PROBING VISUAL PLANNING IN IMAGE EDITING MODELS

Zhimu Zhou[1] **Yanpeng Zhao**[3 †] **Qiuyu Liao**[2] **Bo Zhao**[1] **Xiaojian Ma**[3]
[1]Shanghai Jiao Tong University [2]Renmin University of China
[3]State Key Laboratory of General Artificial Intelligence, BIGAI
https://spatigen.github.io/amaze.io/ https://github.com/spatigen/amaze

ABSTRACT

Visual planning represents a crucial facet of human intelligence, especially in tasks that require complex spatial reasoning and navigation. Yet, in machine learning, this inherently visual problem is often tackled through a verbal-centric lens. While recent research demonstrates the promise of fully visual approaches, they suffer from significant computational inefficiency due to the step-by-step planning-by-generation paradigm. In this work, we present EAR, an editing-as-reasoning paradigm that reformulates visual planning as a single-step image transformation. To isolate intrinsic reasoning from visual recognition, we employ abstract puzzles as probing tasks and introduce AMAZE, a procedurally generated dataset that features the classical Maze and Queen problems, covering distinct, complementary forms of visual planning. The abstract nature of AMAZE also facilitates automatic evaluation of autoregressive and diffusion-based models in terms of both pixel-wise fidelity and logical validity. We assess leading proprietary and open-source editing models. The results show that they all struggle in the zero-shot setting, finetuning on basic scales enables remarkable generalization to larger in-domain scales and out-of-domain scales and geometries. However, our best model that runs on high-end hardware fails to match the zero-shot efficiency of human solvers, highlighting a persistent gap in neural visual reasoning.

1 INTRODUCTION

Spatial reasoning through visual planning is a cornerstone in human intelligence. While humans can navigate complex visual environments intuitively, machine learning models have been predominantly relying on verbal-centric approaches, such as translating these inherently visual reasoning problems into text for large language models (LLMs) (Yang et al., 2022; Wu et al., 2023; Wang et al., 2025a; Dao & Vu, 2025) and framing them as multimodal tasks that rely on vision-language models for text-based chain-of-thought (Li et al., 2023; Xu et al., 2025a; Zhang et al., 2025c;a; Wu et al., 2025b). Recently, reasoning-enhanced generative image models have enabled fully visual alternatives. Some approaches utilize step-wise image-level generation to implement planning but suffer from significant computational inefficiency (Xu et al.,

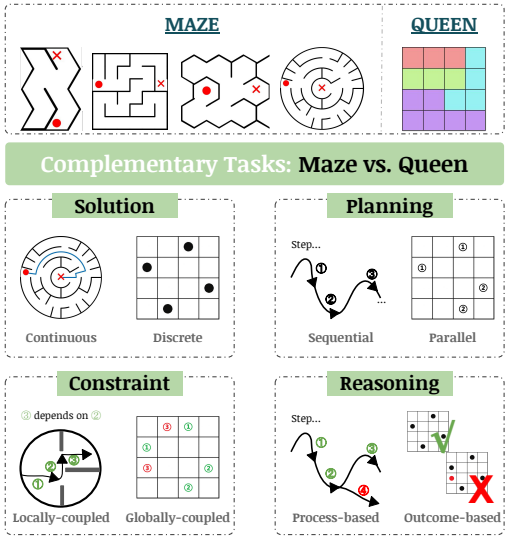


Figure 1: The AMAZE tasks.

†: Project Lead. Contact: piekeniuszwu@gmail.com, yannzhao.ed@gmail.com.

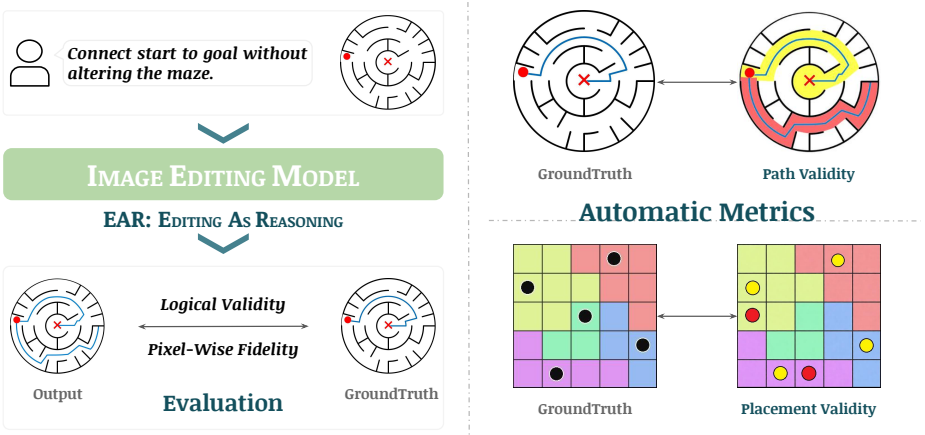


Figure 2: Overview of EAR. Left: the EAR paradigm. Right: automatic evaluation. Yellow and red highlight the generated image's overlap with the solution and non-solution areas, respectively.

2025b); while others attempt direct-generation methods (Wiedemer et al., 2025), yet a comprehensive understanding of the intrinsic visual planning capabilities within these editing-based models remains elusive.

To bridge this gap, we present Editing as Reasoning (EAR), a fully visual reasoning framework that reformulates visual planning as an image editing task. Unlike step-wise approaches, EAR compresses the planning process into an atomic "edit", leveraging the model's internalized spatial and visual priors to produce a complete solution in a single step. By offloading planning to the inherent progressive dynamics of the atomic "edit", EAR eliminates the inductive bias of explicit step-wise modeling, enabling a targeted probing of the intrinsic visual planning capabilities of editing models.

To facilitate in-depth and controlled analysis, we introduce AMAZE, a procedurally generated benchmark for visual planning. AMAZE comprises Maze and Queen tasks that respectively covers two complementary planning paradigms: sequential planning under local constraints and combinatorial planning under global constraints (see Figure 1). AMAZE isolates intrinsic visual reasoning from the confounding factor of complex visual recognition. Its abstract nature enables automatic evaluation metrics that decouple visual reconstruction (pixel-wise fidelity) from logical validity (topological correctness or constraint satisfaction). We incorporate Queen puzzles across different scales (e.g., 4×4 and 10×10), with Maze further featuring diverse geometry types (including triangle, square, hexagon, and circle) to represent varying levels of complexity. This structural diversity enables us to probe the geometric invariance and systematicity of neural visual reasoning, assessing whether models develop generalizable spatial logic or merely exploit local patterns.

