JIGSHAPE: EVALUATING VISUAL-GEOMETRIC REASONING IN VLMS THROUGH JIGSAW PUZZLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Jigsaw puzzle solving requires jointly reasoning about visual content and geometric constraints, yet existing benchmarks use rectangular cuts that create ambiguous ground truth in texture-repeated regions. We introduce *JigShape*, a benchmark with tab-and-blank interlocking pieces where geometric constraints provide strong local compatibility requirements that, combined with visual content, yield unambiguous ground truth. Across 95K instances at four grid densities (4×4 to 16×16), we find that **zero-shot VLMs largely lack geometric reasoning**: only one of five frontier models (GPT-5.5) exceeds random baseline on 4×4 puzzles, while all others perform at chance level. While supervised fine-tuning achieves $>97\%$ on 4×4, demonstrating the task is learnable, **all models collapse on larger grids**: GPT-5.5 drops from 70% to near-random on 8×8, and even fine-tuned models fall below 5% on 12×12. This "scaling cliff" suggests current architectures cannot maintain consistent constraint satisfaction as the number of pieces increases. JigShape establishes scalable geometric reasoning as an open challenge for vision-language models.

Dataset: `https://anonymous-hf.com/a/4pasdiamtue8/`

1 INTRODUCTION

Vision-language models (VLMs) have advanced rapidly in recent years, achieving strong performance on recognition, captioning, visual question answering, and multimodal reasoning (OpenAI, 2023; Anthropic, 2024; Google DeepMind, 2023; Liu et al., 2023b; Li et al., 2023; Dai et al., 2023). These capabilities have enabled applications ranging from document understanding to embodied agents. A natural question arises: *Can VLMs perform spatial reasoning?* Recent studies suggest this remains challenging: VLMs struggle with basic spatial relations (Liu et al., 2023a; Kamath et al., 2023), depth perception (Fu et al., 2024), and 3D understanding (Chen et al., 2024). Spatial reasoning, the ability to infer geometric relationships and reconstruct coherent structures from fragmented observations, is fundamental to many real-world tasks including robotics manipulation, architectural design, and scientific visualization.

Jigsaw puzzles provide a natural testbed for spatial reasoning. Solving a puzzle requires integrating local visual cues (texture, color, object boundaries) with global structural constraints (pieces must tile the image plane without gaps or overlaps). Unlike semantic tasks where partial understanding suffices, jigsaw solving demands precise spatial localization: each piece has exactly one correct position. This makes puzzles an ideal probe for measuring whether VLMs can reason about geometric arrangements rather than merely recognizing visual patterns. Notably, prior work in self-supervised learning has shown that training models to solve jigsaw puzzles yields visual representations that transfer well to downstream tasks such as object detection and image classification (Noroozi & Favaro, 2016; Carlucci et al., 2019; Chen et al., 2023; Ren et al., 2023). This suggests that jigsaw-solving ability is not merely a narrow skill but reflects broader visual reasoning capabilities.

Current Benchmarks. Recent work has begun evaluating VLMs on jigsaw-style tasks. These benchmarks divide images into rectangular patches on a grid (e.g., 2×2 or 3×3), shuffle the patches, and ask models to predict the original arrangement. Jigsaw-Puzzles (Lyu et al., 2025) designs five cognitive tasks around 2×2 and 3×3 grids; the best model (Gemini-2.5-Pro) reaches 77% overall but drops to 30% on open-ended order generation, exposing a gap between multiple-choice recognition and unconstrained spatial reconstruction. The JPwLEG dataset (Song et al., 2023) provides 3×3 and

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.

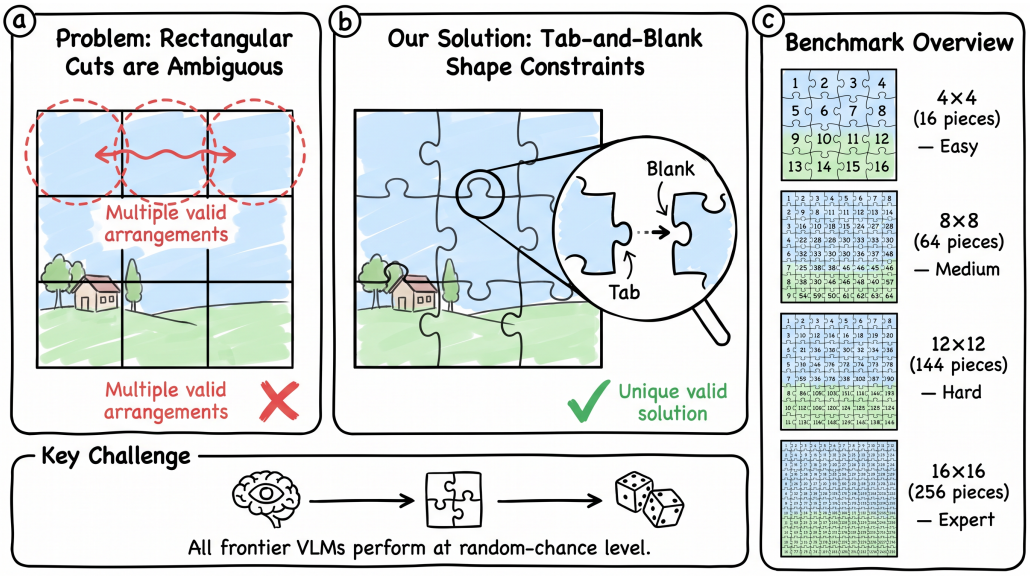


Figure 1: Overview of JigShape. **Left**: Rectangular cuts create ambiguous ground truth in repeated-texture regions; our tab-and-blank constraints enforce local compatibility, yielding unambiguous solutions when combined with visual content. **Right**: Benchmark taxonomy with four grid densities (4×4 to 16×16) and shape vs. no-shape ablation for controlled evaluation.

5×5 puzzles with eroded gaps, on which specialized solvers such as VLHSA (Xu & Liu, 2025) and PuzLM (Elkin et al., 2025) are trained and evaluated. Visual Jigsaw (Wu et al., 2025) uses jigsaw solving as a post-training task to improve MLLM visual understanding across images, videos, and 3D data. These efforts establish spatial reasoning as a challenging frontier for VLMs.

However, existing benchmarks share a common design: they all use simple rectangular cuts. This leads to two fundamental limitations that compromise their diagnostic value (Tab. 1).

- **Ambiguity from rectangular cuts.** When images are divided into rectangular patches, regions with uniform texture (sky, grass, walls) become visually indistinguishable. Consider a 3×3 puzzle of a landscape: the three sky patches may be interchangeable without any perceptual difference. Multiple permutations are thus visually valid, yet ground-truth labels accept only one. This makes evaluation ill-posed: a "wrong" prediction may be perceptually reasonable, conflating genuine reasoning failures with label arbitrariness. Benchmarks using rectangular cuts cannot distinguish whether models fail due to lack of spatial reasoning or due to ambiguous supervision.
- **Low grid density.** Existing benchmarks operate at coarse granularity, typically 2×2 to 5×5 patches (Lyu et al., 2025; Song et al., 2023). A 3×3 grid has only 9! ≈ 363K permutations, while an 8×8 grid has 64! ≈ 10^{89}. Coarse grids limit the difficulty ceiling and may be solvable through simple heuristics (e.g., matching edge colors). Dense grids are necessary to stress-test spatial reasoning at scale and to measure how performance degrades with combinatorial complexity.

