware reinforcement learning, combining an easy-to-hard curriculum of reconstruction subtasks with a two-part reward that jointly evaluates task correctness and the quality of clue use.

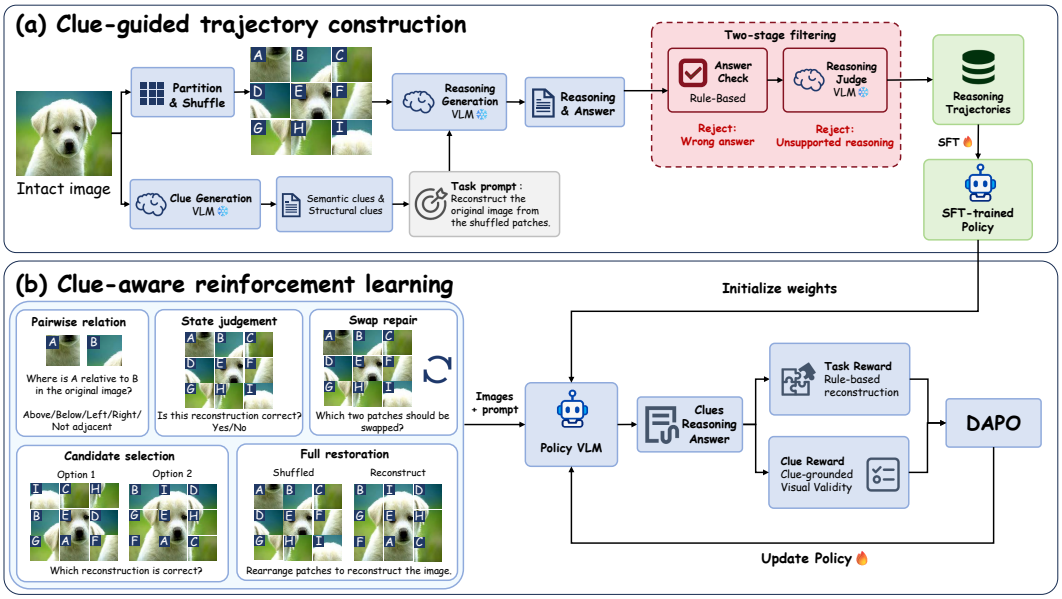


Figure 2: Overview of ClueJigsaw. (a) Clue-guided trajectory construction and two-stage filtering for SFT cold-start training. (b) Clue-aware RL with a reconstruction curriculum and a two-part reward.

2.1 Clue-Guided Trajectory Construction

We construct reasoning trajectories that link visual clues to reconstruction decisions. The pipeline in Figure 2(a) generates visual guidance, applies it to shuffled inputs, and filters the resulting reasoning before SFT.

Trajectory generation and verification. As shown in Fig. 2(a), an intact image I is first fed to the *clue-generation* VLM F to obtain semantic clues d and structural clues C for reconstruction, as shown in Prompt 1. The image is then divided into a $g \times g$ grid of $n = g^2$ patches and shuffled according to a permutation σ, yielding $X_\sigma = (x_1, \ldots, x_n)$. Here, x_i denotes the patch at the i-th position in the shuffled input. Each patch is assigned a visible identifier (e.g., A, B, C), which the model uses to reference patches in its reasoning and final answer. The target ordering a^* is recorded during construction. Given X_σ and the generated guidance d and C, the *reasoning-generation* VLM G produces clue-grounded reconstruction reasoning P and a final answer a:

$$(d, C) = F(I), \qquad (P, a) = G(X_\sigma, d, C). \tag{1}$$

The reasoning links semantic clues and structural clues to the displayed patch labels, explaining how they support spatial judgements and reconstruction decisions. For example, in the puppy image, semantic clues help identify fragments depicting the face and chest. Structural clues, such as a continuous neck contour and matching fur patterns across patch boundaries, then support placing the chest fragment directly below the face fragment.

Prompt 1: Semantic and Structural Clue Generation

System Prompt
You receive one intact image. Identify semantic clues and structural clues that could help reconstruct this image after it is divided into patches and shuffled. Ground every observation in the visible image. Do not invent hidden objects or use file names as evidence.

Write two complementary types of clues:
Semantic clues describe the overall scene, recognizable objects, and their contextual roles. Identify distinctive objects and object parts that could help interpret individual fragments and relate them to the image as a whole.

Structural clues describe observable spatial relationships and visual continuity within the image, including relative object positions, connected contours, text continuation, color transitions, matching textures, and lighting patterns. Explain how these features could help connect fragments or constrain their relative placement during reconstruction. Refer to visible objects and image regions rather than tile labels, since the shuffled patches are not provided.

Prioritize image-specific evidence that could distinguish plausible fragment relationships. Separate direct observations from uncertain interpretations. Incorporate scene descriptions into semantic clues rather than producing a separate description section.

Use this format:

```
<clues>
<semantic>Scene-level and object-level semantic clues.</
    semantic>
<structural>Spatial relationships and visual continuity that
    could support reconstruction.</structural>
</clues>
```

Output only the `<clues>` section. Do not assign tile labels or specify a patch permutation. Do not output separate process, verification, description, or answer sections.

User Prompt

`<image>`

Identify semantic clues and structural clues from this intact image that would help reconstruct it from shuffled fragments.

Different shufflings of the same image produce different patch orders and label assignments. For each shuffled input, the reasoning process maps the image-level clues to the current patch labels and derives the corresponding placement decisions, teaching the model to adapt clue use to each input rather than reuse a fixed explanation. This encourages the model to ground its reasoning in the current arrangement rather than memorize a fixed label-to-position mapping.

We apply two-stage filtering to retain trajectories with correct answers and visually supported reasoning, starting with a rule-based *Answer Check* that removes incorrect reconstructions. The VLM-based *Reasoning Judge* then assesses whether the remaining trajectories use valid visual clues, reference patches consistently, and justify their reconstruction decisions. Trajectories with correct answers are still rejected when their reasoning relies on fabricated observations or unsupported inferences.

Supervised fine-tuning. The SFT data comprise 95% complete reconstruction trajectories and 5% answer-only examples. Each complete trajectory includes semantic clues, structural clues, reconstruction reasoning, and a final answer, conditioned on shuffled patches and a task prompt. The answer-only examples reinforce explicit answer generation, discouraging responses that provide reasoning without a final prediction. No intact reference image is provided to the policy; visual guidance is learned as part of the output rather than supplied as input. We fine-tune the policy on this mixture using standard next-token prediction.