We evaluate representative autoregressive and diffusion-based editing models from both proprietary and open-source domains. Our probing experiments are organized along three primary dimensions: (1) *Generalizability* evaluates how well models transfer to unseen geometry types and scales, including both in-domain and out-of-domain settings. (2) *Scaling effect* investigates the scaling law in fine-tuned models for enhanced visual planning, i.e., relationships between performance and quantities of training data and time. (3) *Human comparison* benchmarks the efficiency of visual planning of editing models against human solvers to reveal the performance gap.

Our evaluation reveals that both proprietary and open-source editing models initially struggle with zero-shot visual planning. On Maze, while proprietary models are relatively stronger, finetuning open-source models like Bagel (Deng et al., 2025) on basic 3×3 mazes improves from 0 to 11.54% (PASS@1), outperforming the best proprietary model by an *absolute* 6.14%, and the fine-tuned models show impressive generalizability to larger scales. Notably, diffusion-based models surpass autoregressive models on both Maze and Queen after fine-tuning, suggesting their effectiveness in developing visual reasoning logic. Moreover, our comparison with human solvers on AMAZE reveals a stark efficiency gap: our best model, when running on a single NVIDIA RTX 5090, still lags behind the near-instantaneous, zero-shot reasoning of human solvers. These findings suggest that while EAR is a promising step toward visual intelligence, current architectures still lack the innate spatial inductive biases of humans.

In summary, our contributions are the following:

- We present EAR, an editing-as-reasoning framework for visual reasoning.
- We introduce AMAZE, an abstract visual planning benchmark that covers two complementary planning forms, alongside automatic metrics for both pixel-wise fidelity and logical validity.
- We design controlled experiments to systematically probe intrinsic visual planning across a diverse suite of image editing models.
- We provide an in-depth analysis of the generalizability, scaling effect, and efficiency gap between neural visual planning and human solvers.

2 THE AMAZE BENCHMARK

We propose the AMAZE benchmark, which consists of the classical Maze and Queen puzzles for assessing and analyzing intrinsic visual planning of image-editing models. The reasons that we choose the Maze and Queen tasks as the testbed are three-fold. First, they respectively cover two complementary paradigms of visual planning: locally-constrained sequential planning and globally-constrained combinatorial planning. Second, they minimize visual recognition complexity—comprising primarily abstract structures—thus allowing for isolating the visual planning ability from multimodal understanding dependencies (§2.1). Third, unlike VLM-based evaluation that focuses more on qualitative assessment, they admit automatic metrics to accurately quantify logical correctness (§2.2).

2.1 AUTOMATIC DATA CURATION

We generate both the Maze and Queen tasks procedurally. The complexity of a task is primarily defined by its scale, which ranges from 3×3 to 16×16 for Maze and 4×4 to 10×10 for Queen. The lower and upper bounds on the scale are chosen to avoid trivial solutions while allowing for efficient task generation. In Maze task, we additionally vary the geometry type to cover circular, hexagonal, square, and triangular geometries (Dawson, 2021), enabling fine-grained analysis. For each combination of maze scale and solution algorithm (including both depth-first and breadth-first search), we generate 50 mazes, totaling 2,800 test examples, that is, 700 per geometry type. For Queen, we randomly sample 50 puzzles per scale, resulting in a total of 350 test examples.

2.2 AUTOMATIC EVALUATION METRICS

A fundamental challenge in generative image tasks is that high-quality visual outputs do not necessarily correspond to the right plan. Traditional metrics such as VLM-based critics (Wang et al., 2025b) and fidelity-oriented metrics (Heusel et al., 2017; Zhang et al., 2018)) are inadequate for assessing the logical correctness of visual planning. Since AMAZE is generated procedurally, it enables rule-based metrics that automatically evaluate the correctness of generated plans. Concretely, we define logical validity as the following:

> **Logical validity** measures whether the generated solution matches the goal solution at the cell level. We compute a COVERAGE ratio, which measures the proportion of the goal solution that is correctly generated, and a VIOLATION ratio, which measures the proportion of the generated solution that deviates from the goal solution. We further define PASS as $\max(0, \text{COVERAGE} - \text{VIOLATION})$. PASS = 1 means the generated solution matches the solution structure exactly.

We further complement logical validity with pixel-wise fidelity, defined as:

> **Pixel-wise fidelity** measures pixel-wise differences that are measured using the mean squared error (MSE) between the generated and ground-truth images. We compute it separately for the solution area (cells covered by the goal solution, indicated by MSE-IN) and the non-solution area (indicated by MSE-OUT).

Model	Continuous (Maze) Task						Discrete (Queen) Task					
	Violation↓	Coverage↑	MSE In↓	MSE Out↓	Pass@1↑	Pass@5↑	Violation↓	Coverage↑	MSE In↓	MSE Out↓	Pass@1↑	Pass@5↑
proprietary models												
GPT-image-1	62.88	58.97	41.16	52.76	5.40	6.06	62.91	37.09	11.84	5.87	0.00	2.28
NanoBanana-Pro	47.76	**64.21**	24.20	17.21	4.82	9.28	32.56	67.43	9.10	1.62	**30.35**	**35.58**
Seedream-4.5	16.90	25.67	28.82	30.96	2.14	3.21	76.86	23.14	11.55	5.95	2.86	2.86
open-source models (w/o chain-of-thought reasoning)												
Flux-Kontext-Dev	23.84	30.24	30.96	18.31	0.36	3.57	78.63	21.37	11.48	7.71	0.92	2.34
Qwen-Image-Edit	19.37	28.51	18.82	5.70	1.43	2.14	69.52	30.47	8.83	5.30	2.86	4.00
Bagel	28.91	27.15	11.64	5.84	0.00	1.00	61.57	38.43	8.94	1.22	0.00	0.00
Janus-Pro	**5.41**	1.85	57.47	76.80	0.00	0.00	84.24	15.76	12.97	9.83	0.00	0.57
Bagel (fine-tuned)	12.21	51.02	**8.66**	**3.07**	11.54	**23.64**	68.27	31.73	**6.05**	**0.63**	14.57	14.29
Janus-Pro (fine-tuned)	35.60	23.33	55.99	50.94	1.43	2.22	**16.07**	**83.93**	7.91	1.38	12.57	13.03
w/ chain-of-thought reasoning												
Bagel	34.06	30.31	14.77	3.97	0.00	0.57	98.41	1.59	9.63	1.40	0.00	0.00
Bagel (fine-tuned)	15.24	44.65	10.17	5.25	**17.90**	18.42	64.22	35.78	6.13	0.72	14.08	14.11
Janus-Pro	6.03	0.89	53.02	73.98	0.00	0.00	82.91	17.09	10.93	8.04	0.00	0.70
Janus-Pro (fine-tuned)	31.23	25.12	56.81	52.28	2.79	4.13	18.52	81.48	6.48	1.67	11.20	13.56

Table 1: Main results (%) on AMAZE, including the Maze and Queen tasks. ↓ indicates lower is better, while ↑ indicates higher is better.