We introduce *JigShape*, a large-scale benchmark that addresses both limitations. JigShape features physically interlocking jigsaw pieces with *tab-and-blank geometric constraints*, as illustrated in Fig. 1. Each piece boundary is either a convex tab or a concave blank; adjacent pieces must have complementary edges (tab fits into blank). This design, inspired by real jigsaw puzzles, provides strong local compatibility constraints that substantially reduce the valid search space. Combined with distinct visual content in each piece, these constraints yield unambiguous ground truth, converting jigsaw evaluation from an ill-posed to a well-posed problem and enabling rigorous measurement of joint visual-geometric reasoning. We verify that every puzzle instance has a unique solution using a classical constraint-based solver (Appendix E.4).

The benchmark spans four grid densities (4×4, 8×8, 12×12, 16×16), with 95,468 total instances generated from 23,742 high-resolution source images (Tab. 2). To isolate the effect of geometric constraints, we additionally generate a *no-shape* (rectangular) version of the test set. Comparing per-

formance with and without shape constraints directly measures whether VLMs can utilize boundary geometry, addressing a central question: *Can VLMs jointly reason over visual content and geometric constraints?*

Our evaluation reveals three key findings. First, **zero-shot geometric reasoning is largely absent**: among five frontier models (six configurations), only GPT-5.5 exceeds random baseline on 4×4 puzzles (69.65% vs. 6.25%), while others, including reasoning-enabled variants like Claude Opus 4.8 and Grok-4.2, fail entirely. Supervised fine-tuning on Qwen3-VL-8B (Bai et al., 2025) and Gemma3-12B (Google DeepMind, 2024) achieves >97% on 4×4, demonstrating the task is learnable. Second, **all models collapse on larger grids**: GPT-5.5 drops from 70% to 4% on 8×8, and even fine-tuned models fall to <5% on 12×12. This "scaling cliff" suggests current architectures cannot maintain consistent constraint satisfaction as the number of pieces increases. Third, **SFT models rely primarily on geometric constraints**: removing tab-blank boundaries causes Qwen3-VL-8B to drop from 97% to 10% PA, barely above random, indicating models learn to exploit shape cues with limited visual content integration.

Our contributions:

- **Problem formalization**: We cast jigsaw puzzle solving as *joint visual-geometric reasoning*, where tab-and-blank boundaries provide local compatibility constraints that, combined with visual content, yield unambiguous ground truth and convert an ill-posed visual matching task into a well-posed reasoning problem.
- **Large-scale benchmark**: JigShape provides 95K+ instances across four grid densities (4×4 to 16×16), with a shape vs. no-shape ablation that isolates the contribution of geometric constraints.
- **Comprehensive evaluation**: We evaluate five frontier VLMs in six zero-shot configurations and two fine-tuned models, revealing that (1) only GPT-5.5 exceeds random baseline on 4×4, (2) fine-tuning achieves near-perfect 4×4 accuracy, proving the task is learnable, and (3) all models collapse on 8×8 and beyond.

2 Related Work

Spatial Reasoning in VLMs. Spatial reasoning has emerged as a key challenge for vision-language models. VSR (Liu et al., 2023a) evaluates binary spatial relations ("is A left of B?") and finds that even large VLMs struggle with basic positional reasoning. SpatialBench (Chen et al., 2024) extends this to 3D spatial understanding, revealing significant gaps in depth and distance estimation. BLINK (Fu et al., 2024) tests relative depth ordering and shows that models often fail on perceptually obvious comparisons. What's Up (Kamath et al., 2023) systematically probes spatial relation understanding, finding that VLMs exhibit consistent biases and fail to generalize across viewpoints. Other benchmarks probe compositional reasoning (Thrush et al., 2022; Parcalabescu et al., 2022) and visual question answering (Johnson et al., 2017; Hudson & Manning, 2019), revealing broader gaps in structured visual understanding. These benchmarks evaluate *recognition* of existing spatial relations. In contrast, JigShape tests *reconstruction*: assembling a coherent structure from fragments by satisfying geometric constraints, a fundamentally harder task requiring integration of local cues with global consistency.

Jigsaw Puzzle Benchmarks for VLMs. Recent work has adopted jigsaw puzzles as a testbed for spatial reasoning in VLMs. Jigsaw-Puzzles (Lyu et al., 2025) designs five cognitive tasks around 2×2 and 3×3 grids on 1,100 real-world images; the best model reaches 77% overall accuracy but only 30% on open-ended order generation. VGRP-Bench (Ren et al., 2025) evaluates 20 rule-based constraint puzzles (e.g., Sudoku, Kakuro, Nonogram) on a grid, testing logical reasoning rather than spatial reconstruction. LEGO-Puzzles (Tang et al., 2025) tests sequential assembly of block structures, requiring step-by-step spatial planning. Visual Jigsaw (Wu et al., 2025) proposes using jigsaw puzzle solving as a self-supervised post-training task to enhance MLLM visual perception; while they demonstrate that training on jigsaw tasks improves downstream performance, their work focuses on training rather than evaluation and uses rectangular cuts without geometric constraints. All existing benchmarks use rectangular cuts, which create ambiguous ground truth when images contain repeated textures (sky, grass, uniform surfaces). Multiple arrangements may be perceptually equivalent, yet only one is marked correct, conflating reasoning failures with label ar-

Table 1: Comparison with existing jigsaw-style benchmarks. Prior work uses rectangular cuts with limited grid sizes and lacks geometric constraints. JigShape introduces tab-and-blank shape constraints that, combined with visual content, ensure unambiguous ground truth, plus dense grids up to 16×16 and controlled shape vs. no-shape ablation.

Benchmark	Task Type	Grid Sizes	Max Pieces	Shape Constraint	Guaranteed Unique	Shape Ablation	Scale
Jigsaw-Puzzles (Lyu et al., 2025)	VLM Eval	2×2, 3×3	9	✗	✗	✗	1.1K
JPwLEG (Song et al., 2023)	Model Training	3×3, 5×5	25	✗	✗	✗	12K
Visual Jigsaw (Wu et al., 2025)	Model Training	3×3	9	✗	✗	✗	118K
Jigsaw-R1 (Wang et al., 2025)	Model Training	2×1 to 2×2	4	✗	✗	✗	83K
JigShape (Ours)	VLM Eval	4×4 to 16×16	**256**	✓	✓	✓	**95K**

bitrariness. JigShape addresses this fundamental limitation through geometric shape constraints that ensure unique solutions.

Self-Supervised Representation Learning. Jigsaw puzzle solving was introduced as a self-supervised pretext task by Noroozi & Favaro (2016) and later shown to improve domain generalization (Carlucci et al., 2019) and Vision Transformer robustness (Chen et al., 2023; Ren et al., 2023). More recently, jigsaw puzzles have emerged as a training signal for large vision-language models: Visual Jigsaw (Wu et al., 2025) uses jigsaw solving as RL post-training, improving fine-grained perception and 3D understanding; Jigsaw-R1 (Wang et al., 2025) finds that rule-based visual RL on jigsaw generalizes more effectively than supervised fine-tuning; PuzzleCraft (Jeddi et al., 2025) introduces exploration-aware curriculum learning with puzzle-based RLVR; and Spatial-SSRL (Liu et al., 2025) derives self-supervised spatial signals: including shuffled patch reordering, to improve spatial understanding via GRPO. These works demonstrate that jigsaw training improves general model capabilities beyond puzzle solving, motivating benchmarks like JigShape that serve both as evaluation instruments and as training data sources for spatial reasoning.

Computational Jigsaw Puzzle Solving. The computational puzzle-solving literature spans decades. Classical approaches use pairwise compatibility measures based on color gradients and texture continuity (Pomeranz et al., 2011), genetic algorithms for large puzzles (Sholomon et al., 2013), and loop constraints for global consistency (Son et al., 2014). Recent deep learning methods include SD2RL (Song et al., 2023), which introduces the JPwLEG dataset and a Siamese DQN for puzzles with eroded gaps, VLHSA (Xu & Liu, 2025), which applies vision-language alignment to the same setting, PuzLM (Elkin et al., 2025), which formulates puzzle solving as sequence-to-sequence prediction, and DiffAssemble (Scarpellini et al., 2024), which uses graph diffusion for 2D and 3D reassembly. These methods train specialized models on puzzle data (typically 3×3 to 5×5 grids). Our work differs in two ways: (1) we evaluate *zero-shot* reasoning in general-purpose VLMs rather than training puzzle-specific models, and (2) we introduce geometric shape constraints as a benchmark dimension, enabling controlled ablation of visual vs. geometric reasoning.