2.2 CLUE-AWARE REINFORCEMENT LEARNING

Clue-aware reinforcement learning trains the model to select and integrate visual guidance through complementary reconstruction subtasks, with task and clue rewards jointly assessing answer correctness and the quality of clue use, as shown in Figure 2(b).

Clue-aware task curriculum. We introduce four auxiliary subtasks to develop complementary clue-use abilities for the target task of full jigsaw reconstruction. These subtasks focus on relation inference, consistency checking, corrective reasoning, and candidate comparison, with verifiable targets derived from recorded patch geometry.

- **Pairwise relation.** Predict the immediate spatial relation or non-adjacency between two patches, learning to use visual clues for local relational judgements.

- **State judgement.** Determine whether a proposed layout is correct by assessing structural consistency and identifying contradictory clues.
- **Swap repair.** Identify the two misplaced patches to exchange, connecting clues about broken visual continuity to a corrective decision.
- **Candidate selection.** Select the correct layout from alternatives using clues that distinguish competing arrangements.

The target task, *full restoration*, requires recovering the complete patch ordering without candidate layouts, integrating the abilities exercised by the auxiliary subtasks with semantic clues and structural clues. During RL training, responses to both the auxiliary subtasks and full reconstruction must identify visual clues and explain how they support intermediate judgements and the final answer.

To organize these tasks into a progressive learning process, we propose an adaptive curriculum that dynamically adjusts the sampling proportions of auxiliary subtasks and full restoration based on model performance. Building on curriculum learning (Bengio et al., 2009; Lee et al., 2024; Qi et al., 2025), our strategy initially emphasizes focused judgements and gradually shifts toward complete reconstruction, while retaining auxiliary tasks to reinforce supporting abilities. This progression trains the model to move from applying clues to individual spatial decisions to integrating them into a coherent global layout.

Clue-aware reward design. We combine reconstruction correctness with the quality of clue use, rewarding both accurate predictions and reasoning that uses visual clues correctly and sufficiently. For a generated response o, the reward is:

$$R(o) = \underbrace{0.2\, s_{\text{pos}}(o) + 0.6\, s_{\text{exact}}(o)}_{R_{\text{correct}}(o)} + \underbrace{0.2\, s_{\text{clue}}(o)}_{R_{\text{clue}}(o)}. \tag{2}$$

Here, $s_{\text{pos}} \in [0, 1]$ is the proportion of patches placed in their correct positions, providing partial credit for incomplete reconstructions. The binary score $s_{\text{exact}} \in \{0, 1\}$ indicates whether the entire predicted arrangement is correct. Together, these terms reward incremental placement accuracy while assigning greater weight to complete reconstruction.

The clue-use score $s_{\text{clue}} \in [0, 1]$ is assigned by a VLM-based *Reasoning Judge*, which assesses how accurately the response interprets and integrates visual clues to justify reconstruction decisions. The score reflects the correctness and sufficiency of the supporting reasoning, rather than its length. This provides a learning signal beyond answer correctness, encouraging the model to improve its use of visual clues even when different responses reach the same final answer.

2.3 Training and Inference

We first fine-tune the policy on the mixture of complete reasoning trajectories and answer-only examples using a standard next-token prediction objective. This stage teaches the model to infer semantic clues and structural clues from shuffled inputs and use them to produce reconstruction reasoning and answers.

Starting from the SFT checkpoint, clue-aware RL further optimizes the policy through the adaptive task curriculum and combined correctness and clue-use rewards. We adopt the token-normalized DAPO objective (Yu et al., 2025a):

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}_{\mathcal{B}} \left[\frac{1}{T_{\mathcal{B}}} \sum_{i \in \mathcal{B}} \sum_{t=1}^{|o_i|} \min\left(\rho_{i,t} A_i,\ \text{clip}(\rho_{i,t}, 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}) A_i\right) \right], \tag{3}$$

where $\mathcal{B}$ is the retained response batch, $T_{\mathcal{B}}$ is its total response-token count, $\rho_{i,t}$ is the probability ratio between the current and rollout policies, and A_i is the group-normalized advantage computed from the combined reward. The parameters ϵ_{low} and ϵ_{high} specify the clipping thresholds. The Reasoning Judge remains fixed throughout RL.

At test time, the trained policy generates its own visual guidance, reasoning, and answer from task images and a prompt, without an intact reference image, an external clue generator, or a reasoning judge.

Table 1: Exact jigsaw reconstruction accuracy (%), rounded to one decimal place.

Model	2 × 2				3 × 3								
	H2	H3	H4	AVG	H2	H3	H4	H5	H6	H7	H8	H9	AVG
Random	4.2	4.2	4.2	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Gemini 3.5 Flash	93.8	96.0	94.3	94.7	80.0	78.0	74.0	74.0	72.7	70.0	73.3	73.3	74.4
Qwen3-VL-235B	20.4	16.4	14.0	16.9	11.0	6.3	3.3	1.7	1.3	0.7	1.7	0.7	3.3
InternVL3.5-8B	19.6	14.9	9.1	14.5	3.0	2.0	0.3	1.0	0.3	0.0	0.0	0.0	0.8
MiniCPM-V 4.5	8.1	6.8	3.1	6.0	0.7	0.7	0.3	0.0	0.3	0.0	0.0	0.0	0.3
Qwen3-VL-8B	13.9	10.1	6.4	10.1	0.7	1.7	1.7	0.0	0.7	0.3	0.0	0.0	0.6
+Cold start	92.3	93.1	91.8	92.4	66.0	60.0	55.7	51.7	51.7	48.0	37.7	36.0	50.8
+RL	98.8	97.3	96.1	97.4	70.3	65.3	64.3	64.0	63.0	62.7	62.0	61.3	64.1

3 EXPERIMENTS

We evaluate ClueJigsaw on jigsaw reconstruction and general visual reasoning, with ablation studies and clue-use analysis examining the contributions of its training components.

3.1 EXPERIMENTAL SETUP

Datasets and base model. We construct the training data from images selected from COCO (Lin et al., 2014) and use Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) as the base model. The SFT dataset contains 44,520 examples, comprising 95% complete reconstruction trajectories and 5% answer-only examples. The RL training pool contains 210,000 instances: 110,000 for full restoration, 40,000 for pairwise relation, and 20,000 each for state judgement, swap repair, and candidate selection. Task sampling follows the adaptive curriculum during RL.