2.3 CONSENSUS WITH HUMAN JUDGES

To validate the reliability of our proposed automatic metric for logical validity, we measure the agreement between it and human judges. Specifically, we randomly sample 50 images per task for each evaluated model. Three human annotators were tasked with binary classification: check whether the generated solution successfully matches the ground truth image without any violations to the naked eye. We compare the results of human judges against our PASS rate; the agreement rate is 98%, implying the high reliability of our automatic metric. Regarding the 2% discrepancy, we find that it primarily arises from two scenarios: (1) complex tasks that cause human perception errors; and (2) overly faint solutions or altered non-solution areas. Fortunately, our automatic metric for pixel-wise fidelity helps detect these failure cases.

3 EXPERIMENT

3.1 EXPERIMENTAL SETUP

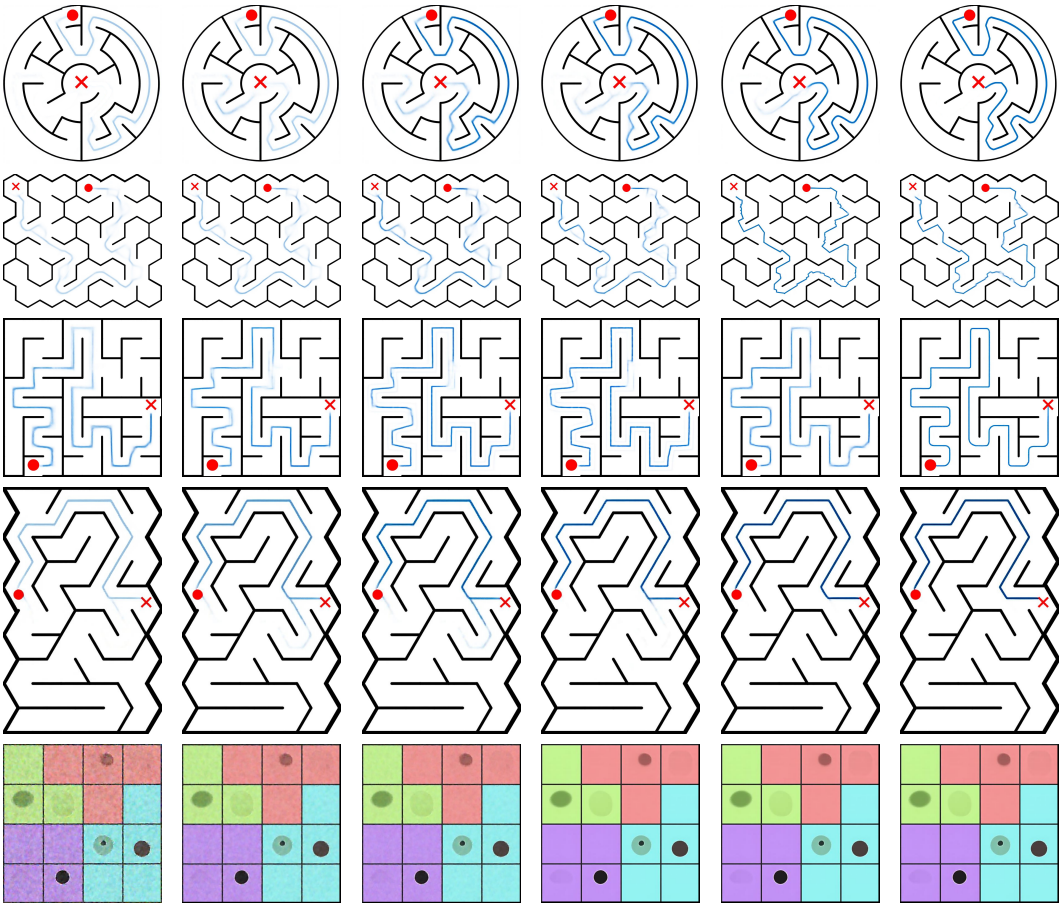
Evaluated models. To investigate the intrinsic visual planning capabilities of current image editing models, we benchmark representative models from two dominant generative paradigms: diffusion-based and autoregressive models. To reveal the gaps between the proprietary and open-source domains, we consider the following image editing models:

- **Proprietary domain** includes frontier models like GPT-Image-1 (OpenAI, 2025), NanoBanana-Pro (DeepMind, 2025)[1] and Seedream-4.5 (Seedream et al., 2025).
- **Open-source domain** includes Qwen-Image-Edit (Wu et al., 2025a), Flux-Kontext-Dev (Labs et al., 2025), Bagel (Deng et al., 2025) and Janus-Pro-7B (Chen et al., 2025). Among them, Qwen-Image-Edit, Flux-Kontext-Dev, and Bagel are diffusion-based and Janus-Pro-7B is an autoregressive model.

Evaluation method. We directly prompt models to draw out the required solution. We keep the prompt concise and clear and apply the same prompt to all models, minimizing variances arising from prompts (see the example evaluations in the Figure 2). The complete prompts are provided in Appendix A.

Measures. We evaluate each model 5 times and report the average PASS@5, MSE-IN, MSE-OUT, and COVERAGE and VIOLATION ratios (see §2.2). We also supplement with PASS@1 using the first-round image generation.

[1]For NanoBanana-Pro, the use of Chain-of-Thought (CoT) is not publicly reported.



(a) $t = 1$ (b) $t = 2$ (c) $t = 4$ (d) $t = 6$ (e) $t = 8$ (f) $t = 10$

Figure 3: Solutions from different denoising steps (t) of a fine-tuned Bagel on Maze (first four rows) and Queen (last row) task.

3.2 Main Results

Our initial results show that Bagel and Janus-Pro struggle in the zero-shot setting, i.e., they fail to follow the instruction and generate valid solutions, likely because these are out-of-domain scenarios for them (see Table 1). Thus, to investigate their potential in acquiring the visual planning ability, we apply supervised fine-tuning. We curate a training set consisting of the simplest scale: 3×3 mazes spanning all four geometry types (circle, hexagon, square, and triangle) and 4-Queens puzzles. The training set comprises 800 samples per geometry type and 800 4-Queens puzzles, accompanied by a separate held-out set for validation. We train each model for up to 8 epochs, and apply early stopping when the MSE loss on the validation set plateaus.