3 BENCHMARK

JigShape is a geometry-centric benchmark where adjacent pieces must have complementary edge shapes (tab matches blank). These geometric constraints, combined with distinct visual content in each piece, yield unambiguous ground truth. This section describes the formal task definition, shape constraint mechanism, and dataset construction.

3.1 TASK FORMULATION

Given a shuffled set of jigsaw pieces with visible shape boundaries, the task is to predict the correct grid position for each piece. Formally, a puzzle instance is a tuple $P = (I, G, E, \pi)$, where I is the source image, G is the $N \times N$ grid, $E : \mathcal{E} \to \{\text{tab}, \text{blank}, \text{flat}\}$ assigns edge types to piece boundaries, and $\pi : \{0, \ldots, N^2 - 1\} \to \{0, \ldots, N^2 - 1\}$ is a random permutation mapping piece IDs to grid positions.

The model receives an image showing all N^2 pieces arranged in a grid layout ordered by piece ID (not by correct position), with each piece clearly labeled with its ID number. The model must output a mapping from each piece ID to its predicted (row, col) position in the solved puzzle.

The *geometric compatibility constraint* requires that adjacent pieces have complementary touching edges:

$$E(p_i, \text{right}) = \text{tab} \Leftrightarrow E(p_j, \text{left}) = \text{blank}. \quad (1)$$

Shape constraints drastically reduce the valid search space by enforcing local compatibility. Combined with distinct visual content in each piece, these constraints yield a unique ground truth. Unlike rectangular-cut puzzles where repeated textures create multiple valid-looking arrangements, our tab-and-blank design ensures every instance has exactly one correct solution.

3.2 Shape Constraint Mechanism

Real jigsaw puzzles use interlocking shapes to ensure pieces fit together unambiguously. We implement this mechanism with three edge types:

- **Tab**: A convex semicircular protrusion extending outward from the piece boundary. Tabs carry visual content from the neighboring piece, creating a natural "lock" that must fit into a corresponding blank.
- **Blank**: A concave semicircular indentation cut into the piece boundary. Blanks receive the tab from an adjacent piece, forming the complementary half of the interlock.
- **Flat**: A straight edge appearing only on the outer boundary of the puzzle. Corner pieces have two flat edges; edge pieces have one; interior pieces have none.

Each piece's *edge signature* $(e_{\text{top}}, e_{\text{right}}, e_{\text{bottom}}, e_{\text{left}})$ encodes its geometric constraints. For an $N \times N$ puzzle, there are $(N-1) \times N$ horizontal internal edges and $N \times (N-1)$ vertical internal edges, each randomly assigned tab-blank polarity during generation. This randomization ensures that edge signatures vary across instances, preventing models from memorizing positional patterns.

The shape constraints provide two types of information for solving: (1) *local compatibility*, where adjacent pieces must have complementary edges, and (2) *global position cues*, where flat edges indicate boundary positions and the number of flat edges distinguishes corners (2), edges (1), and interior pieces (0).

3.3 Grid Configurations

We construct puzzles at four density levels to measure how spatial reasoning scales with complexity (Tab. 2). The grid sizes are chosen to span a wide range of combinatorial difficulty while remaining visually parseable:

- **4×4 (16 pieces)**: Entry-level difficulty with $\sim 2 \times 10^{13}$ permutations. Pieces are large enough that visual content provides strong cues. This setting tests basic spatial reasoning.
- **8×8 (64 pieces)**: Medium difficulty with $\sim 10^{89}$ permutations. Individual pieces contain less distinctive content, increasing reliance on shape constraints and local edge matching.
- **12×12 (144 pieces)**: High difficulty with $\sim 10^{249}$ permutations. Pieces are small and often visually similar, making geometric constraints essential for disambiguation.
- **16×16 (256 pieces)**: Extreme difficulty with $\sim 10^{507}$ permutations. This configuration stress-tests the limits of current VLM architectures.

Tab geometry is scaled with grid density: the tab ratio (tab depth as a fraction of piece side length) increases from 0.15 at 4×4 to 0.28 at 16×16, ensuring tabs remain visually perceptible as piece size decreases. The final layout sizes range from 724×744 pixels (4×4) to 3292×3312 pixels (16×16), all within the input resolution limits of modern VLMs.

3.4 Source Images and Statistics

High-resolution source images are essential for dense grid configurations, as each piece must contain sufficient visual detail for potential matching. We curate images from three sources:

- **DIV2K** (Agustsson & Timofte, 2017): 900 images at ∼2K resolution, originally designed for super-resolution research. These images feature diverse natural scenes with rich textures.

Table 2: Grid configurations in JigShape. Tab ratio indicates tab depth as a fraction of piece side length. Permutations show the combinatorial search space.

Grid	Pieces	Tab Ratio	Layout Size	Permutations	Difficulty
4×4	16	0.15	724×744	$\sim 2 \times 10^{13}$	Easy
8×8	64	0.20	1532×1552	$\sim 10^{89}$	Medium
12×12	144	0.24	2412×2432	$\sim 10^{249}$	Hard
16×16	256	0.28	3292×3312	$\sim 10^{507}$	Expert



Figure 2: Example 12×12 puzzle instance from JigShape. Left: shuffled layout with 144 labeled pieces showing tab-and-blank edge shapes. Right: original source image representing ground truth. Geometric constraints (tabs must match blanks) enforce local compatibility, while visual content disambiguates pieces with identical edge signatures. See Appendix F for examples at other grid sizes.

- **DIV8K** (Gu et al., 2019): 1,500 images at up to 8K resolution, providing extremely high detail suitable for our densest 16×16 configuration.
- **Unsplash** (Unsplash Inc., 2024): 21,342 curated high-resolution photographs covering landscapes, architecture, objects, and abstract patterns. This forms the majority of our dataset and ensures broad visual diversity.

In total, 23,742 unique source images are used. Each image generates one puzzle instance per grid configuration, yielding 95,468 total instances. We split the data by source image to prevent train-test leakage: the same image never appears in both training and evaluation sets, even at different grid sizes (Tab. 3). All experimental results reported in this paper are on the **evaluation set**; the held-out **test set** is reserved for a future public competition to enable fair comparison of methods developed after publication.

No-shape version for ablation. We create a **no-shape** ablation set of 500 paired instances at 4×4 resolution. Each pair shares the same source image and piece ID permutation, differing only in edge type: the shape version uses tab-and-blank edges, while the no-shape version uses rectangular cuts. This paired design enables controlled measurement of whether models utilize geometric boundary information. We restrict this ablation to 4×4 because larger grids with rectangular cuts lack unique ground truth: regions with repeated textures (sky, grass, walls) become interchangeable, making multiple arrangements visually valid. See Appendix A for generation details.

3.5 Generation Pipeline

Image Preprocessing. The source image is resized so both dimensions are divisible by the grid size N, yielding piece dimensions $h \times w$. We use high-quality Lanczos resampling (Lanczos, 1956) to preserve detail.

Table 3: Benchmark split statistics. Each source image generates one instance per grid size. For ablation, a no-shape (rectangular) version is provided only for 4×4, as larger grids without shape constraints have ambiguous ground truth. The held-out test set is reserved for a future competition.