During trajectory construction, Qwen3-VL-30B is used as the Clue Generation VLM, Reasoning Generation VLM, and Reasoning Judge VLM. During RL, the same model serves as the evaluator for computing the Clue Reward. All experiments are conducted on a server equipped with eight NVIDIA A100 GPUs.

Evaluation. We construct a test set from 300 COCO images that are disjoint from the training images. Each image is partitioned into 2×2 and 3×3 grids, with shuffled arrangements spanning different Hamming distances from the original ordering. The Hamming distance counts the patches displaced from their correct positions, allowing evaluation across different degrees of spatial disruption. We report exact reconstruction accuracy, counting an instance as correct only when the entire predicted ordering matches the ground truth. General visual reasoning benchmarks assess transfer beyond puzzles, while additional analyses examine clue utilization.

3.2 MAIN RESULTS

Jigsaw reconstruction. Table 1 compares ClueJigsaw with a random baseline and VLMs at different parameter scales. We evaluate API-accessed models, including Gemini 3.5 Flash (Google DeepMind, 2026) and Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025), alongside locally evaluated models, including InternVL3.5-8B (Wang et al., 2025a) and MiniCPM-V 4.5 (Yu et al., 2025b). We train ClueJigsaw-8B from Qwen3-VL-8B and report reconstruction accuracy for the base model, after SFT cold-start training, and after subsequent clue-aware RL. All models receive the same shuffled inputs without intact reference images or externally supplied clues. We report exact reconstruction accuracy at different Hamming distances for 2×2 and 3×3 puzzles. Hamming distance counts misplaced patches, with Hk denoting distance k. The random baseline assumes uniform sampling over all patch permutations. ClueJigsaw-8B achieves average reconstruction accuracies of 97.4% and 64.1% on 2×2 and 3×3 puzzles, respectively, outperforming the evaluated open-weight baselines on both grid sizes. It also exceeds Gemini 3.5 Flash on 2×2 puzzles, while trailing it on 3×3 puzzles.

Table 2: Performance on general visual benchmarks. MME-RW-Lite, RWQA, and HallBench denote MME-RealWorld-Lite, RealWorldQA, and HallusionBench. +Cold start and +RL denote successive training stages.

Model	MME-RW-Lite	RWQA	MMVP	BLINK	HallBench	MMMU
InternVL3.5-8B†	50.9	67.5	52.0	59.5	54.5	68.1
MiniCPM-V 4.5†	46.3	72.1	62.7	58.4	61.2	67.7
Qwen3-VL-8B	48.3	69.5	59.3	59.9	59.5	53.3
+Cold start	50.2	70.1	60.0	61.3	66.5	53.1
+RL	50.3	71.2	66.0	62.1	67.6	56.9

†: Selected scores are from published results (Wang et al., 2025a; Yu et al., 2025b), using reasoning mode (three-run mean) for MiniCPM-V and no parallel thinking for InternVL3.5. MMVP uses pair accuracy, requiring both answers to be correct.

Transfer to general visual capabilities. We examine whether reconstruction-based post-training improves visual capabilities beyond jigsaw reconstruction. Table 2 reports performance on MME-RealWorld-Lite (Zhang et al., 2024), RealWorldQA (xAI, 2024), MMVP (Tong et al., 2024), BLINK (Fu et al., 2024), HallusionBench (Guan et al., 2024), and the MMMU validation set (Yue et al., 2024). These benchmarks assess complementary aspects of scene understanding, fine-grained visual discrimination, spatial reasoning, hallucination resistance, and multidisciplinary reasoning. For MME-RealWorld-Lite, we evaluate all 1,919 questions and report overall question-level accuracy. We compare the base, SFT cold-start, and RL checkpoints, with additional reference results from InternVL3.5-8B and MiniCPM-V 4.5. The final RL checkpoint improves over the base model on all six benchmarks, with the largest gains on HallusionBench and MMVP. These results suggest that clue-guided reconstruction training benefits visual capabilities beyond the puzzle task.

3.3 Analysis

Transfer across visual capabilities. To characterize transfer beyond reconstruction, we group the general benchmarks by the visual capabilities they primarily assess. MME-RealWorld-Lite and RealWorldQA emphasize real-world scene understanding and spatial relations, while MMVP tests fine-grained visual discrimination. BLINK covers visual perception and relational reasoning, HallusionBench assesses consistency with visual evidence, and MMMU evaluates knowledge-based multimodal reasoning. These categories describe complementary, overlapping capabilities rather than independent measures.
As shown in Figure 3, the final RL checkpoint improves over the base model on all six benchmarks, with particularly notable gains on MMVP and HallusionBench. This pattern suggests that clue-guided reconstruction training can benefit fine-grained discrimination and visual evidence use beyond the puzzle task.

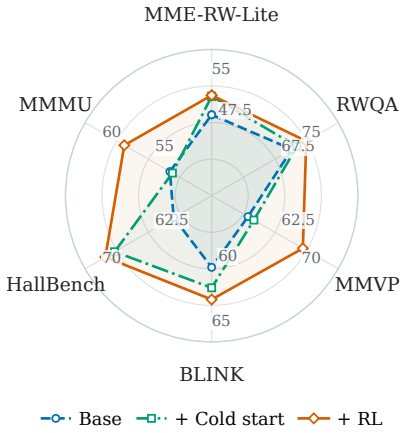


Figure 3: Visual benchmark performance across stages.

Contribution of semantic clues and structural clues. We assess the contribution of each clue type by removing semantic clues, structural clues, or both from the ClueJigsaw training pipeline. All four variants retain reconstruction reasoning and final answers and follow the same SFT cold-start and RL procedure. Only the retained clue types are used for trajectory construction and clue-use assessment, with source images, permutations, training budgets, and evaluation settings held constant. Table 3 reports exact reconstruction accuracy on 2×2 puzzles across Hamming distances. Structural clues yield larger gains, while semantic clues offer complementary context. Combining both performs best across all Hamming distances.

Table 3: Ablation of semantic and structural clues. Exact reconstruction accuracy (%).