Frontier proprietary editing models have limited capacity in abstract visual planning. On Maze, proprietary editing models achieve the best Pass@1 of 5.4%, exhibiting limited zero-shot proficiency. They often fail to respect maze boundaries, generating paths that cut through walls. Among them, GPT-Image-1 exhibits the worst instruction-following capability, with a violation rate of 62.88%. While NanoBanana-Pro performs best in terms of pixel-wise fidelity, it tends to over-generate paths that traverse the entire maze, indicated by its high violation rate (e.g., 47.76%). Seedream-4.5 appears to respect the constraints ($< 20\%$ violation), but this is through the shortcut of under-generation, i.e., it can hardly generate a complete path. On Queen, while the best performing NanoBanana-Pro shows a high Pass@1 of 30.35%, all other proprietary models demonstrate nearly zero Pass@1 in the zero-shot setting. The surprisingly high performance of NanoBanana-Pro indicates that it may have seen similar tasks during training.

Diffusion-based models may be more effective at developing visual reasoning logic than autoregressive models. We analyze which learning paradigm is better at developing visual reasoning logic.

To do so, we compare Bagel (Deng et al., 2025) and Janus-Pro (Chen et al., 2025), two representatives of diffusion-based editing models, respectively. Without fine-tuning, both have a zero PASS@1; after fine-tuning, Bagel improves PASS@1 from 0 to 11.54% on Maze, but the performance of Janus-Pro is only 1.43%. A similar trend is observed on Queen: the fine-tuned Bagel achieves 14.57% PASS@1, while the fine-tuned Janus-Pro lags, reaching only 12.57%. Though the lack of transparency regarding training precludes a definitive conclusion, these findings suggest that diffusion-based modeling may be more effective at developing visual reasoning logic. We hypothesize that the progressive denoising in diffusion models fosters a global structural awareness that is beneficial for visual planning. Conversely, the sequential, token-based nature of autoregressive models lacks this global perspective, as generation is constrained by a local, raster-scan order.

Chain-of-Thought prompting is not always helpful. We further evaluate the models using Chain-of-Thought (CoT) prompting, but the results are mixed. For unified multimodal architectures such as Bagel and Janus-Pro, CoT provides negligible benefits in the zero-shot regime. However, it yields marginal improvements following fine-tuning, suggesting that the models must first internalize the task's underlying logic before they can effectively leverage intermediate reasoning steps.

Qualitative studies of "visual planning". We provide a qualitative study of the "planning" process of a fine-tuned Bagel on both Maze and Queen tasks (see Figure 3). On Maze, the model exhibits a clear global-planning behavior. The overall solution path emerges at early denoising steps (e.g., $t = 1, 2, 4$) with low confidence, indicated by faint trajectories, and is progressively refined over time. Incorrect subpaths are gradually corrected (e.g., $t = 8$), leading to a valid solution at later steps (e.g., $t = 10$). This coarse-to-fine trajectory construction aligns with the denoising nature of diffusion models, where the global structure is iteratively improved. On Queen, we observe a distinct planning pattern: a coarse global configuration of placements is established in the initial steps, followed by fine-grained adjustments. This contrast highlights the differences between the two paradigms. While sequential tasks like Maze are amenable to iterative and local refinements, combinatorial tasks like Queen necessitate significant global updates. Such global coordination remains a formidable challenge for current editing models.

3.3 GENERALIZABILITY

We further investigate how well editing models can generalize to unseen geometry types and scales. For this study, we use a fine-tuned Bagel as it demonstrates non-trivial visual planning capabilities. For Maze, we evaluate generalization across both geometry types and scales. The test set covers scales from 3×3 to 16×16, with 50 mazes sampled per scale for each geometry type. For Queen, we evaluate across scales from 4×4 to 10×10, with 50 samples per scale.

3.3.1 CROSS-GEOMETRY GENERALIZATION

Fine-tuning on ⬡ yields the best generalization across other geometry types. We evaluate Bagel's zero-shot generalization across geometry types (See Figure 4 (left)), revealing an asymmetric transfer pattern: training on complex geometries (e.g., hexagons) yields better performance on simpler ones than vice-versa. Notably, the hexagon-trained model generalizes best—achieving 40.14% on triangles and 30.00% on squares, outperforming in-domain baselines. We attribute this to the more variable directions in hexagonal mazes. Their action space functions as a superset that encompasses the geometric constraints of both square and triangular mazes. This suggests that the models have learned fundamental path-finding logic that transcends specific geometries.

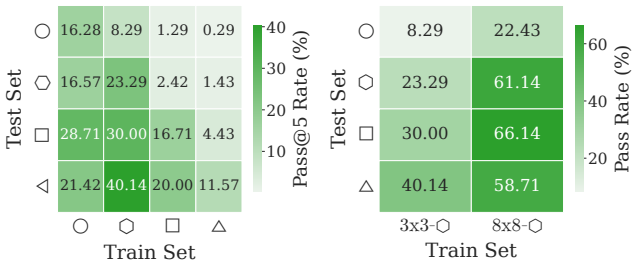


Figure 4: Zero-shot generalization. **Left:** PASS@5 matrix for 3×3 models. **Right:** Comparison between 3×3 and 8×8 ⬡ training.

Fine-tuning on larger-scale ⬡ enhances cross-geometry generalization. To further explore if increased training complexity reinforces cross-geometry generalization, we extend our study to a

8×8 training setting. As illustrated in Figure 4 (right), increasing the scale of the training mazes leads to a substantial leap in generalization performance across all test domains, which indicates that exposure to larger-scale problems forces the model to transition from learning in-domain geometric patterns to out-of-domain visual planning capabilities.

3.3.2 Cross-Scale Generalization

Fine-tuning on 3×3 mazes yields generalization to larger scales up to 16×16. We further investigate cross-scale generalization. For this analysis, we fine-tuned Bagel on ⎔ mazes since they induce the best cross-geometry generalization. Surprisingly, fine-tuning on simple 3×3 mazes enables generalization to larger scales up to 16×16 (see Figure 5). We again extend our study to the 8×8 training setting. Unsurprisingly, more complex training mazes lead to better zero-shot cross-scale generalization. However, while 8×8 trained model excels at maintaining local structural constraints, indicated by the low violation rate, it still struggles with the most complex mazes. We find that, when the scale increases, the model often generates perfect local paths near the starting and end points of the maze but fails to connect them in the middle, leading to near-zero success rate, presumably because the path length increases with the scale, making it more challenging for the model to maintain a growing long-distance dependency.