Split	4×4	8×8	12×12	16×16	Total
Train	22,992	22,992	22,992	22,992	91,968
Eval	250	250	250	250	1,000
Test (held-out)	500	500	500	500	2,000
Test (No-Shape)	500	—	—	—	500
Total	24,242	23,742	23,742	23,742	**95,468**

Edge Assignment. For each internal edge between adjacent grid cells, we randomly assign tab-blank polarity with equal probability, while boundary edges are set to flat. This random assignment ensures diverse edge configurations across instances, preventing models from memorizing fixed patterns.

Shape Mask and Content Extraction. We generate an alpha mask for each piece encoding its shape. Tabs are semicircular protrusions with radius $r = \text{tab_ratio} \times \min(h, w)/2$, centered on the edge midpoint; blanks are corresponding indentations. Piece content is extracted using these masks. Crucially, tab regions carry visual content from the neighboring cell, mimicking real jigsaw puzzles where the protruding tab shows part of the adjacent image region. This creates natural visual continuity cues that a reasoning model could exploit.

Layout Composition. Piece IDs (0 to $N^2 - 1$) are randomly permuted to prevent any correlation between ID and grid position. Pieces are arranged in a regular grid layout ordered by ID (not by correct position), with each piece clearly labeled with its numeric ID. Importantly, pieces are presented in their original orientation without rotation; this ensures a unique ground-truth solution exists, as allowing rotations would introduce ambiguity when pieces have symmetric shapes. The pipeline is fully deterministic given a random seed, enabling reproducibility. We verify correctness by confirming that pieces can be recomposed to reconstruct the original image exactly (Fig. 2; additional examples in Appendix F).

4 Evaluation Protocol

4.1 Metrics

We report four metrics (formal definitions in Appendix D): **Piece Accuracy (PA)**: fraction of pieces placed in correct positions. **Exact Match (EM)**: fraction of instances with all pieces correct (PA = 100%). **Adjacency Accuracy (AA)**: fraction of ground-truth adjacent piece pairs that are also adjacent in the prediction, measuring local coherence. **Shape Compatibility (SC)**: fraction of predicted adjacencies with geometrically compatible edges (tab matches blank). The random baseline is grid-specific due to flat boundary edges (e.g., 31% for 4×4, 44% for 16×16; see Appendix D for exact derivation); SC above this baseline indicates shape constraint utilization.

5 Experiments

Models. We evaluate five frontier VLMs in six zero-shot configurations: **GPT-5.5** and **GPT-5.4-mini** (OpenAI, 2023); **Grok-4.2** in reasoning and non-reasoning modes; **Llama-4-Maverick**; and **Claude Opus 4.8** (Anthropic, 2024). Additionally, we fine-tune **Qwen3-VL-8B** (Bai et al., 2025) and **Gemma3-12B** (Google DeepMind, 2024) on mixed-grid training data (4×4: 3K, 8×8: 2K, 12×12: 1K, 16×16: 0.5K; 6.5K total) as supervised baselines.

Setup. All models are evaluated on 250 eval samples per grid with standardized prompts (Appendix B). We report PA, EM, AA, SC (Sec. 4.1; formal definitions in Appendix D). Random baseline: PA = $1/N^2$, AA = $4/[N(N+1)]$, SC is grid-specific due to flat boundary edges (30.8%–44.1%;

Table 4: Piece Accuracy (PA) and Exact Match (EM) on JigShape. Higher is better. All models evaluated on 250 eval samples per grid. Only GPT-5.5 exceeds random baseline significantly on 4×4; all models fail on larger grids.

	4×4		8×8		12×12		16×16	
Model	PA	EM	PA	EM	PA	EM	PA	EM
Frontier VLMs (Zero-Shot)								
GPT-5.5	**69.65**	**26.40**	**4.37**	0.00	0.67	0.00	—	—
Claude Opus 4.8	8.50	0.00	1.38	0.00	0.38	0.00	—	—
Grok-4.2-reasoning	11.97	0.00	1.66	0.00	0.67	0.00	—	—
Grok-4.2 (non-reasoning)	6.90	0.00	1.54	0.00	0.64	0.00	—	—
GPT-5.4-mini	6.75	0.00	1.57	0.00	0.52	0.00	—	—
Llama-4-Maverick	6.62	0.00	1.57	0.00	**0.69**	0.00	—	—
Supervised Fine-Tuned Models (mixed-grid training)								
Qwen3-VL-8B (SFT)	97.28	**90.80**	27.34	0.00	3.44	0.00	**0.40**	0.00
Gemma3-12B (SFT)	**97.82**	89.20	**33.58**	**0.40**	**4.57**	0.00	0.35	0.00
Random Baseline	6.25	0.00	1.56	0.00	0.69	0.00	0.39	0.00

Table 5: Adjacency Accuracy (AA) and Shape Compatibility (SC) on JigShape. AA measures local coherence; SC measures geometric constraint utilization. Random baselines are grid-specific (see Appendix D).

	4×4		8×8		12×12		16×16	
Model	AA	SC	AA	SC	AA	SC	AA	SC
Frontier VLMs (Zero-Shot)								
GPT-5.5	**54.22**	**71.33**	**10.58**	**42.47**	2.90	42.28	—	—
Claude Opus 4.8	18.02	33.78	4.92	39.42	2.65	42.14	—	—
Grok-4.2-reasoning	22.22	37.62	5.50	39.58	2.66	**42.41**	—	—
Grok-4.2 (non-reasoning)	17.22	32.88	5.75	39.76	2.74	42.18	—	—
GPT-5.4-mini	18.48	33.09	4.92	39.58	2.58	42.34	—	—
Llama-4-Maverick	16.40	33.60	5.06	39.46	**2.85**	42.24	—	—
Supervised Fine-Tuned Models (mixed-grid training)								
Qwen3-VL-8B (SFT)	95.77	98.40	15.58	70.75	0.59	52.11	**0.00**	**45.02**
Gemma3-12B (SFT)	**97.35**	**98.95**	**38.44**	**87.06**	**10.88**	**62.10**	0.00	45.12
Random Baseline	20.00	30.83	5.56	38.99	2.56	42.34	1.47	44.13

see Appendix D for exact derivation). We omit 16×16 evaluation for frontier VLMs: since all zero-shot models already perform at random chance on 12×12 (PA $< 1\%$), further evaluation on 16×16 would yield no additional insight. See Appendix C for full experiment details.

5.1 Main Results

GPT-5.5 is the only zero-shot model to exceed random baseline. Tab. 4 shows that GPT-5.5 achieves 69.65% PA and 26.40% EM on 4×4 puzzles, far exceeding random baseline (6.25% PA). Tab. 5 confirms this with high AA (54.22% vs. 20.0% random) and SC (71.33% vs. 30.8% random), indicating GPT-5.5 leverages both visual content and geometric constraints. In contrast, all other zero-shot models, including Claude Opus 4.8, Grok-4.2-reasoning, and Llama-4-Maverick, achieve near-random PA (6.62–11.97%) with 0% EM. Their 4×4 SC values (32.9–37.6%) are close to the random baseline (30.8%), indicating they do not meaningfully utilize shape constraints.

Extended reasoning provides limited benefit. Grok-4.2-reasoning outperforms its non-reasoning variant by 5 percentage points on 4×4 (11.97% vs. 6.90% PA), but this remains far below GPT-5.5 and insufficient for reliable puzzle solving. The improvement vanishes on larger grids. This suggests that current reasoning mechanisms help process visual-spatial information but do not enable geometric constraint satisfaction.

All zero-shot models fail on larger grids. GPT-5.5's advantage disappears at higher complexity: 4.37% PA on 8×8 and 0.67% PA on 12×12, both near random baseline. All zero-shot models converge to chance-level performance as grid density increases. On 8×8 and 12×12, SC values (39.4–42.5%) match the grid-specific random baselines (39.0% and 42.3%, respectively), confirming that models place pieces without regard for edge compatibility on larger grids.