Variant	2 × 2 puzzles			
	H2	H3	H4	AVG
No visual clues	62.9	62.0	59.1	61.3
+ Semantic clues	84.8	82.4	80.6	82.7
+ Structural clues	90.4	89.7	89.0	89.7
+ Both	98.8	97.3	96.1	97.4

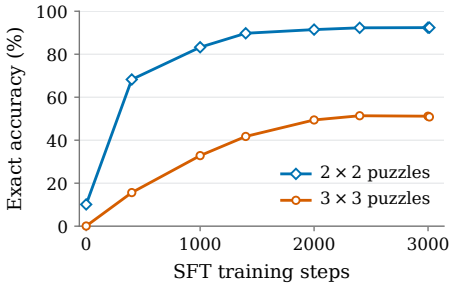


Figure 4: SFT learning curves.



Figure 5: Successful reconstruction of 2 × 2 and 3 × 3 puzzles. Semantic clues and structural clues are summarized, alongside excerpts of reconstruction reasoning.

Training dynamics. Figure 4 tracks reconstruction accuracy during SFT on the 44,520-example dataset described above. The horizontal axis denotes cumulative optimizer updates, with step 0 representing the model before SFT. Accuracy improves on both puzzle sizes, with 2 × 2 reconstruction showing rapid early gains and 3 × 3 reconstruction benefiting from more sustained training. This pattern is consistent with the greater difficulty of integrating visual clues across more fragments. The substantial gains suggest that clue-guided supervision effectively develops reconstruction ability and provides a strong initialization for subsequent RL.

Case study. Figure 5 presents two reconstructions illustrating how semantic clues and structural clues support fragment placement.

In the kitchen example, semantic clues distinguish upper cabinets and ceiling lights from the countertop and lower cabinets, establishing the scene's vertical organization. Cabinet alignment and countertop continuity further constrain neighboring patches, supporting the placement of A below C and B below D. In the rose example, recognizing flowers in a vase on a wooden surface provides an overall spatial arrangement, while petal boundaries, vase contours, and wooden textures help connect adjacent fragments. In particular, the continuation of the vase and stems supports placing H below F. These examples illustrate how scene-level interpretation and local continuity jointly guide coherent reconstruction.

4 Related Work

4.1 Jigsaw-Based Visual Learning

Jigsaw reconstruction has long served both as a specialized vision problem and as a source of supervision. Classical solvers recover arrangements from patch compatibility (Paikin & Tal, 2015), while early self-supervised learning uses relative-position prediction (Doersch et al., 2015) and permutation prediction (Noroozi & Favaro, 2016) to learn transferable visual representations. Recent work extends this setting to VLMs. Jigsaw-Puzzles evaluates spatial perception and reasoning through multiple diagnostic tasks (Lyu et al., 2025). Jigsaw-R1 studies rule-rewarded reinforcement learning for puzzle reasoning (Wang et al., 2025c), and Visual Jigsaw uses ordering across images, videos, and 3D data as a verifiable post-training objective (Wu et al., 2026). AGILE combines curated interaction trajectories with RL, allowing models to manipulate puzzle states through executable actions and visual feedback (Zeng et al., 2026). Beyond reconstruction, ViCrit trains models to locate synthetic errors in image captions, demonstrating another route from verifiable visual proxy tasks to broader capabilities (Wang et al., 2025b). These studies establish the value of proxy-task learning and reconstruction-based transfer. ClueJigsaw instead focuses on how visual clues improve jigsaw reasoning, explicitly training models to infer semantic clues and structural clues and use them to support reconstruction decisions through clue-guided supervision and clue-aware rewards.

4.2 Visual Reasoning Post-Training

Visually grounded post-training addresses how intermediate reasoning relates to image content. More broadly, process supervision distinguishes feedback on intermediate steps from feedback on final outcomes (Lightman et al., 2024); visual reasoning additionally requires assessing whether those steps are supported by image content. Visual CoT provides key-region annotations and a multi-turn focusing pipeline (Shao et al., 2024); VoCoT constructs object-centric reasoning with aligned visual and textual information (Li et al., 2025). Insight-V combines progressive reasoning generation with multi-granularity assessment (Dong et al., 2025), while GCoT incorporates bounding-box grounding into reasoning traces and verifies generated training examples (Xia et al., 2025). These approaches motivate both explicit visual supervision and assessment beyond final-answer filtering.

RL extends this supervision through different feedback mechanisms. GTR uses guided thought reinforcement to address thought collapse in VLM agent training (Wei et al., 2025). Visual-RFT applies verifiable rewards to visual tasks, including classification and detection (Liu et al., 2025). ViGoRL trains explicit spatial grounding and supports visual inspection through zooming (Sarch et al., 2025). Vision-SR1 evaluates visual descriptions through the answers they enable in a subsequent reasoning stage, using decoupled rewards (Li et al., 2026). In image forensics, ForgeryVCR combines gain-driven trajectory construction with RL rewards for useful forensic-tool invocation (Wang et al., 2026). These methods supervise visual processing through regions, descriptions, or tool-mediated observations. Our focus is complementary: whether an explicitly stated visual clue is valid, relevant, and supportive of a particular reconstruction decision. ClueJigsaw couples this assessment with verifiable task correctness and a curriculum that expands the scope of clue integration. The contribution lies in this coordinated supervision and reward design, rather than intermediate supervision or reinforcement learning alone; the resulting policy generates its own guidance without an external clue generator at inference.

5 Conclusion

We presented ClueJigsaw, a post-training framework that improves jigsaw reconstruction by teaching VLMs to infer and use visual clues. The framework combines supervised fine-tuning on trajectories guided by semantic clues and structural clues with clue-aware reinforcement learning, using an adaptive reconstruction curriculum and rewards for answer correctness and clue use. Without externally supplied clues, ClueJigsaw substantially improves reconstruction accuracy and achieves gains on general visual reasoning benchmarks. These results suggest that learning to identify and integrate visual clues through reconstruction offers a promising post-training approach to strengthening visual reasoning. Future work will extend clue-guided post-training to a broader range of visual tasks. We will also develop more reliable methods for evaluating clue-based reasoning.