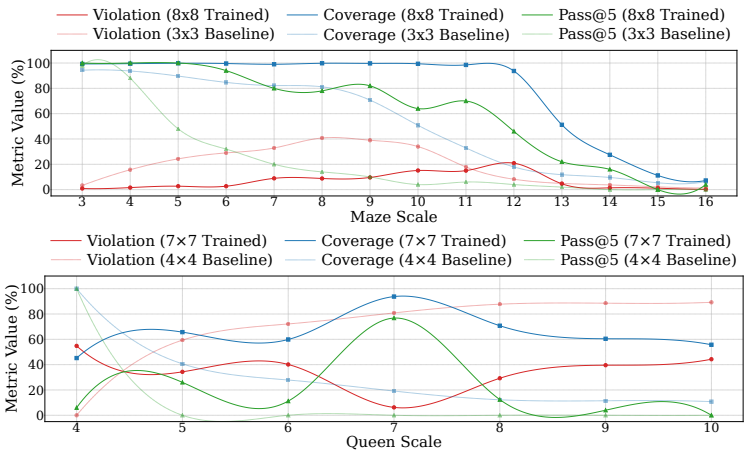


Figure 5: Generalization across scales for Maze (top) and Queen (bottom) tasks. The dotted line and the solid line respectively represent baseline and trained model.

Queen relies on more complex training scales for non-trivial cross-scale generalization. As shown at the bottom of the Figure 5, unlike Maze, which induces non-trivial cross-scale generalization when training from the smallest 3×3 scale, fine-tuning on the smallest 4×4 scale yields perfect in-domain performance but no generalization to larger scales, indicating a strong memorization. Consistent with our observations on Maze, fine-tuning on larger scales (e.g., 7×7) yields better, non-trivial cross-scale generalization. This suggests that for combinatorial visual planning, exposure to larger training scales is crucial to acquiring scale-invariant reasoning capabilities.

3.4 Scaling Effect

Next, we study if scaling up the training data and compute improves visual planning. For this analysis, we fine-tune Bagel on 8×8 ⎔ mazes (representing the best performing geometry), 8×8 ○ mazes (representing the hardest geometry), and 7-Queens, respectively, and test on all scales of the same geometry type.

Scaling up training data. We analyze the effect of data scaling with $N \in \{800, 1600, 3200, 6400\}$ under a fixed compute budget of 1000 training steps. In general, scaling up training data initially yields slight improvements on all tasks, but the gains become marginal after $N > 1600$ (see Figure 6). On Maze, data scaling results in a quick performance saturation on both ⎔ and ○ geometries, e.g., while the performance on ⎔ improves from 65.2% to 68.1% when increasing N from 800 to 1600, it then stays

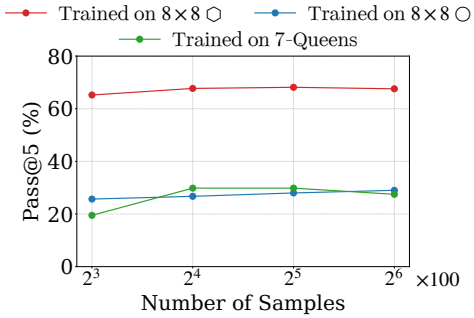


Figure 6: Data scaling.

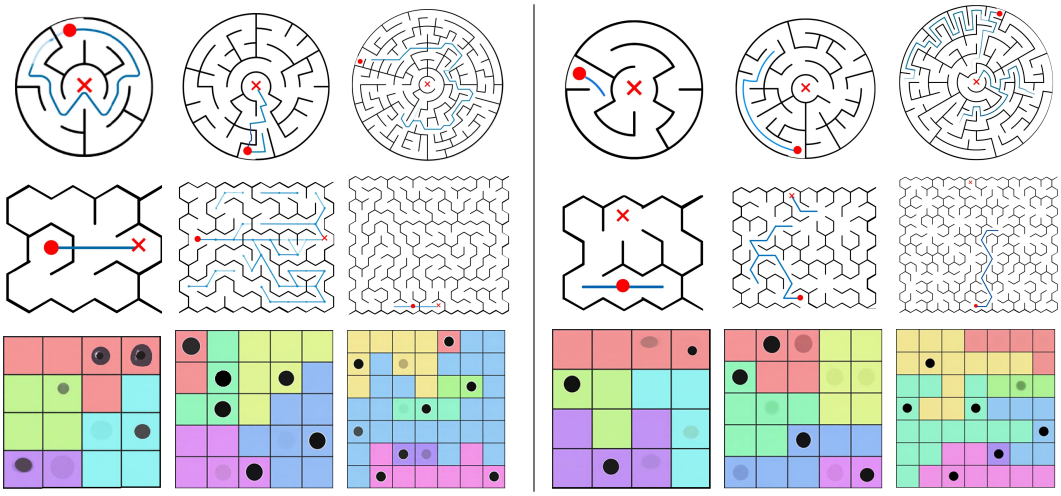


Figure 8: Examples of failure modes in Maze (first two rows) and Queen (last row). **Left:** constraint violation; **Right:** incomplete solution. Examples from other maze geometries can be found in Appendix D.

plateaued, suggesting that scaling up training data mainly improves robustness to scale variation rather than the intrinsic sequential planning ability. Though the trend on 7-Queens is similar to that on Maze, scaling up training data from 800 to 1600 yields a much larger initial gain (+10.3%), indicating that combinatorial tasks like Queen benefit a lot more from the highly diverse solution patterns. We also provide a more detailed analysis of data scaling on cross-domain geometries for Maze task in Appendix B.

Scaling up training compute. We double the training duration from 500 to 1000 steps (equivalent to increasing from 2.5 to 5 epochs) while maintaining a fixed training set of 6400 samples. Overall, scaling up training compute yields consistent improvements except for slight drops on Maze at step 800 and on Queen at step 700. Interestingly, gains are generally marginal over 500–700 steps and become more pronounced from step 700 onward. For example, the performance on ⎔ improves by 6.1% over 500–700 steps and by 15.8% over 700–1000 steps. Given the upward momentum in performance, we hypothesize that extended training will yield further gains. A more detailed analysis of the interaction between data and compute is provided in Appendix C.

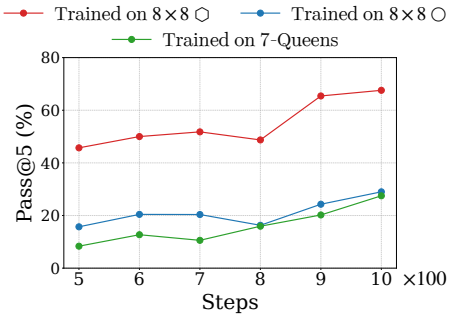


Figure 7: Compute scaling.

3.5 Error Analysis

We further analyze model failures, which can be broadly categorized into two modes: constraint violation and incomplete solution (see Figure 8). Additional Maze cases across geometries provided in Appendix D.

Constraint violation refers to instances where the generated solution fails to adhere to task-specific requirements, reflecting the model's deficit in instruction-following. On Maze, these violations manifest as invalid trajectories that cross boundaries or connect the start and end points directly–a failure mode that becomes particularly pronounced in complex geometries like ○ and ⎔. On Queen, this is characterized by erroneous placements that break the global constraint.