SFT models learn the task but exhibit a scaling cliff. Both SFT models achieve >97% PA and ∼90% EM on 4×4, confirming the task is learnable with supervision. Gemma3-12B outperforms Qwen3-VL-8B across all metrics, particularly on 8×8 (33.58% vs. 27.34% PA; 38.44% vs. 15.58% AA). However, both exhibit sharp performance degradation: 8×8 drops to 27–34% PA, 12×12 to 3–5% PA, and 16×16 to near-random (0.35–0.40% PA vs. 0.39% baseline).

5.2 Ablation Study

We examine which cues support reconstruction and whether changes to training or additional guidance alleviate the degradation on larger grids. The following ablations and diagnostic analyses summarize the main findings; detailed setups and results are provided in Appendix E.

Shape vs. no-shape. Removing tab-and-blank boundaries substantially reduces 4×4 accuracy, with a larger drop for SFT models than for GPT-5.5. Strong performance on shaped puzzles therefore does not transfer reliably to rectangular pieces, suggesting substantial dependence on boundary cues. This comparison also changes the input distribution, so it does not by itself establish that models ignore visual content (Appendix E.1, Tab. 6).

Boundary vs. interior pieces. Position-based analysis shows that corner pieces remain easier than interior pieces as grid size increases. This pattern is consistent with models benefiting from flat-edge position cues while struggling to place pieces that require matching internal boundaries and visual content. It helps localize the loss in accuracy beyond the aggregate scores (Appendix E.2, Tab. 7).

Emphasizing harder training samples. An alternative SFT configuration that upweights low-accuracy grids performs worse than baseline SFT on PA across all grid sizes. Thus, this attempt to emphasize harder puzzles does not overcome the scaling difficulty. Since sample count and training duration also change, the experiment does not isolate the effect of weighting or establish an architectural limit (Appendix E.3, Tab. 8).

Partial ground-truth positions. Providing correct positions for most pieces does not make the remaining placements reliable: with 60 of 64 positions revealed, both models still have low PA on the four unknown pieces. Partial solutions alone are therefore insufficient under the tested prompting protocol, although this experiment does not separate clue interpretation from spatial reasoning (Appendix E.5, Tab. 9).

Visual similarity hints. Grouping pieces by visual similarity reduces Qwen SFT's 8×8 PA by 8.80 percentage points. This finding shows that the tested textual grouping hints do not translate into better reconstruction; it leaves open whether alternative representations of the same information would help (Appendix E.5, Tab. 10).

6 Conclusion

We introduced *JigShape*, a jigsaw benchmark for visual-geometric reasoning with interlocking shape constraints and grids from 4×4 to 16×16. Zero-shot performance is limited for most evaluated VLMs; supervised fine-tuning enables strong small-grid performance but does not prevent sharp degradation as grids grow. Removing shape cues substantially reduces accuracy, especially for fine-tuned models. These findings motivate methods that integrate visual content with geometric constraints and sustain consistent reconstruction at larger scales.

7 AI USAGE STATEMENT

We used generative AI tools to assist with method implementation and dataset cleaning and reformatting. During manuscript revision, we also used these tools to provide feedback on experimental methodology and result interpretation, and to help correct the mathematical derivation of the random shape-compatibility baseline. Additionally, we used generative AI tools to edit the manuscript for readability and refine its organization. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

8 ETHICS STATEMENT & REPRODUCIBILITY STATEMENT

All source images are from publicly available datasets (DIV2K, DIV8K under academic license; Unsplash under permissive license). We release the complete generation pipeline, evaluation code, and configuration to facilitate reproducibility and further research.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Anthropic Technical Report*, 2024.

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

Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2229–2238, 2019.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024.

Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan A. K. Suykens. Jigsaw-ViT: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023.

Gur Elkin, Ofir Itzhak Shahar, and Ohad Ben-Shahar. PuzLM: Solving jigsaw puzzles with sequence-to-sequence language models. *arXiv preprint arXiv:2511.06315*, 2025.

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. *arXiv preprint arXiv:2404.12390*, 2024.

Google DeepMind. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Google DeepMind. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. DIV8K: Diverse 8k resolution image dataset. In *ICCV Workshops*, 2019.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. *CVPR*, 2019.

Ahmadreza Jeddi, Hakki Can Karaimer, Hue Nguyen, Zhongling Wang, Ke Zhao, Javad Rajabi, Ran Zhang, Raghav Goyal, Konstantinos G. Derpanis, Babak Taati, and Radek Grzeszczuk. Puzzlecraft: Exploration-aware curriculum learning for puzzle-based RLVR in VLMs. *arXiv preprint arXiv:2512.14944*, 2025.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CVPR*, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023.

Cornelius Lanczos. Applied analysis. *Prentice-Hall*, 1956.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. In *Transactions of the Association for Computational Linguistics*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *NeurIPS*, 2023b.

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-SSRL: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025.

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, 2025. URL `https://arxiv.org/abs/2505.20728`.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pp. 69–84, 2016.

OpenAI. GPT-4V(ision) system card. *OpenAI Technical Report*, 2023.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *ACL*, pp. 8253–8280, May 2022.

Dolev Pomeranz, Michal Shemesh, and Ohad Ben-Shahar. A fully automated greedy square jigsaw puzzle solver. *CVPR*, 2011.

Bin Ren, Yahui Liu, Yue Song, Wei Bi, Rita Cucchiara, Nicu Sebe, and Wei Wang. Masked jigsaw puzzle: A versatile position embedding for vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20382–20391, 2023.

Yufan Ren, Konstantinos Tertikas, Shalini Maiti, Junlin Han, Tong Zhang, Sabine Süsstrunk, and Filippos Kokkinos. VGRP-Bench: Visual grid reasoning puzzle benchmark for large vision-language models. *arXiv preprint arXiv:2503.23064*, 2025.

Gianluca Scarpellini, Stefano Fiorini, Francesco Giuliari, Pietro Morerio, and Alessio Del Bue. DiffAssemble: A unified graph-diffusion model for 2d and 3d reassembly. In *CVPR*, 2024.

Dror Sholomon, Omid David, and Nathan S Netanyahu. A genetic algorithm-based solver for very large jigsaw puzzles. *CVPR*, 2013.

Kilho Son, James Hays, and David B Cooper. Solving square jigsaw puzzles with loop constraints. *ECCV*, 2014.

Xingke Song, Jiahuan Jin, Chenglin Yao, Shihe Wang, Jianfeng Ren, and Ruibin Bai. Siamese-discriminant deep reinforcement learning for solving jigsaw puzzles with large eroded gaps. *AAAI*, 37(2):2303–2311, Jun. 2023.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. LEGO-Puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. *CVPR*, 2022.

Unsplash Inc. Unsplash. `https://unsplash.com`, 2024. High-resolution photographs under Unsplash License.

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*, 2025.

Penghao Wu, Yushan Zhang, Haiwen Diao, Bo Li, Lewei Lu, and Ziwei Liu. Visual jigsaw post-training improves MLLMs. *CoRR*, abs/2509.25190, 2025.

Zhuoning Xu and Xinyan Liu. VLHSA: Vision-language hierarchical semantic alignment for jigsaw puzzle solving with eroded gaps. *arXiv preprint arXiv:2509.25202*, 2025.

A Generation Pipeline Details

Shape version. Given source image I and grid size N, we resize I so dimensions are divisible by N. For each internal edge, we randomly sample polarity (tab-blank or blank-tab); boundary edges are flat. Tabs are semicircular with radius $r = \text{tab_ratio} \times \min(h, w)/2$. Content is extracted with tab regions carrying neighboring pixels. Piece IDs (0 to N^2-1) are randomly permuted, and pieces are arranged by ID with numeric labels. The pipeline is deterministic given a seed.