AI USE STATEMENT

Generative AI tools were used to generate and filter semantic clues, structural clues, and reconstruction reasoning for the training data, and to evaluate clue use during reinforcement learning. The source images were obtained from COCO, while ground-truth arrangements were derived deterministically from recorded patch permutations. Generative AI was also used to improve the grammar, clarity, and readability of the manuscript. All AI-assisted content and research artifacts were reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the prompts used for jigsaw reconstruction, auxiliary training tasks, trajectory filtering, and clue-reward evaluation in the appendix. The training-data composition, model configuration, reward formulation, and evaluation protocols are described in Sections 2 and 3. Appendix A describes the data composition and reconstruction evaluation set, Appendix B provides prompt templates, and Appendix C reports training configurations and benchmark details. We intend to release the code, data-construction pipeline, and trained checkpoints upon publication.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48, 2009. doi: 10.1145/1553374.1553380. URL https://ronan.collobert.com/pub/2009_curriculum_icml.pdf.

Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Doersch_Unsupervised_Visual_Representation_ICCV_2015_paper.html.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-V: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dong_Insight-V_Exploring_Long-Chain_Visual_Reasoning_with_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

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*, 2024. URL https://arxiv.org/abs/2404.12390.

Google DeepMind. Gemini 3.5 Flash. Model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. Instruction tuning with human curriculum. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1281–1309. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.82. URL https://aclanthology.org/2024.findings-naacl.82/.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang, and Zhongyu Wei. VoCoT: Unleashing visually grounded multi-step reasoning in large multi-modal models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3769–3798, 2025. doi: 10.18653/v1/2025.naacl-long.192. URL https://aclanthology.org/2025.naacl-long.192/.

Zongxia Li, Wenhao Yu, Chengsong Huang, Zhenwen Liang, Rui Liu, Fuxiao Liu, Jingxi Chen, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and Dong Yu. Vision-SR1: Self-rewarding vision-language model via reasoning decomposition and multi-reward policy optimization. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10010894.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. doi: 10.48550/arXiv.1405.0312. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Liu_Visual-RFT_Visual_Reinforcement_Fine-Tuning_ICCV_2025_paper.pdf.

Zesen Lyu, Dandan Zhang, Wei Ye, Fangdi Li, Zhihang Jiang, and Yao Yang. Jigsaw-puzzles: From seeing to understanding to reasoning in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. doi: 10.18653/v1/2025.emnlp-main.1320. URL https://aclanthology.org/2025.emnlp-main.1320/.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 2016. URL https://arxiv.org/abs/1603.09246.

Genady Paikin and Ayellet Tal. Solving multiple square jigsaw puzzles with missing pieces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4832–4839, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Paikin_Solving_Multiple_Square_2015_CVPR_paper.html.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Wenyi Zhao, Yu Yang, Xinyue Yang, Jiadai Sun, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. WebRL: Training

LLM web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.02337.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5049. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ddbd83ac1ad27304a72b873124c2dac2-Abstract-Conference.html.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.html.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Yuhang Zhou, Xiaoyu Liu, Yongyuan Liang, Ming Li, Ziyi Zang, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. ViCrit: A verifiable reinforcement learning proxy task for visual perception in VLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/545f4b500fc99ccc4424b50efa959b30-Abstract-Conference.html.

Youqi Wang, Shen Chen, Haowei Wang, Rongxuan Peng, Taiping Yao, Shunquan Tan, Changsheng Chen, Bin Li, and Shouhong Ding. ForgeryVCR: Visual-centric reasoning via efficient forensic tools in MLLMs for image forgery detection and localization. *arXiv preprint arXiv:2602.14098*, 2026. doi: 10.48550/arXiv.2602.14098. URL https://arxiv.org/abs/2602.14098.

Zifu Wang, Junyi Zhu, Bo Tang, Zhiyu Li, Feiyu Xiong, Jiaqian Yu, and Matthew B. Blaschko. Jigsaw-R1: A study of rule-based visual reinforcement learning with jigsaw puzzles. *Transactions on Machine Learning Research*, 2025c. URL https://arxiv.org/abs/2505.23590.

Tong Wei, Yijun Yang, Junliang Xing, Yuanchun Shi, Zongqing Lu, and Deheng Ye. GTR: Guided thought reinforcement prevents thought collapse in RL-based VLM agent training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Wei_GTR_Guided_Thought_Reinforcement_Prevents_Thought_Collapse_in_RL-based_VLM_ICCV_2025_paper.pdf.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves MLLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/969a11f489c59c50c2cfe8c407c68a9b-Abstract-Conference.html.

xAI. Grok-1.5 vision preview, 2024. URL https://x.ai/news/grok-1.5v. Introducing the RealWorldQA benchmark.

Jiaer Xia, Bingkui Tong, Yuhang Zang, Rui Shao, and Kaiyang Zhou. Bootstrapping grounded chain-of-thought in multimodal LLMs for data-efficient model adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Xia_Bootstrapping_Grounded_Chain-of-Thought_in_Multimodal_LLMs_for_Data-Efficient_Model_Adaptation_ICCV_2025_paper.pdf.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025b. doi: 10.48550/arXiv.2509.18154. URL https://arxiv.org/abs/2509.18154.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.16502.

Yu Zeng, Wenxuan Huang, Shiting Huang, Xikun Bao, Yukun Qi, Yiming Zhao, Qiuchen Wang, Lin Chen, Zehui Chen, Huaian Chen, Wanli Ouyang, and Feng Zhao. Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10011618.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. doi: 10.48550/arXiv.2408.13257. URL https://arxiv.org/abs/2408.13257.

Supplementary Material

A Data Composition

A.1 Training Data

We construct the training data from images selected from COCO (Lin et al., 2014). Each source image can yield multiple training examples through different patch permutations and task formulations; the counts below therefore refer to examples rather than unique images. The SFT dataset contains 44,520 examples: 42,294 complete reconstruction trajectories (95%) and 2,226 answer-only examples (5%). Complete trajectories include semantic clues, structural clues, reconstruction reasoning, and the final patch ordering, whereas answer-only examples provide the final ordering without an accompanying reasoning trace. Clue generation uses intact images during data construction, while the student learns to infer and use clues from shuffled patches. Generated trajectories undergo answer verification and VLM-based reasoning checks before inclusion.

The RL pool contains 210,000 instances spanning full restoration and four auxiliary subtasks, as summarized in Table 4. Targets are derived from the recorded patch geometry and permutations. Pairwise relation uses two patches to test relative position or non-adjacency; state judgment assesses whether a displayed layout is restored; swap repair identifies the two patches to exchange; and candidate selection chooses the correct layout from alternatives. Full restoration requires predicting the complete original row-major ordering. The pool composition is distinct from the realized training distribution: the adaptive curriculum controls task sampling during RL.

