BUILD THE SCAFFOLD, THEN GENERATE: EXPLICIT SPATIAL PLANNING FOR SPATIALLY FAITHFUL IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image models produce high-fidelity images but remain unreliable in realizing complex spatial compositions. This limitation is more pronounced in open-source models, with a notable gap to proprietary models. We propose SCAFFOLDGEN, an agent-driven framework that separates spatial planning from visual synthesis. Given a prompt, the planning agent parses constraints from the prompt and constructs a global layout plan in HTML, which is rendered into a visual scaffold. The generator is guided by the visual scaffold to synthesize detailed visual appearance. For systematic evaluation, we introduce SLAB, a Spatial LAyout Benchmark covering ten task categories involving complex spatial composition and layout organization. The benchmark assesses semantic and spatial constraint satisfaction through dependency-aware question-answer accuracy (QAA). On SLAB, SCAFFOLDGEN outperforms all evaluated open-source methods and achieves competitive performance with certain proprietary models. Further evaluation on SpatialGenEval also shows consistent improvements over the baseline, confirming that the enhanced spatial awareness generalizes beyond SLAB. Together, the results demonstrate the effectiveness of SCAFFOLDGEN and point to a broader research direction: leveraging the spatial reasoning capabilities of LLM agents to compensate for the limited spatial perception of generation models.

1 INTRODUCTION

Modern text-to-image models have demonstrated remarkable proficiency in general instruction following and high-fidelity visual synthesis Black Forest Labs (2024; 2025a); OpenAI (2025a; 2026b); Google (2025a); HiDream-ai (2026); Wu et al. (2025); Tongyi-MAI (2025). However, generation remains challenging for tasks demanding intricate *spatial composition*, which entail the joint satisfaction of a global layout specification and fine-grained semantic and spatial constraints. Figure 1 presents representative examples of such tasks, showing that advanced image generators struggle to satisfy all specified requirements. This challenge is particularly pronounced for open-source models. According to our evaluation (Table 1), the leading proprietary model GPT-Image-2 OpenAI (2026b) achieves a question-answer accuracy above 90%, whereas the evaluated open-source models Black Forest Labs (2024; 2026; 2025a); HiDream-ai (2026); Tongyi-MAI (2025); Wu et al. (2025) trail behind by more than 20 percentage points. These results reveal a substantial disparity in spatial perception between open-source and proprietary generators.

This diagnosis motivates us to harness large language models (LLMs) to compensate for the image generator's deficiency in spatial perception. We propose SCAFFOLDGEN, an agent-driven framework that decouples spatial planning from visual synthesis. Specifically, An LLM agent first parses the compositional requirements from the prompt and translates them into a spatial plan. This spatial plan specifies the global arrangement of the image with regions and semantic labels. We adopt HTML as an intermediate representation that bridges language and vision. The resulting spatial plan is rendered into a *visual scaffold* that guides the subsequent generation. Much like the scaffolding erected around a building under construction, this scaffold outlines the compositional skeleton without dictating the final appearance. The generator, in turn, derives structural cues from the scaffold and relies on the textual prompt to render detailed visual appearance.

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.