No-shape version (for ablation). For the ablation study, we generate a **no-shape** (rectangular) version of the 4×4 test set. The no-shape pipeline uses the same source images, the same random seed (thus identical piece ID permutations), and the same grid layout, but replaces tab-and-blank edges with straight cuts. This produces rectangular pieces without geometric interlocking. The *only* difference between shape and no-shape versions is the presence of tab-blank boundaries, enabling controlled measurement of whether models utilize geometric constraints. We restrict no-shape generation to 4×4 because larger grids with rectangular cuts have ambiguous ground truth: visually similar regions (e.g., sky, grass) become interchangeable, making multiple arrangements equally valid.

B Prompt Templates

All models receive identical prompts consisting of a system prompt and a user prompt. The system prompt establishes the task context and emphasizes that piece IDs are randomly assigned. The user prompt specifies the grid size and output format. We provide separate prompts for shape-constrained and no-shape (ablation) conditions.

Shape-Constrained Prompts. System Prompt:

```
You are an expert visual puzzle solver. You must carefully analyze the image provided and
    determine the correct arrangement of puzzle pieces based on:
1. The visual content of each piece (colors, textures, patterns, objects)
2. The shape of piece edges (tabs and blanks must match between adjacent pieces)
3. Corner pieces have 2 flat edges, edge pieces have 1 flat edge

CRITICAL: The pieces are RANDOMLY SHUFFLED with RANDOM labels. The piece labeled "0" is NOT
    necessarily at position [0,0]. The piece labeled "3" is NOT necessarily a corner piece.
    You MUST look at the ACTUAL IMAGE of each piece to determine its correct position. Do
    NOT assume any relationship between piece numbers and positions.
```

User Prompt:

```
This is a NxN jigsaw puzzle with N^2 pieces labeled 0 to N^2-1.

Analyze the image carefully:
1. Identify the 4 corner pieces (2 flat edges each)
2. Identify the edge pieces (1 flat edge each)
3. Match pieces by their visual content and edge shapes

Output ONLY a JSON object mapping piece_id to [row, col]:
{"0": [row, col], "1": [row, col], ...}

Rows and columns are 0-indexed from top-left.
```

No-Shape Prompts (Ablation). System Prompt:

```
You are an expert visual puzzle solver. You must carefully analyze the image provided and
     determine the correct arrangement of puzzle pieces based on:
1. The visual content of each piece (colors, textures, patterns, objects)
2. How adjacent pieces connect visually (matching edges, continuous patterns, color
     continuity)

NOTE: These puzzle pieces are simple rectangles without interlocking shapes. You must rely
     ONLY on visual content to determine the correct arrangement.

CRITICAL: The pieces are RANDOMLY SHUFFLED with RANDOM labels. The piece labeled "0" is NOT
     necessarily at position [0,0]. You MUST look at the ACTUAL IMAGE of each piece to
     determine its correct position. Do NOT assume any relationship between piece numbers and
     positions.
```

User Prompt:

```
This is a NxN jigsaw puzzle with N^2 rectangular pieces labeled 0 to N^2-1.

Analyze the image carefully:
1. Look at the visual content of each piece (colors, textures, patterns, objects)
2. Determine how pieces connect by matching visual content at edges
3. Reconstruct the original image by finding the correct position for each piece

Output ONLY a JSON object mapping piece_id to [row, col]:
{"0": [row, col], "1": [row, col], ...}

Rows and columns are 0-indexed from top-left.
```

C Experiment Details

Frontier VLMs (Zero-Shot). We evaluate five frontier VLMs in six zero-shot configurations with identical prompts: GPT-5.5 and GPT-5.4-mini (OpenAI, May 2025); Grok-4.2 with and without extended reasoning (xAI); Llama-4-Maverick-17B-128E-Instruct-FP8 (Meta); Claude Opus 4.8 (Anthropic). All models support high-resolution image input and 128K+ token context windows, sufficient for processing 16×16 puzzle layouts with 256 labeled pieces. For models with reasoning capabilities, we use their reasoning-enabled modes: GPT-5.5 performs internal chain-of-thought reasoning before outputting JSON (similar to o1/o3 models), and Grok-4.2-reasoning uses `reasoning_effort: medium`. While our prompt requests JSON-only output for parsing, these models perform extended reasoning internally. We separately evaluate Grok without reasoning mode as a control, showing only modest improvement from explicit reasoning (11.97% vs. 6.90% PA on 4×4).

Fine-Tuned Models. We fine-tune Qwen3-VL-8B and Gemma3-12B on a mixed-grid training set with weighted sampling (4×4: 3K, 8×8: 2K, 12×12: 1K, 16×16: 0.5K; 6.5K total). Training uses 1 epoch, learning rate 1×10^{-5}, effective batch size 64, BF16 mixed precision, and Flash Attention 2.

Evaluation Protocol. All models are evaluated on the eval set (250 instances per grid size, 1,000 total). Model outputs are parsed to extract piece ID to position mappings; responses with invalid JSON receive zero accuracy on all metrics.

D Evaluation Metrics

We define four evaluation metrics to comprehensively assess model performance on jigsaw puzzle solving. Let N denote the grid size, $\hat{p} : \{0, \ldots, N^2 - 1\} \to \{0, \ldots, N - 1\}^2$ denote the predicted position mapping, and p^* denote the ground-truth mapping.

Piece Accuracy (PA). The fraction of pieces placed in their correct positions:

$$\text{PA} = \frac{1}{N^2} \sum_{i=0}^{N^2-1} \mathbf{1}[\hat{p}(i) = p^*(i)] \tag{2}$$

PA measures per-piece correctness without considering spatial relationships between pieces.

Exact Match (EM). A binary indicator of whether the entire puzzle is solved correctly:

$$\text{EM} = \mathbf{1}[\forall i : \hat{p}(i) = p^*(i)] \tag{3}$$

EM equals 1 only when all N^2 pieces are in their correct positions (PA = 100%).

Adjacency Accuracy (AA). The fraction of correctly predicted adjacent piece pairs. We define the ground-truth adjacency set $\mathcal{A}^*$ as all piece ID pairs (i, j) where i and j are horizontally or vertically adjacent in the solved puzzle. For an $N \times N$ grid, $|\mathcal{A}^*| = 2N(N - 1)$. The predicted adjacency set $\hat{\mathcal{A}}$ is computed analogously from $\hat{p}$:

$$\text{AA} = \frac{|\hat{\mathcal{A}} \cap \mathcal{A}^*|}{|\mathcal{A}^*|} \tag{4}$$

AA measures local coherence: a model can achieve high AA by correctly grouping pieces locally even if global positions are wrong. Under a uniformly random permutation, the probability that any given ground-truth adjacent pair (i, j) lands on an adjacent position pair equals the number of adjacent position pairs divided by the total number of ordered position pairs: $\mathbb{E}[\text{AA}] = \frac{2 \cdot 2N(N-1)}{N^2(N^2-1)} = \frac{4}{N(N+1)}$, yielding baselines of 20.0% (4×4), 5.6% (8×8), 2.6% (12×12), and 1.5% (16×16).