Incomplete solution refers to cases where the model produces only a partial solution, reflecting a conservative generation strategy. On Maze, we observe that the model often generates a valid prefix

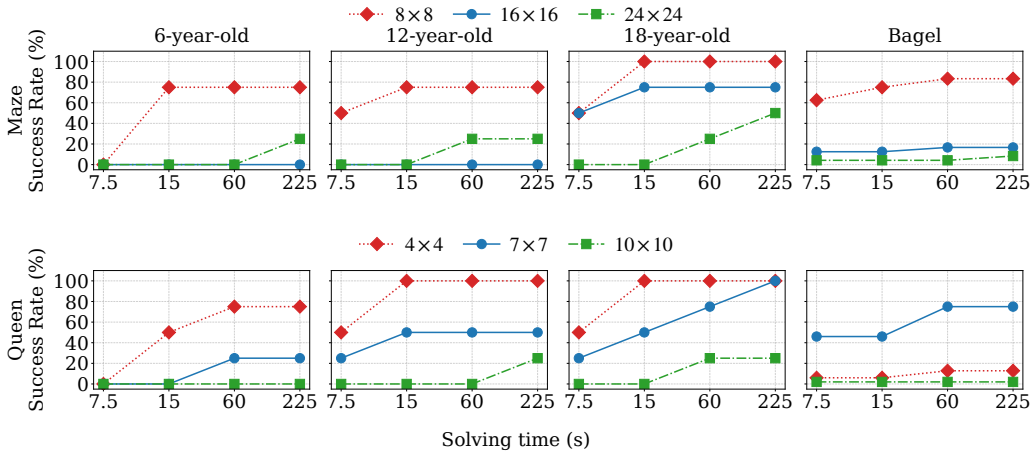


Figure 9: Success rates of humans and Bagel under different time budgets.

path from the start point but stops early before reaching the end point–a tendency that is particularly pronounced in larger scales or out-of-domain geometries. On Queen, this corresponds to instances where the model completes only a subset of goal placements. On both tasks, this failure mode results in locally valid but globally incomplete solutions.

3.6 Human Studies

We also conducted a comparative study between the model and humans.

Settings. We use Bagel as the model representative that is fine-tuned on 8×8 ⎔ mazes and 7-Queens, respectively. For human solvers, we recruited volunteers from three different age groups, representing different stages of cognitive development:

- **6-year-old** represents early childhood, where basic visual planning skills are developed but complex logical planning is still forming.
- **12-year-old** represents the transition to formal operational thoughts, where abstract reasoning and visual planning are largely consolidated.
- **18-year-old** represents the adult baseline for fully mature visual planning.

Each age group consists of four participants; each individual is assigned ⎔ mazes across three scales (8×8, 16×16, and 24×24) and Queen puzzles of scales 4, 7, and 10. These scales are selected to provide a balanced coverage of difficulty levels within each task, representing easy, moderate, and hard levels respectively. This configuration yields 12 trials per age group for each task, facilitating controlled analysis across task complexity and cognitive development stage.

We provide participants with unlimited time for mental reasoning prior to drawing their solutions. To align with the model's inference process, participants are required to complete their drawings in a single, continuous attempt—prohibiting erasing, backtracking, or restarts. We record the total latencies for both the reasoning and drawing phases. To ensure a fair comparison, the model is allocated a time budget equivalent to that of human participants, during which it may generate as many candidate solutions as the budget allows.[2]

The success rate of humans is more positively correlated with the time permitted than that of the model. Unsurprisingly, with increasing time allowed, human solvers tend to achieve a higher success rate (see Figure 9), particularly in harder tasks. In contrast, the performance of the model remains relatively flat regardless of the time allowed. Moreover, the elder group demonstrates better leverage of extra time; for example, the 18-year-old group achieves a perfect score on 7-Queens within 225 seconds, presumably because their visual planning ability has matured.

[2]A single 'drawing' takes the model about 7.5 seconds, averaged over 20 runs.

The visual planning ability of the model resembles that of the 6-year-old on Queen and that of the 18-year-old on Maze. In general, across tasks of varying difficulty levels, the trend of the model performance does not resemble the same age group (see Figure 9). To better understand their relationship, we estimate the Pearson correlation between the model and each human group on each task (see Figure 10). On Maze, we observe that the model correlates best with the 18-year-old, but on Queen, it correlates best with the 6-year-old, probably because that combinatorial planning under the global constraint is generally harder.

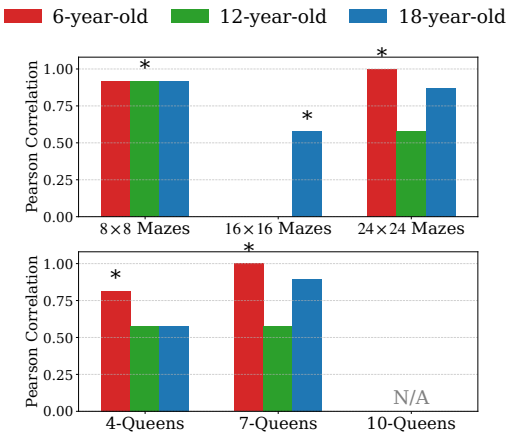


Figure 10: Correlation between model and human group. Stars mark the highest correlation per task. "N/A" indicates that the correlation is undefined because the model has zero success rates.

4 Related Work

Spatial reasoning. Spatial reasoning via visual planning requires a deep understanding of topological properties and logical rules. Existing paradigms either rely fully on textual reasoning as a proxy (Ivanitskiy et al., 2023; Dao & Vu, 2025) or integrate chain-of-thought prompting into visual reasoning (Wu et al., 2025b; Li et al., 2025; Zhang et al., 2025c). While there has been work exploring fully visual approaches, without relying on textual reasoning (Xu et al., 2025c; Zhang et al., 2025b), they only consider simple grid-based topologies and use costly step-wise image generation to model sequential planning. In contrast, we curate a set of spatial reasoning tasks of diverse visual geometries and propose an efficient editing-as-reasoning framework.

Image editing models. The goal of image editing is to transform an input image per the given instruction. Existing image editing models generally fall into two main streams: (1) autoregressive models that rely on token-based image representations for causal language-like modeling (Chen et al., 2025; Team, 2024), and (2) diffusion-based models that foster global structural awareness by simultaneously refining the entire image manifold through iterative denoising (Lipman et al., 2023; Deng et al., 2025). Early work learns a standalone editing model (Brooks et al., 2023) while recent research focuses on developing unified multimodal models capable of both image understanding and generation (Team, 2024; Chen et al., 2025; Deng et al., 2025). We formulate visual spatial reasoning as an editing task and repurpose recent strong editing models for it.