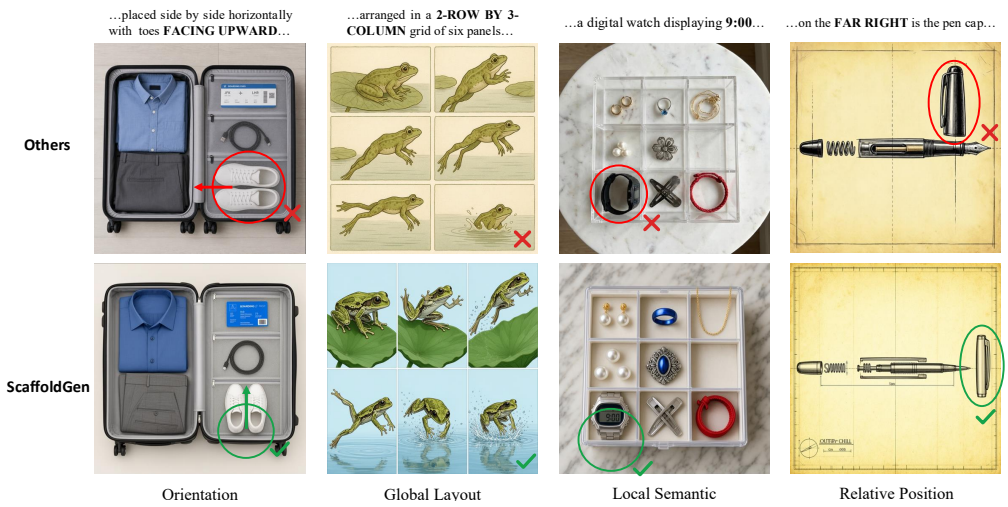


Figure 1: Examples illustrating the difficulty of satisfying complex spatial constraints for image generators. The prompt requirements are shown at the top. The first row presents failure cases from GPT-Image-2 OpenAI (2026b), GPT-Image OpenAI (2025a), Nano-Banana-Pro Google (2025a), and Qwen-Image Wu et al. (2025), from left to right, illustrating distinct error types. The second row shows corresponding results from SCAFFOLDGEN that satisfy the stated requirements.

To systematically assess performance in spatially demanding scenarios, we introduce SLAB (**S**patial **LA**yout **B**enchmark), a dedicated benchmark spanning ten task categories that cover geometric arrangements, region organization, and cross-region correspondence. As illustrated in Figure 3, each task requires models to fulfill detailed semantic and spatial requirements within a prescribed global composition. For evaluation, we decompose each prompt into a set of binary atomic questions, each verifying a single requirement. Performance is measured by dependency-aware question-answer accuracy (QAA). Local questions are linked to their necessary binding prerequisites and receive credit only when these prerequisites are also satisfied.

On SLAB, SCAFFOLDGEN substantially improves the QAA score from 60.90% for the baseline to 80.80%. It outperforms all evaluated open-source models and achieves competitive performance with certain proprietary systems. We further evaluate on SpatialGenEval, a related benchmark focusing on object-level spatial relations. The results likewise show an improvement over the baseline, with overall accuracy increasing from 61.9% to 63.8%. Together, these results support the effectiveness of SCAFFOLDGEN in enhancing spatial perception of image generator and verfity the generalization to related tasks beyond SLAB.

Our contributions are summarized as follows:

- We propose SCAFFOLDGEN, an agent-driven framework that decouples spatial planning from visual synthesis: an LLM agent converts spatial requirements into an spatial plan, which is rendered into a visual scaffold that supplies structural guidance to the generator.
- We introduce SLAB (**S**patial **LA**yout **B**enchmark), a benchmark comprising ten categories of spatially demanding tasks paired with a dependency-aware evaluation protocol.
- Experiments on SLAB and SpatialGenEval demonstrate clear improvements over the baseline, supporting the effectiveness of SCAFFOLDGEN for spatially faithful image generation.

2 RELATED WORK

2.1 BENCHMARKS FOR COMPOSITIONAL AND SPATIAL IMAGE GENERATION

Spatial relations are commonly included as one evaluation dimension in general text-to-image benchmarks, such as GenEval series (Ghosh et al., 2023; Kamath et al., 2025), T2I-CompBench series (Huang et al., 2023; 2025), ConceptMix (Wu et al., 2024), DPG-Bench (Hu et al., 2024),

OneIG-Bench (Chang et al., 2025), TIIF-Bench (Wei et al., 2025), and T2I-CoReBench (Li et al., 2026). Within these evaluations, spatial assessment tends to emphasize basic object-level relations, with more limited coverage of complex scene organization. A complementary line of benchmarks places spatial and structural fidelity at the center of evaluation. GenSpace (Wang et al., 2025b) and SpatialGenEval (Wang et al., 2026) examine relative position, orientation, and distance in greater depth, while StructBench (Zhuo et al., 2026) addresses factual and structural correctness in structured image generation and editing. Along this line, SLAB examines *global composition*, where the satisfaction of local requirements depends on the prescribed arrangement of regions, hierarchies, and cross-region correspondences.

2.2 Spatially Controllable Image Generation

ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) introduce a direct way to guide pretrained diffusion models with visual conditions. Sketches and depth maps, for example, provide geometric cues that guide the spatial organization of images. Building on this principle of spatial conditioning, a line of work on layout-to-image generation explores explicit control over scene structure through prescribed layouts. GLIGEN (Li et al., 2023), BoxDiff (Xie et al., 2023), and InstanceDiffusion (Wang et al., 2024) condition synthesis on object categories, bounding boxes, and other instance-level signals. EchoGen (Zou et al., 2026) couples layout-conditioned synthesis with image grounding through cycle-consistent learning, while OcclusionFormer (Li & Ding, 2026) models occlusion order explicitly. These methods primarily focus on placing objects at prescribed locations in the image, with limited emphasis on autonomously optimizing the layout itself.