Shape Compatibility (SC). The fraction of adjacent piece placements with geometrically compatible edges. Each piece carries edge types from its *original* grid position: internal edges are randomly assigned tab or blank, while boundary edges are flat (§3.2). Let $\mathcal{E}^{\hat{p}}$ be the set of internal edges in the predicted arrangement, and $c(e) = 1$ if the two facing edges are complementary (tab–blank):

$$\text{SC} = \frac{1}{|\mathcal{E}^{\hat{p}}|} \sum_{e \in \mathcal{E}^{\hat{p}}} c(e) \tag{5}$$

Under random placement, SC $\neq$ 50% because boundary pieces carry flat edges that are always incompatible when placed at internal positions. Consider a random internal grid edge where piece i contributes its right edge and piece j contributes its left edge. Let $M = N(N-1)$ be the number of horizontal internal edges in the original puzzle, and let T_R and B_L count tab-right and blank-left pieces. Each tab-right edge is paired with a blank-left edge, so $T_R = B_L$ for every generated puzzle; their common count is random, with $T_R \sim \text{Binomial}(M, 1/2)$ and $\mathbb{E}[T_R] = M/2$. Let X count pieces that are *both* tab-right and blank-left. Of the $T_R B_L = T_R^2$ ordered tab-right/blank-left pairs, exactly X use the same piece twice and must be excluded. The reverse polarity has the same expected count by symmetry. Thus, averaging over both random edge assignments and uniformly random permutations,

$$\mathbb{E}[\text{SC}] = \frac{2\,\mathbb{E}[T_R^2 - X]}{N^2(N^2 - 1)} = \frac{M^2 + N}{2N^2(N^2 - 1)}, \tag{6}$$

where $\mathbb{E}[T_R^2] = (M^2 + M)/4$ and $\mathbb{E}[X] = N(N-2)/4$. The same derivation applies to vertical edges. This yields baselines of 30.8% (4×4), 39.0% (8×8), 42.3% (12×12), and 44.1% (16×16), confirmed by Monte Carlo simulation (2×10^6 trials). The simple approximation $\frac{1}{2}((N-1)/N)^2$ underestimates the baseline by ignoring the finite-grid correction from paired edge assignments and sampling without replacement. SC above the grid-specific baseline indicates the model utilizes shape constraints.

Table 6: Shape vs. No-Shape ablation on 4×4 puzzles. "Shape" uses tab-blank constraints; "No-Shape" uses rectangular pieces. PA: Piece Accuracy (%), AA: Adjacency Accuracy (%), EM: Exact Match (%).

	Shape			No-Shape		
Model	PA	AA	EM	PA	AA	EM
Frontier VLMs (Zero-Shot)						
GPT-5.5	69.65	54.22	26.40	23.90	25.69	0.40
Supervised Fine-Tuned Models						
Qwen3-VL-8B (SFT)	97.28	95.77	90.80	10.00	30.97	0.00
Gemma3-12B (SFT)	97.82	97.35	89.20	13.78	35.48	0.00
Random Baseline	6.25	20.00	0.00	6.25	20.00	0.00

Table 7: Position-based accuracy (%) for SFT models. Corner pieces (2 flat edges) are more accurate than interior pieces (0 flat edges) across all tested grid sizes.

	Qwen3-VL-8B			Gemma3-12B		
Grid	Corner	Edge	Interior	Corner	Edge	Interior
4×4	99.70	97.00	95.40	99.40	97.90	96.10
8×8	66.60	37.15	16.43	78.80	44.30	21.41
12×12	34.20	6.28	1.08	39.70	8.03	1.78
16×16	5.10	0.54	0.26	1.00	0.29	0.35

E Additional Results

E.1 Shape vs. No-Shape Ablation

We compare tab-and-blank pieces with rectangular pieces on 4×4 puzzles using the same source images and piece ID permutations. Shape-constrained and no-shape prompts are provided in Appendix B. Tab. 6 reports PA, AA, and EM for both conditions.

Sensitivity to removing shape cues. GPT-5.5's PA decreases from 69.65% to 23.90%, a drop of 45.75 percentage points, while EM decreases from 26.40% to 0.40%. Its no-shape PA remains above the random baseline of 6.25%, consistent with an ability to use visual content without interlocking boundaries. SFT models show larger PA drops: Qwen3-VL-8B decreases from 97.28% to 10.00%, and Gemma3-12B from 97.82% to 13.78%. These results indicate that their high accuracy on shaped puzzles is sensitive to the removal of boundary cues.

Interpretation. Removing shape boundaries eliminates both flat-edge position cues and tab-blank compatibility, changes the pixels assigned to individual pieces, and introduces an input distribution shift for shape-trained models. The performance difference therefore measures the combined effect of these changes rather than isolating a single reasoning mechanism. In particular, the drop alone does not establish that SFT models ignore visual content.

E.2 Position-Based Error Analysis

We decompose SFT piece accuracy into corners (two flat edges), non-corner boundary pieces (one flat edge), and interior pieces (no flat edges). Tab. 7 reports the results across all four grid sizes.

On 4×4 puzzles, both models exceed 99% accuracy on corners and reach 95–96% on interior pieces. At 12×12, Gemma3-12B retains 39.70% corner accuracy but only 1.78% interior accuracy; Qwen3-VL-8B exhibits a similar disparity. With fixed piece orientations, flat edges constrain positions directly, whereas interior placements require compatibility and visual matching. This pattern is consistent with a dependence on boundary cues, but position groups also differ in the number of candidate locations, so their raw accuracies do not isolate the contribution of each cue. At 16×16, accuracy is low across all position groups.

Table 8: Comparison of baseline SFT vs. weighted SFT on Qwen3-VL-8B across all metrics. PA: Piece Accuracy (%), EM: Exact Match (%), AA: Adjacency Accuracy (%), SC: Shape Compatibility (%). The weighted configuration emphasizes harder grids but yields lower PA at every grid size.

	Baseline SFT				Weighted SFT			
Grid	PA	EM	AA	SC	PA	EM	AA	SC
4×4	97.28	90.80	95.77	98.40	56.40	0.80	31.68	67.90
8×8	27.34	0.00	15.58	70.75	9.12	0.00	1.09	48.39
12×12	3.44	0.00	0.59	52.11	0.89	0.00	0.00	50.00
16×16	0.40	0.00	0.00	45.02	0.00	0.00	0.00	50.00

E.3 Weighted SFT Training

We explore whether emphasizing harder training samples during fine-tuning can improve performance on larger grid sizes. We implement an inverse-accuracy weighting scheme where training samples are weighted by $(1-\text{PA}_{\text{grid}})$, upweighting harder grids (12×12, 16×16) while downweighting easier grids (4×4) where baseline performance is already high. We train Qwen3-VL-8B for 3 epochs on a mixed-grid training set (20K samples) with these sample weights applied to the loss function.

Tab. 8 shows lower PA under the weighted configuration at every grid size, including a decrease from 27.34% to 9.12% on 8×8. This configuration does not alleviate the observed scaling difficulty. However, it differs from baseline SFT in sample count (20K vs. 6.5K), training duration (3 vs. 1 epoch), and loss weighting. The comparison therefore cannot attribute the degradation to weighting alone or rule out improvements from other training strategies.

E.4 Classical Solver Baseline

To verify that every puzzle instance has a unique solution, we implement a classical constraint-based solver that uses exact edge signature information. Each piece has four edge types (tab, blank, or flat). The solver operates as follows:

Edge Compatibility Rules. Adjacent pieces must have complementary edges: a *tab* fits into a *blank*, while *flat* edges occur only at puzzle boundaries (corners have two flat edges, border pieces have one).

Algorithm. We use backtracking search with constraint propagation. First, we precompute candidate pieces for each grid position based on boundary constraints (e.g., corner positions require exactly two flat edges). Then we fill positions row by row, checking at each step that the placed piece is compatible with all already-placed neighbors. If no valid piece exists, we backtrack.

Results. The solver achieves 100% accuracy across all grid sizes (4×4, 8×8, 12×12, 16×16) on all 500 test instances per size, confirming that every puzzle in JigShape has a unique solution.

E.5 Auxiliary Hints

We investigate whether providing auxiliary information can help models solve larger puzzles. We design two types of hints: (1) partial ground truth positions, and (2) visual similarity groupings.

Partial Ground Truth (Clue Benchmark). We test whether models can leverage partial solutions by revealing some pieces' correct positions upfront. Tab. 9 shows results on 8×8 puzzles with varying amounts of known information.