Evaluations of image editing models. Evaluations of image editing models assess whether the transformed image aligns with the given instruction. They have been through visual question-answering-based checks (Antol et al., 2015; Goyal et al., 2017), vision-language models based judges (Chen et al., 2024), and image-text alignment scoring (Watanabe et al., 2023; Kim et al., 2025), but these evaluation paradigms often prioritize semantic fidelity or consistency over logical correctness (Tong et al., 2024; Yu et al., 2025), thus inadequate for tasks that emphasize logical validity. To address the gap, we curate a set of abstract reasoning tasks devoid of perceptual complexity, accompany it with reliable and rule-based automatic metrics, and evaluate the intrinsic visual planning in image editing models.

5 Conclusion

We have proposed EAR, an editing-as-reasoning paradigm that reformulates visual planning as a single-step image-editing task. To benchmark editing models on visual planning, we develop AMAZE, a set of abstract visual planning tasks that consist of Maze and Queen, covering two complementary paradigms of visual planning. AMAZE is designed to be devoid of perceptual complexity, enabling a focused study of models' intrinsic visual planning and facilitating reliable and automatic evaluation. We empirically find that existing editing models are still limited in abstract visual planning. While supervised fine-tuning on simple tasks yields remarkable improvements, the best fine-tuned model still falls short of the instantaneous, nearly zero-shot reasoning of human solvers.

6 ACKNOWLEDGEMENTS

Yanpeng Zhao acknowledges the support of the National Natural Science Foundation of China (12574467). We would like to thank Chenghao Liu for their assistance with the experiments and helpful suggestions.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=dbFEFHAD79.

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

Alan Dao and Dinh Bach Vu. Alphamaze: Enhancing large language models' spatial intelligence via grpo, 2025. URL https://arxiv.org/abs/2502.14669.

Rob Dawson. mazes. https://github.com/codebox/mazes, 2021.

Google DeepMind. Gemini 3 pro image (nano banana pro). Web page, 2025. URL https://deepmind.google/models/gemini-image/pro/.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Michael Igorevich Ivanitskiy, Alex F. Spies, Tilman Räuker, Guillaume Corlouer, Chris Mathwin, Lucia Quirke, Can Rager, Rusheb Shah, Dan Valentine, Cecilia Diniz Behn, Katsumi Inoue, and Samy Wu Fung. Structured world representations in maze-solving transformers, 2023. URL https://arxiv.org/abs/2312.02566.

Yoonjeon Kim, Soohyun Ryu, Yeonsung Jung, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang, and Eunho Yang. Preserve or modify? context-aware evaluation for balancing preservation and modification in text-guided image editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23474–23483, 2025. doi: 10.1109/CVPR52734.2025.02186.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36340–36364. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25cz.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

OpenAI. Gpt image 1: State-of-the-art image generation model. Web page, 2025. URL https://platform.openai.com/docs/models/gpt-image-1.

Team Seedream, :, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. Seedream 4.0: Toward next-generation multimodal image generation, 2025. URL https://arxiv.org/abs/2509.20427.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. doi: 10.48550/arXiv.2405.09818. URL https://github.com/facebookresearch/chameleon.

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 (CVPR)*, pp. 9568–9578, June 2024.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. MathCoder-VL: Bridging vision and code for enhanced multimodal mathematical reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2505–2534, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.128. URL https://aclanthology.org/2025.findings-acl.128/.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.

Yuto Watanabe, Ren Togo, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Manipulation direction: Evaluating text-guided image manipulation based on similarity between changes in image and text modalities. *Sensors*, 23(22), 2023. ISSN 1424-8220. doi: 10.3390/s23229287. URL https://www.mdpi.com/1424-8220/23/22/9287.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025. URL https://arxiv.org/abs/2509.20328.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *CoRR*, abs/2303.04671, 2023. doi: 10.48550/ARXIV.2303.04671. URL https://doi.org/10.48550/arXiv.2303.04671.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a. URL https://arxiv.org/abs/2508.02324.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing Spatial Reasoning in Vision-Language Models with Interwoven Thinking and Visual Drawing. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

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

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. In *Workshop on Foundation Models Meet Embodied Agents at CVPR 2025*, 2025b. URL https://openreview.net/forum?id=ELIt3v3S1J.

Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images, 2025c. URL https://arxiv.org/abs/2505.11409.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 3081–3089, 2022.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, Zhedong Zheng, Zhipeng Zhang, Yifan Wang, Lin Song, Lijun Wang, Yanwei Li, Ying Shan, and Huchuan Lu. How far are vlms from visual spatial intelligence? a benchmark-driven perspective, 2025. URL https://arxiv.org/abs/2509.18905.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.82. URL https://aclanthology.org/2025.acl-long.82/.

Tao Zhang, Jia-Shu Pan, Ruiqi Feng, and Tailin Wu. Vfscale: Intrinsic reasoning through verifier-free test-time scalable diffusion model, 2025b. URL https://arxiv.org/abs/2502.01989.

Yu Zhang, Yunqi Li, Yifan Yang, Rui Wang, Yuqing Yang, Dai Qi, Jianmin Bao, Dongdong Chen, Chong Luo, and Lili Qiu. Reasongen-r1: Cot for autoregressive image generation models through sft and rl, 2025c. URL https://arxiv.org/abs/2505.24875.

A Complete Prompts for Maze and Queen Tasks

We provide the complete prompts used for the Maze and Queen tasks. For each task, we include the prompts with and without Chain-of-Thought (CoT) as follows:

A.1 Prompts without Chain-of-Thought

Maze task (without CoT) requires generating a valid path from the entrance to the exit while strictly following geometric constraints.

> Add the blue solution path for the maze, connect start point (solid red circle) to end point (red 'X' mark). Ensure all original maze elements (walls, points, etc.) remain unchanged—only add the path.

Queen task (without CoT) requires placing all queens such that no conflicts occur across rows, columns, and different color regions.

> Generate the solved board by placing one queen (represented by a solid black circle in the center of a grid cell) in each row, column, and colored region while ensuring queens do not touch in 8-neighborhood.

A.2 Prompts with Chain-of-Thought (CoT)

The CoT-augmented prompts explicitly encourage the model to perform intermediate reasoning before producing the final output.

Maze Task (CoT) augments the instruction with an additional prompt, as shown below:

> Add the blue solution path for the maze, connect start point (solid red circle) to end point (red 'X' mark). Ensure all original maze elements (walls, points, etc.) remain unchanged—only add the path.
> You should first think about the planning process in the mind. The planning process must be enclosed within `<think>` and `</think>` tags.

Queen Task (CoT) uses a similar process, as shown below:

> Generate the solved board by placing one queen (represented by a solid black circle in the center of a grid cell) in each row, column, and colored region while ensuring queens do not touch in 8-neighborhood.
> You should first think about the planning process in the mind. The planning process must be enclosed within `<think>` and `</think>` tags.