A complementary line of work integrates the layout generation into the method. LayoutGPT (Feng et al., 2023) and CreatiLayout (Zhang et al., 2025) employ language-guided layout construction, SKE-Layout (Wang et al., 2025a) incorporates spatial knowledge, and LayoutLLM-T2I (Qu et al., 2023) and LLM-grounded Diffusion (Lian et al., 2023) connect text-derived object layouts to conditional synthesis. Like ScaffoldGen, these approaches first plan a layout and then use it to guide image generation. Their layouts, however, primarily specify the positions and relations of individual object instances. ScaffoldGen addresses more complex compositions, where local requirements must be jointly satisfied under global constraints on region organization and cross-region correspondence.

2.3 LLM-Guided Image Generation

Beyond instance-level layouts, LLM-guided generation also explores more expressive plans that combine semantic content with spatial organization. LLM Blueprint (Gani et al., 2023) constructs scene blueprints, while RPG (Yang et al., 2024) and Ranni (Feng et al., 2024) organize generation through regional plans and semantic panels, respectively. LayoutAgent (Fan et al., 2025) and Plan2Gen (Zhou et al., 2025) use agents to plan spatial layouts and interpret complex generation instructions. PlanGen (He et al., 2025) jointly models layout planning and image generation within an autoregressive model, and 3D Space as a Scratchpad (Saha et al., 2026) introduces an editable geometric workspace. ScaffoldGen shares the use of LLMs as planning agents, but emphasizes adapting this paradigm to spatial composition tasks. Accordingly, we adopt HTML as a bridge between prompt interpretation and image synthesis: the agent decomposes semantic and spatial constraints into a structured plan, and provides explicit visual guidance to the generator.

3 Method

3.1 Plan-and-Generate Framework

As illustrated in Figure 2(a), ScaffoldGen employs a two-stage generation framework. In the first stage, a planning agent extracts the global composition and spatial constraints from the prompt and formulates a coarse spatial plan in HTML. HTML is a text-based markup representation suited to language model generation that can encode layout structure and be rendered into a visual condition. The spatial plan is then rendered into a visual scaffold that provides explicit structural guidance for the generator to synthesize visual details. To construct this interface between planning and generation, we instruct the planning agent to represent the image with only simple geometric shapes and concise semantic labels. The shapes encode region boundaries and spatial relationships, while the labels specify the intended semantic content of each region.

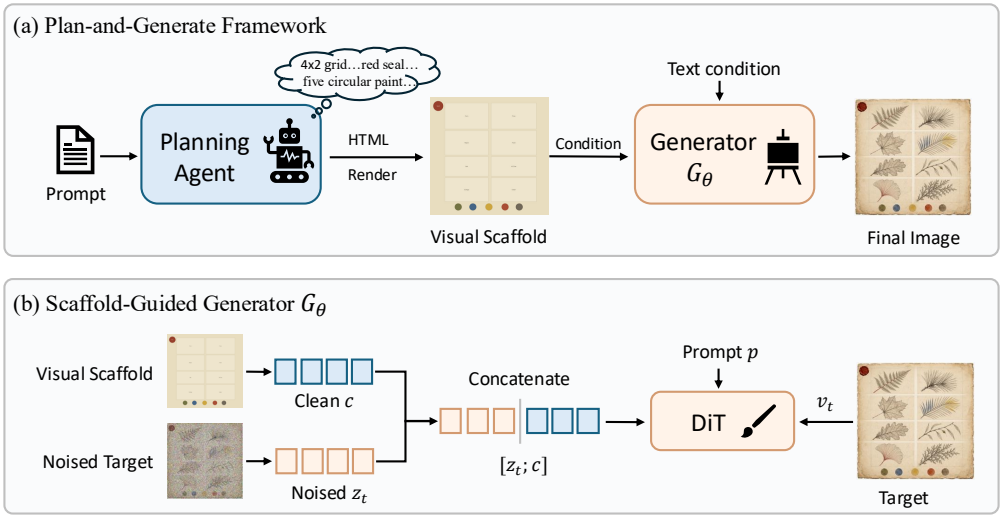


Figure 2: Overview of SCAFFOLDGEN. (a) The plan-and-generate framework, with a rendered visual scaffold linking spatial planning to image synthesis. (b) Scaffold-guided generation pipeline: the visual scaffold is encoded into clean tokens, and then concatenated with noisy target tokens.