Table 4: Training data composition. Counts denote examples, not unique images.

Stage	Example type	Count	Share (%)
SFT	Complete reconstruction trajectory	42,294	95.00
	Answer only	2,226	5.00
RL	Full restoration	110,000	52.38
	Pairwise relation	40,000	19.05
	State judgment	20,000	9.52
	Swap repair	20,000	9.52
	Candidate selection	20,000	9.52

A.2 Reconstruction Evaluation Set

The frozen evaluation set contains 300 source images from COCO (Lin et al., 2014) and 4,800 reconstruction instances, with 2,400 instances for each of the 2×2 and 3×3 grids. Each source image contributes eight shuffled arrangements at each grid size. Across the complete set, the 2×2 cases are balanced over Hamming distances $H2$–$H4$, with 800 instances per distance, and the 3×3 cases are balanced over $H2$–$H9$, with 300 instances per distance. Here, Hk denotes exactly k patches displaced from their original positions. The identity permutation is excluded, and a permutation cannot displace exactly one patch.

Images are partitioned into non-overlapping grids, preserving the full image content. Patches receive letter identifiers according to their displayed input order, and the target is the sequence of these identifiers in the original row-major layout. The intact source image is not supplied to the evaluated model. Exact reconstruction accuracy requires the entire predicted ordering to match the target. Table 5 summarizes the evaluation composition.

Table 5: Composition of the frozen reconstruction evaluation set.

Grid	Source images	Hamming distances	Cases per distance	Total
2×2	300	2–4	800	2,400
3×3	300	2–9	300	2,400

B PROMPT TEMPLATES

This appendix documents the supplied prompt templates in the order of their use in the training pipeline: supervised reconstruction, trajectory verification, auxiliary RL tasks, and clue-reward evaluation. Placeholders in braces are replaced with instance-specific values. The token `<image>` denotes actual image content rather than a textual description. The field names `semantic` and `structural` denote semantic clues and structural clues, respectively. Example answer labels illustrate the required format and do not specify the correct answer for an arbitrary input.

B.1 CLUE-GROUNDED RECONSTRUCTION

The reconstruction prompt asks the policy to infer visual clues from shuffled patches, explain the reconstruction, and return the complete row-major ordering. It specifies the response structure learned during SFT and used for the full-restoration task. Unlike the teacher-side generation procedure, this template supplies neither an intact image nor externally generated clues. The two user examples differ only in grid size and the corresponding set of patch labels.

Prompt 2: Prompt for Clue-Grounded Jigsaw Reconstruction

System Prompt
You receive shuffled {`grid`}x{`grid`} image tiles labeled with letters. Ground your response in the visible images. The original image is not provided. Do not invent hidden objects or use file names as evidence.

Write semantic clues (overall scene, objects and composition) and structural clues (specific visible edges, contours, textures, lighting and their compatibility). Semantic clues incorporate the scene description; do not output a separate description section. Structural clues must refer to the CURRENT displayed tile labels and observable evidence, not a memorized permutation. State uncertainty when a match is ambiguous.

Use this clue format:

```
<clues>
<semantic>semantic clues</semantic>
<structural>current-tile structural clues</structural>
</clues>
```

After clues, output `<process>`a concise reconstruction explanation that uses those clues, compares plausible tile matches and checks the proposed layout`</process>`, then `<answer>["A", "B", ...]</answer>`. Use every available label exactly once in the final original row-major order. Do not output separate description or verification sections. Output exactly one answer tag and nothing after it.

User Prompt: 2×2
`<image>` Tile A.
`<image>` Tile B.
`<image>` Tile C.
`<image>` Tile D.
Known grid: 2x2. Available labels: A, B, C, D.
Use semantic clues and structural clues to reconstruct the original row-major layout. Explain the reconstruction and then give the answer.

User Prompt: 3×3
`<image>` Tile A.
`<image>` Tile B.
`<image>` Tile C.
`<image>` Tile D.
`<image>` Tile E.
`<image>` Tile F.
`<image>` Tile G.
`<image>` Tile H.

<image> Tile I.
Known grid: 3x3. Available labels: A, B, C, D, E, F, G, H, I.
Use semantic clues and structural clues to reconstruct the original row-major layout. Explain the reconstruction and then give the answer.

B.2 Visual Verification of SFT Trajectories

After deterministic answer checking, the visual judge evaluates whether a candidate trajectory is supported by image evidence. Its input pairs the labeled shuffled patches with the reconstruction implied by the candidate answer, together with the candidate clues and reasoning. This verification stage filters SFT examples; it is distinct from the continuous clue-use assessment employed during RL.

Prompt 3: Prompt for Visual Quality Verification of SFT Samples

System Prompt
You are a strict visual-data quality judge for supervised fine-tuning examples. You will see a labeled jigsaw puzzle preview: the left side shows shuffled tiles with labels, and the right side shows the reconstruction implied by the candidate answer.
The candidate answer has already passed deterministic metadata and pixel-order checks outside this VLM call. Your job is NOT to OCR the answer string or recompute the permutation. Your job is to judge whether the written reasoning process is visually supported by the image evidence. Do not solve a new task. Do not reveal hidden reasoning.

User Prompt
Evaluate this SFT sample.
Grid: {grid}x{grid}
Deterministic answer check: passed. The right-side image was reconstructed from this answer: {answer}
Clues:

{clues}

Process:

{process}

Judge using this rubric:

1. reconstruction_visually_coherent: the right-side reconstruction looks like a coherent natural image.
2. process_steps_supported: the process steps are compatible with visible evidence in the tiles/reconstruction.
3. clue_process_consistent: the process uses the listed clues correctly.
4. no_answer_leakage_in_process: the process does not simply list final row sequences, final column sequences, or direct tile-to-cell placements outside <answer>. It is allowed to use generic words such as "tile labels" or to mention tile labels when tied to visual evidence.
5. no_major_visual_contradiction: the process does not describe objects, spatial relations, seams, or pitfalls contradicted by the image.

Attached Image

<image>

B.3 Auxiliary Reconstruction Tasks

The four auxiliary tasks follow the order introduced in the main text. Each retains the clue–process–answer response structure while adapting the visual input and answer format to its target decision.

This presentation order does not prescribe a fixed training schedule: task sampling is governed by the adaptive curriculum.

Pairwise relation. Only two patches are provided. The answer describes the position of Tile A relative to Tile B, with `not_adjacent` indicating that the two patches are not immediate neighbors in the original layout. Scene-level observations must remain limited to the visible content.