For models that do not natively support joint text-and-image generation (e.g., Janus-Pro), we adopt a two-stage inference prompts. Prompts for these models consists of two stages: text generation and image generation. This formulation is shared across both the Maze and Queen tasks. We illustrate the prompt using the Maze task as an example; the same formulation applies to the Queen task.

Prompt for text generation requires model to output text CoT, shown as follows:

Add the blue solution path for the maze, connect start point (solid red circle) to end point (red 'X' mark). Ensure all original maze elements (walls, points, etc.) remain unchanged—only add the path.
You should first think about the planning process in the mind. The planning process is enclosed within <think> </think> tags.

Prompt for image generation requires model to output final image only. The ellipsis denotes the model's reasoning in text generation. The prompt is shown as follows:

Add the blue solution path for the maze, connect start point (solid red circle) to end point (red 'X' mark). Ensure all original maze elements (walls, points, etc.) remain unchanged—only add the path.
<think>..... </think>
According to your thinking process, output the image only.

B Scaling Up Training Data on Cross-Domain Performance

We further investigate how scaling the training data affects cross-domain performance, where models are trained on a single geometry and evaluated across different geometries. We train models on 8×8 ⬡ mazes and 8×8 ○ mazes with fixed steps (500), and evaluate cross-domain performance on all geometry types across all scales from 3×3 to 16×16.

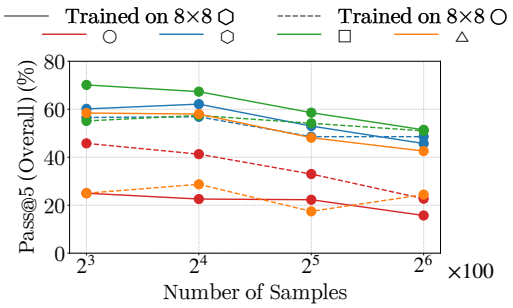


Figure 11: Data scaling on cross-domain performance.

As shown in Figure 11, the topology of the training geometry plays a critical role in determining transferability. Models trained on ⬡ mazes (solid line) exhibit robust performance across all tested shapes, whereas those trained on ○ mazes (dotted line) show weaker transferability. This is primarily because the ⬡ mazes allow the model to learn stable, translation-invariant navigation strategies, compared to ○ mazes with arbitrary action spaces. Interestingly, all models have best performance in □ mazes, which is possibly because its trade-off between action space and topological constraints.

Notably, the degradation at larger training data indicates a tendency toward geometry-specific overfitting: as the training distribution becomes denser, the model increasingly specializes to the source geometry, reducing its ability to generalize to structurally different domains.

C Extended Analysis of Data–Compute Scaling

We further analyze how data scaling interacts with compute budgets. The training and evaluation settings are the same as in §3.4. As shown in Figure 12, the effect of increasing training data is highly dependent on the available compute, exhibiting a clear coupled behavior.

For Maze task (left), performance consistently improves with more training steps, while the benefit of increasing data is conditional: moderate scaling ($N \leq 3200$) helps, but larger datasets often yield diminishing performance. For Queen task (right), the dependence on compute is more pronounced. Higher-step models benefit more consistently from larger datasets, whereas low-step models exhibit unstable and inconsistent scaling trends.

These results reveal a strong coupling between data and compute. Effective scaling requires a balanced regime where both data and optimization steps are sufficiently large. This suggests that the

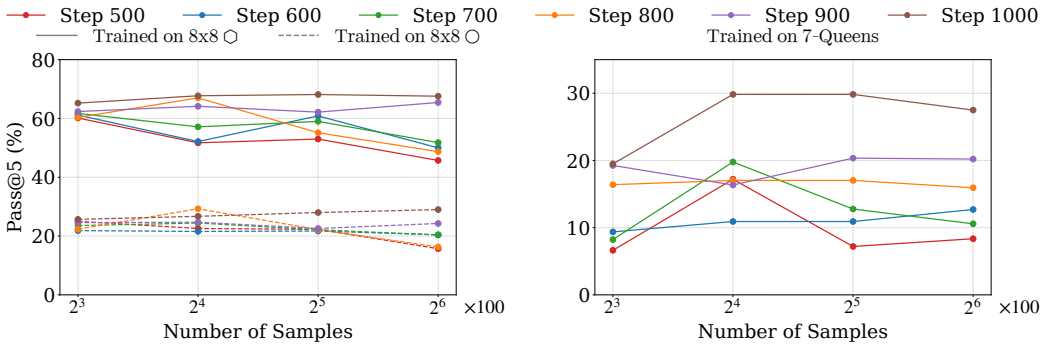


Figure 12: Joint Scaling of Data and Compute.

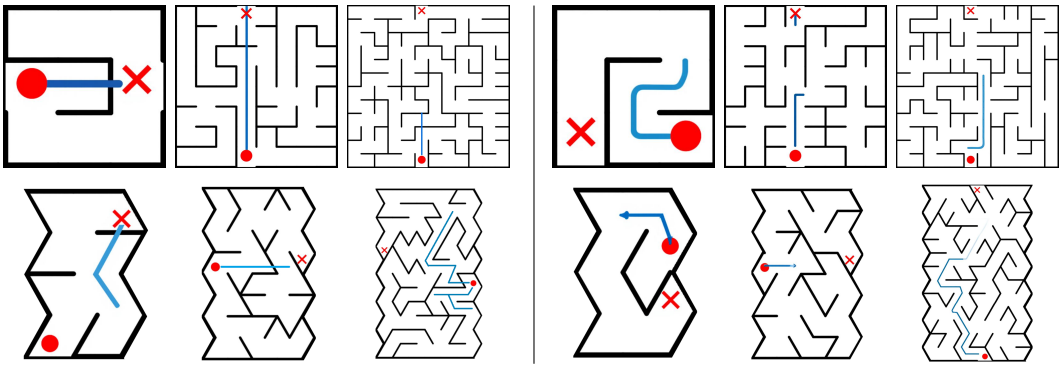


Figure 13: Fatal cases for □ and △ mazes. **Left:** boundary violation; **Right:** incomplete paths.

bottleneck of visual planning is jointly constrained by optimization capacity and the ability to fully absorb the training distribution.

D Additional Error Cases for Maze Task

We provide an additional set of examples across different geometry types, including □, and △ mazes. Constraint violations are more frequent when the action space is different from the training distribution (out-of-domain geometries), while incomplete solutions are more prevalent in larger-scale instances, where long-range dependencies are required to connect distant regions. These results further support that the observed failure modes reflect a general limitation in maintaining both local validity and global consistency during visual planning.