Building on this design, we implement two variants of the planning agent. For the *proprietary agent*, we directly adopt GPT-5.6-Sol OpenAI (2026a) and design a dedicated instruction to prompt the agent to construct spatial plans in the required format. Beyond the proprietary agent, we also explore a lightweight open-source alternative, Qwen3.8-27B Qwen Team (2026) as a *fine-tuned agent*. Given the capability gap between open-source models of this scale and their proprietary counterparts, zero-shot prompting alone proves insufficient for producing reliable spatial plans. We therefore perform task-specific supervised fine-tuning (SFT) on a small set of prompt-HTML pairs to adapt the agent to our planning task.

3.2 SCAFFOLD-GUIDED GENERATOR

As shown in Figure 2(b), we encode the visual scaffold with a VAE encoder and concatenate the clean latent features c with the noisy target tokens z_t. The diffusion transformer processes this joint sequence conditioned on the text prompt p. We adapt the generator to visual scaffold conditioning with a flow-matching objective:

$$\mathcal{L}(\theta) = \mathbb{E}[\mathrm{MSE}(\hat{v}_\theta(z_t, t; p, c), v_t)]\,, \tag{1}$$

where $\hat{v}_\theta$ predicts the velocity of the target tokens, v_t is the flow-matching target velocity, t is the time variable, p is the text prompt Noise and supervision are applied only to the target segment; the visaul scaffold remains a clean condition.

We adopt FLUX.2-klein-base-9B (Black Forest Labs, 2026) as the baseline generator. Our formulation shares the input interface of image editing, but differs in the role of the reference image. In image editing, the reference image anchors the appearance of the output, and the prompt specifies the desired modifications. In our setting, the visual scaffold conveys only the global structure without detailed semantic information. The prompt instead serves as the primary source of visual details. In other words, an editing model preserves the appearance of its reference, whereas our generator inherits only the structure from the reference and synthesizes the appearance from the prompt.

3.3 TRAINING DATA CURATION

To adapt the generator to scaffold-guided generation, we construct SCAFFOLDGEN-10K, a dataset tailored for spatially controllable image generation. The construction pipeline begins with prompt synthesis on ten subtasks. The prompts are then used to generate corresponding images with GPT-Image-2 (OpenAI, 2026b). Given each prompt-image pair, we prompt GPT-5.6-Sol (OpenAI, 2026a) to produce a spatial plan, jointly considering the semantic requirements from the prompt and the

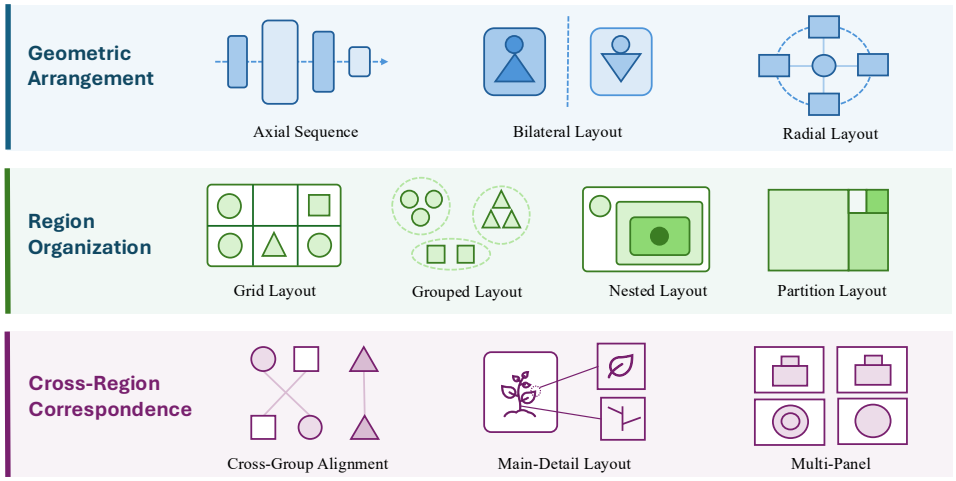


Figure 3: Task taxonomy of SLAB. Ten layout categories are organized into three complementary dimensions: geometric arrangement, region organization, and cross-region correspondence. The schematic examples illustrate the spatial organization characteristic of each category.

spatial structure from the visual scaffold. After post-filtering, we finally curate 10,000 valid triples (l, x_0, p), each comprising a visual scaffold, a target image, and a prompt. The training prompts share the task categories of SLAB but do not overlap with its prompts. The agent undergoes SFT with LoRA Hu et al. (2021) on prompt-HTML pairs.