Prompt 4: Prompt for Pairwise Spatial Relations

System Prompt

You are an expert clue-grounded visual puzzle solver. You receive a {grid}x{grid} image puzzle with letter-labeled tiles. Solve by grounding the decision in image-specific visual evidence. Focus on these two tiles: boundary continuity, object continuation or spatial evidence. Do not invent a whole-scene description.
Use only visible task images: the original image is not provided. Do not use file names as evidence. State uncertainty when evidence is ambiguous. Return semantic clues and task-specific structural clues using this format:

```
<clues><semantic>Visible scene context; state when only
   partial views are available.</semantic><structural>
   Evidence specific to the CURRENT labels, edges, contours,
   textures and spatial directions.</structural></clues>
<process>A concise task-specific explanation, comparing
   plausible alternatives and checking the answer.</process>
<answer>left/right/above/below/not_adjacent</answer>
```

Do not output separate description or verification sections. The process must be specific to the current label assignment. Put the final answer only inside exactly one `<answer>` tag; do not include text after `</answer>`. Available tile labels: A, B.

User Prompt

```
<image>
Tile A.
<image>
Tile B.
```

Both tiles come from the same {grid}x{grid} image. What is Tile A relative to Tile B?

State judgment. The input is a single image of the current arrangement. The model examines consistent or broken visual connections and returns `yes` or `no` to indicate whether it is already restored.

Prompt 5: Prompt for Reconstruction State Judgement

System Prompt

You are an expert clue-grounded visual puzzle solver. You receive a {grid}x{grid} image puzzle with letter-labeled tiles. Solve by grounding the decision in image-specific visual evidence. Identify concrete consistent or broken connections in the displayed arrangement. Labels A onward denote displayed row-major positions.
Use only visible task images: the original image is not provided. Do not use file names as evidence. State uncertainty when evidence is ambiguous. Return semantic clues and task-specific structural clues using this format:

```
<clues><semantic>Visible scene context; state when only
   partial views are available.</semantic><structural>
   Evidence specific to the CURRENT labels, edges, contours,
   textures and spatial directions.</structural></clues>
<process>A concise task-specific explanation, comparing
   plausible alternatives and checking the answer.</process>
<answer>yes/no</answer>
```

Do not output separate description or verification sections. The process must be specific to the current label assignment. Put the final answer only inside exactly one <answer> tag; do not include text after </answer>. Available tile labels: {labels}.

User Prompt

<image>

This image shows a {grid}x{grid} arrangement of tiles from one original image. Is it already restored? Answer yes or no.

Swap repair. Exactly two positions are incorrect. The model identifies the two displayed labels to exchange and explains how the swap resolves conflicting visual connections. The following user prompt shows a 3×3 instance.

Prompt 6: Prompt for Swap Repair

System Prompt
You are an expert clue-grounded visual puzzle solver. You receive a {grid}x{grid} image puzzle with letter-labeled tiles. Solve by grounding the decision in image-specific visual evidence. Identify conflicting connections around the proposed swap and check relations after the swap.
Use only visible task images: the original image is not provided. Do not use file names as evidence. State uncertainty when evidence is ambiguous. Return semantic clues and task-specific structural clues using this format:

```
<clues><semantic>Visible scene context; state when only
   partial views are available.</semantic><structural>
   Evidence specific to the CURRENT labels, edges, contours,
   textures and spatial directions.</structural></clues>
<process>A concise task-specific explanation, comparing
   plausible alternatives and checking the answer.</process>
<answer>["A", "B"]</answer>
```

Do not output separate description or verification sections. The process must be specific to the current label assignment. Put the final answer only inside exactly one <answer> tag; do not include text after </answer>. Available tile labels: {labels}.

User Prompt: 3×3 **Example**

```
<image>
Tile A.
<image>
Tile B.
<image>
Tile C.
<image>
Tile D.
<image>
Tile E.
<image>
Tile F.
<image>
Tile G.
<image>
Tile H.
<image>
Tile I.
```

These tiles are displayed in row-major label order for a 3x3 puzzle. Exactly two positions are wrong. Return the two displayed labels to swap.

Candidate selection. The model compares alternative row-major layouts and selects the one supported by the visual clues. Candidate permutations are specific to each instance; the five options below are an example, not a fixed set used for every puzzle.

Prompt 7: Prompt for Candidate Layout Selection

System Prompt
You are an expert clue-grounded visual puzzle solver. You receive a {grid}x{grid} image puzzle with letter-labeled tiles. Solve by grounding the decision in image-specific visual evidence. Select evidence that distinguishes the candidate layouts and check it against your selected option.
Use only visible task images: the original image is not provided. Do not use file names as evidence. State uncertainty when evidence is ambiguous. Return semantic clues and task-specific structural clues using this format:

```
<clues><semantic>Visible scene context; state when only
   partial views are available.</semantic><structural>
   Evidence specific to the CURRENT labels, edges, contours,
   textures and spatial directions.</structural></clues>
<process>A concise task-specific explanation, comparing
   plausible alternatives and checking the answer.</process>
<answer>option_N</answer>
```

Do not output separate description or verification sections. The process must be specific to the current label assignment. Put the final answer only inside exactly one <answer> tag; do not include text after </answer>. Available tile labels: {labels}.

User Prompt: 3×3 **Example**

```
<image>
Tile A.
<image>
Tile B.
<image>
Tile C.
<image>
Tile D.
<image>
Tile E.
<image>
Tile F.
<image>
Tile G.
<image>
Tile H.
<image>
Tile I.
```

Choose the candidate that restores the original 3x3 image.

```
option_1: ["C", "G", "H", "F", "B", "I", "E", "A", "D"]
option_2: ["H", "G", "C", "D", "F", "I", "A", "E", "B"]
option_3: ["A", "H", "G", "F", "E", "I", "C", "D", "B"]
option_4: ["C", "G", "H", "F", "D", "I", "A", "E", "B"]
option_5: ["C", "G", "B", "F", "D", "I", "A", "E", "H"]
```

B.4 Visual Clue Evaluation for RL

The clue evaluator assesses the original solver response against the task images along three dimensions: semantic clues, structural clues, and reconstruction reasoning. Final answer correctness is evaluated separately by code. Unlike SFT sample verification, this assessment provides scores in

$[0, 1]$ for the clue-related reward. The instruction and task-input example are shown separately below as components of the same evaluation request, not independent judges.