In the extreme60 setting, 60 out of 64 pieces' correct positions are provided, leaving only 4 pieces to place. PA on these remaining pieces is 25.0% for Qwen SFT and 2.50% for GPT-5.5 (Tab. 9). These results show that providing most of the solution does not yield reliable placement of the remaining pieces under this evaluation protocol; they do not establish whether the errors arise from interpreting the clues or from spatial reasoning.

Table 9: Clue benchmark results on 8×8 puzzles. We provide models with n pieces already at correct positions and measure PA only on the remaining puzzle.

Clue Type	Known	Unknown	Qwen SFT	GPT-5.5
corners	4	60	26.74%	2.31%
edges	28	36	18.76%	3.76%
diagonal	16	48	22%	2.02%
two_rows	16	48	23%	2.62%
center4x4	16	48	21%	3.02%
extreme60	60	4	25.0%	2.50%

Table 10: Effect of visual similarity hints on 8×8 puzzles (Qwen SFT).

Condition	PA
Without hint	27.34%
With hint	18.54%
Difference	−8.80 pp

Visual Similarity Hints. Standard segmentation maps have limited applicability to jigsaw puzzles since we need piece-level grouping rather than pixel-level segmentation. We design an alternative approach: clustering pieces by visual similarity (color histograms in LAB space) and conveying this grouping through natural language to avoid adding visual processing burden. Fig. 3 illustrates the hint visualization.

As shown in Tab. 10, the visual similarity hint reduces PA from 27.34% to 18.54%, a decrease of 8.80 percentage points. We hypothesize two reasons for this failure. First, VLM pretraining data rarely contains such visually dense, spatially structured images. A single 8×8 puzzle packs 64 distinct regions with complex geometric relationships, far outside typical training distributions. Second, the image already encodes rich visual content plus geometric cues (tabs, blanks, edge shapes). Adding textual hints creates information overload: the model must simultaneously process dense visual features, geometric constraints, and auxiliary guidance, exceeding its effective integration capacity.

Together, these experiments show that the tested forms of auxiliary guidance do not yield reliable reconstruction on larger puzzles. They leave open whether failures reflect hint representation, instruction following, visual processing, or constraint reasoning; alternative guidance formats remain to be evaluated.

F Dataset Examples

We present example instances from JigShape at different grid sizes. Each row shows the shuffled puzzle layout (left) that serves as input to the model, and the original source image (right) that represents the ground-truth solution. Piece IDs are randomly assigned and displayed on each piece. The model must determine the correct position mapping from piece ID to grid location. See Fig. 2 in the main text for a 12×12 example.

G Dataset Licensing

DIV2K/DIV8K are released by ETH Zürich under academic license for research purposes. **Unsplash** images are licensed under the Unsplash License, permitting free use.

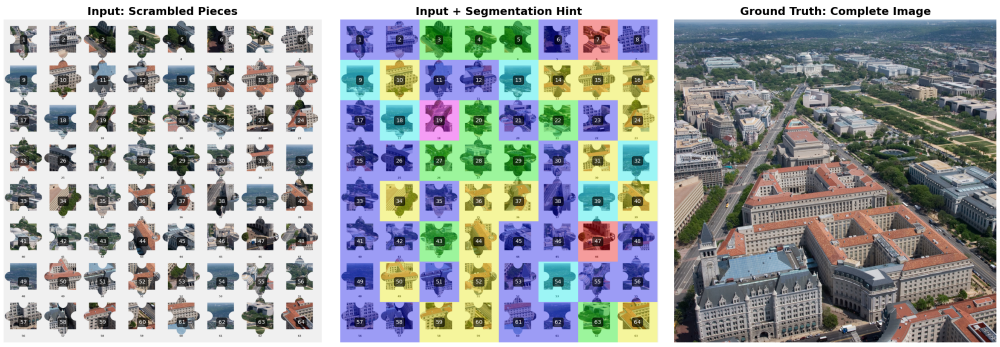
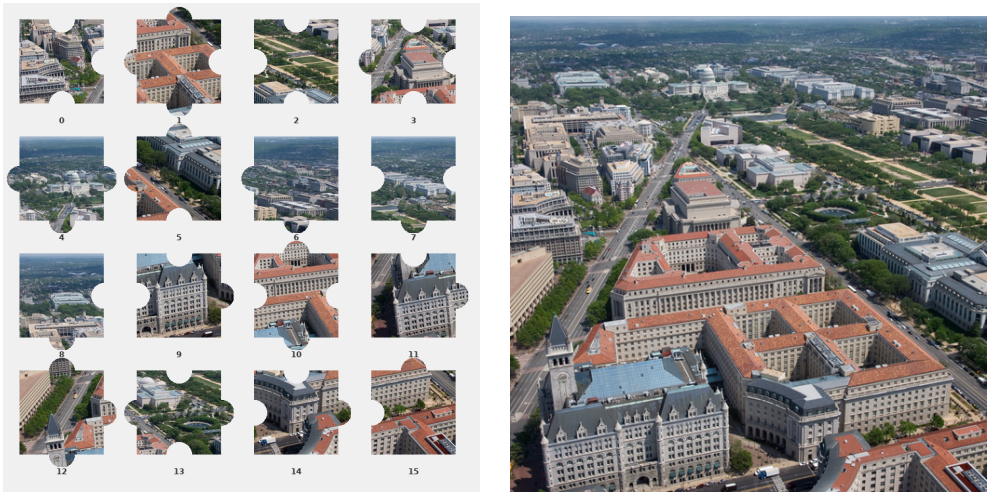
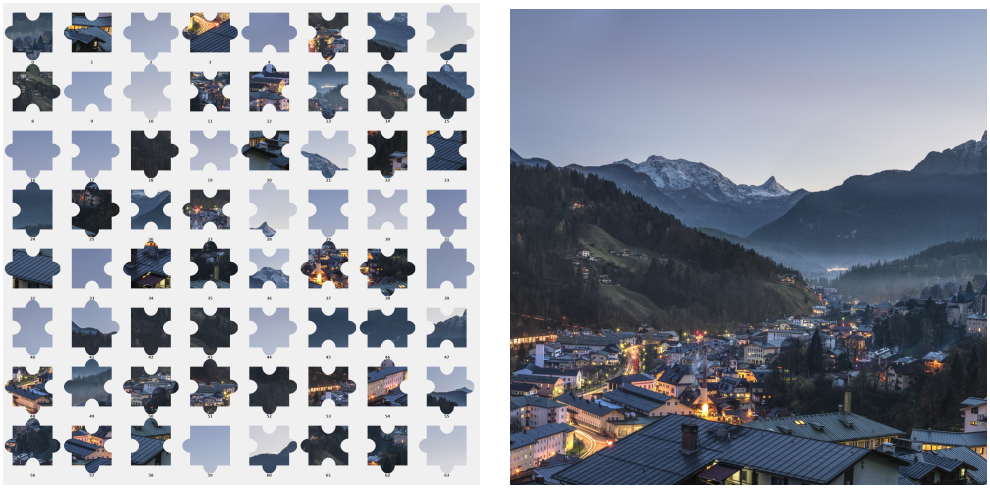


Figure 3: Visual similarity hint. Left: original scrambled pieces. Middle: pieces colored by cluster membership (same color = visually similar). Right: ground truth.



(a) 4×4 grid: 16 pieces



(b) 8×8 grid: 64 pieces



(c) 16×16 grid: 256 pieces

Figure 4: Example instances from JigShape. Left: shuffled puzzle layout with labeled pieces. Right: original source image. As grid size increases, the number of pieces grows quadratically, making constraint satisfaction exponentially harder. At 16×16, models must reason about 256 pieces with 480 internal edges.