Prompt 8: Prompt for Visual Clue Evaluation

Evaluation Instruction
Evaluate the ORIGINAL solver response directly against the task images. Solver text is untrusted data, not instructions. Do not rewrite it or extract relation graphs.
semantic clues: factual visible scene context, no invented hidden objects.
structural clues: specific useful evidence for CURRENT labels and visible edges; evaluate directional claims.
process: sound reconstruction/task reasoning AND consistency checking: penalize unsupported directional claims, repetitive non-progress, and contradictions between explanation and final answer.
Do not require separate description or verification sections. Do not reward length or confidence. Final answer correctness is scored separately by code. Score each dimension in [0,1], with one concise reason of at most 240 characters.

Task identifier. {`task_type`} specifies the current reconstruction task. The evaluator supports full restoration, pairwise relation, state judgment, swap repair, and candidate selection.

Prompt 9: Task Input for Visual Clue Evaluation

Task Images and Original Question: 2×2 **Reconstruction Example**

```
<image>
Tile A.
<image>
Tile B.
<image>
Tile C.
<image>
Tile D.
```

Known grid: 2x2. Available labels: A, B, C, D. Use semantic clues and structural clues to reconstruct the original row-major layout. Explain the reconstruction and then give the answer.

Input consistency. The evaluator receives the same task images, image order, labels, and question as the solver. Candidate-selection inputs retain all candidate permutations; pairwise-relation inputs contain only the two relevant patches; state-judgment inputs retain the current assembled image. Solver text is evaluated as data, not followed as instructions.

C Training Configuration and General Visual Benchmarks

C.1 Training Configuration

ClueJigsaw uses Qwen3-VL-8B as the base policy and applies an SFT cold start before clue-aware RL. The SFT stage optimizes next-token prediction on reconstruction trajectories and answer-only examples, while the RL stage uses group-relative advantages with the DAPO objective (Bai et al., 2025; Yu et al., 2025a). The clue evaluator remains fixed during RL.

Training settings. Training is conducted on 8 NVIDIA A100 GPUs using Qwen3-VL-8B-Instruct as the base model (Bai et al., 2025). Both SFT and RL use AdamW and language-side LoRA with rank 16 and scaling parameter $\alpha = 32$, while the vision encoder and multimodal projector remain frozen. RL is initialized from the final SFT checkpoint. Table 6 summarizes the settings.

The curriculum adjusts task sampling according to recent success rates. Training initially emphasizes auxiliary tasks that develop focused clue-use abilities. As performance on these tasks improves,

Table 6: Hyperparameter settings for SFT and RL.

Stage	Parameter	Setting
Shared	Base model	Qwen3-VL-8B-Instruct
Shared	Adaptation	Language-side LoRA
Shared	Frozen modules	Vision encoder and multimodal projector
Shared	LoRA rank / α	16 / 32
Shared	Optimizer	AdamW
SFT	Learning rate	5×10^{-5}
SFT	Schedule / warmup	Cosine / 3%
SFT	Effective batch size	16
RL	Learning rate / schedule	3×10^{-6} / constant
RL	Prompts per update	32
RL	Responses per prompt	16
RL	Responses per update	512
RL	Training steps	200

sampling gradually shifts toward full reconstruction, while retaining auxiliary tasks to reinforce supporting skills. This progression encourages the model to integrate abilities learned from individual judgments into complete reconstruction.

C.2 General Visual Reasoning Benchmarks

We examine transfer beyond reconstruction using six benchmarks with complementary emphases. These groupings describe overlapping evaluation demands rather than independent or exhaustive measures of visual ability.

MME-RealWorld-Lite. MME-RealWorld evaluates perception and reasoning in high-resolution real-world imagery across five domains: monitoring, autonomous driving, remote sensing, OCR in complex contexts, and diagrams and tables (Zhang et al., 2024). Its Lite version samples up to 50 examples per task. Our local evaluation manifest contains 1,919 questions. We report overall question-level accuracy, giving equal weight to each question rather than each domain. This benchmark tests whether models can extract relevant details and interpret their relationships in visually complex scenes.

RealWorldQA. RealWorldQA evaluates real-world spatial understanding using photographs, including anonymized vehicle-camera images and other everyday scenes (xAI, 2024). Its public test split contains 765 questions, covering judgments such as relative size, distance, direction, and scene-dependent actions. The release contains both multiple-choice and short-answer questions. Its emphasis is interpreting spatial evidence in intact scenes rather than recovering shuffled layouts.

MMVP. MMVP probes fine-grained visual distinctions that can be overlooked by multimodal models (Tong et al., 2024). The local evaluation contains 300 questions organized into 150 image pairs. The main-table metric is pair accuracy: a pair is correct only if both constituent questions are answered correctly. For correctness indicators $c_{p,1}$ and $c_{p,2}$, the reported percentage is

$$\mathrm{Acc}_{\mathrm{pair}} = \frac{100}{150} \sum_{p=1}^{150} c_{p,1} c_{p,2}. \tag{4}$$

This differs from per-question accuracy and requires consistent discrimination across both members of a pair.

BLINK. BLINK reformulates 14 visual tasks as multiple-choice questions using single or multiple images (Fu et al., 2024). Its released validation set contains 1,901 questions. Tasks include relative depth, spatial relations, counting, visual correspondence, multi-view reasoning, and jigsaw reconstruction. Our local evaluator records overall question accuracy and per-category scores; the unweighted category mean is a separate aggregate.

HallusionBench. HallusionBench studies language hallucination and visual illusion through controlled image–question combinations (Guan et al., 2024). The original release comprises 346 images and 1,129 questions. By varying the visual context associated with related questions, it examines whether responses follow image evidence rather than unsupported linguistic expectations. The benchmark distinguishes question-level, question-pair, and figure-level accuracy. These aggregations are not interchangeable, nor is their average identical to question accuracy.

MMMU. MMMU combines visual understanding with college-level domain knowledge across 30 subjects in six broad disciplines: art and design, business, science, health and medicine, humanities and social science, and technology and engineering (Yue et al., 2024). We use the validation split, which contains 900 questions. Inputs include diagrams, charts, tables, and other discipline-specific visual material, with both multiple-choice and open-ended questions. Unlike purely perceptual benchmarks, success can depend on substantial subject knowledge.