4 SLAB: A Diagnostic Benchmark for Spatial Layout

4.1 Task Overview

We introduce SLAB to evaluate the spatial composition capabilities of image generation models. Moving beyond pairwise object positioning, the benchmark spans diverse compositional scenarios to examine whether models can faithfully realize detailed semantic requirements under complex spatial constraints. We propose 10 task categories along three complementary dimensions, illustrated in Figure 3. *Geometric arrangement* concerns the organization of elements relative to shared axes, symmetry, or a common center. *Region organization* captures the grouping, partitioning, and hierarchical containment of content. *Cross-region correspondence* addresses the alignment and consistency of content across distinct groups, scales, and views. Together, these dimensions connect local semantic fidelity with the structural coherence of the overall composition.

The prompts in SLAB are constructed through a semi-automatic synthesis pipeline. Starting from ten manually defined task categories, we generate scenarios for each category and specify their overall spatial layouts and subjects as seeds. Each seed is then automatically expanded into a detailed prompt, with explicit instructions to describe spatial relations concretely and unambiguously. The generated prompts undergo manual review, with those that deviate from the intended task design returned for revision. After this whole process, we curate a set of 200 prompts evenly covering the ten categories.

4.2 Question Design

Figure 4 illustrates four cases demonstrating prompts and representative questions. We decompose each prompt into a set of binary questions of three types: *layout* questions verify the global structure, *spatial* questions verify local spatial relations, and *content* questions verify visual attributes. Under a fixed number of questions per prompt, Layout and spatial constraints are verified by a dedicated atomic question, whereas compatible visual attributes are merged into a single check. In this way, we allocate the question budget toward spatial structure while still covering semantic fidelity.



Figure 4: Evaluation examples from SLAB. Each case shows three selected questions with their judgments. The highlighted errors cover position, count, orientation, and relative relations.

4.3 Dependency-aware Scoring

Beyond the correctness, our scoring also accounts for the dependencies among questions. For example, given *a red cup in the upper-left corner*, the color requirement is meaningful only if the cup appears in the correct position. We therefore organize the questions of each prompt p into a directed acyclic graph $\mathcal{G}_p = (\mathcal{V}_p, \mathcal{E}_p)$. Each node denotes a question and each edge (j, i) indicates the prerequisite between question j and i. Layout questions serve as root nodes, while spatial questions bind objects to regions and act as prerequisites for subsequent checks. Traversing $\mathcal{G}_p$ in topological order, we compute the effective judgment of each question and average it over all questions:

$$\tilde{a}_{p,i} = a_{p,i} \prod_{j \in D_{p,i}} \tilde{a}_{p,j}, \qquad \text{QAA} = \frac{1}{M} \sum_{p,i} \tilde{a}_{p,i}, \tag{2}$$

where $a_{p,i} \in \{0, 1\}$ is the raw judgment for question i, $D_{p,i} = \{j \mid (j, i) \in \mathcal{E}_p\}$ denotes its direct prerequisites, and M is the total number of questions. A question thus receives credit only if it and all its ancestors are judged correct, so errors propagate downstream but never upstream.

4.4 Automatic Evaluation

We use Gemini-3.1-pro Team et al. (2023) as the automatic evaluator to evaluate a group of questions based on the original prompt and the generated image. Through prompt engineering, we instruct the evaluator to independently check the each question, without relying on other questions or answers. To reduce judgment ambiguity and encourage precise assessment, we restrict each answer to `true`/`false` and require a concrete explanation for negative judgments.

To validate the reliability of the automatic evaluator, we construct a validation subset of 20 images generated by GPT-Image OpenAI (2025a), and manually annotate all 291 associated questions. The evaluator's judgments agree with human annotations on 90.03% of these questions. Considering a certain degree of ambiguity of human annotations, this level of agreement suggests that the evaluator serves as a reliable proxy for human assessment.

5 Experiments

5.1 Evaluation on SLAB

We test open-source methods with their officially recommended configuration at 1024×1024 resolution on SLAB. For ScaffoldGen, we evaluate three variants. The default ScaffoldGen

Table 1: Comparison with other methods on SLAB. We report dependency-aware QAA (%) and Δ_{dep}, the drop from raw accuracy to dependency-aware QAA score. For limited space, we only report scores on five representative subtasks; Overall aggregates questions from all ten categories.

Model	Grid Layout	Grouped Layout	Multi-Panel	Radial Layout	Cross-Group Alignment	Overall	
						QAA (%) ↑	Δ_{dep}
Closed-source generators							
Nano-Banana-Pro (Google, 2025a)	80.37	73.56	70.03	69.34	74.63	79.22	7.73
GPT-Image-1.5 (OpenAI, 2025b)	88.33	85.92	67.15	69.34	64.01	75.22	9.52
GPT-Image-2 (OpenAI, 2026b)	94.96	97.41	89.34	90.51	89.38	91.51	1.40
Seedream-5-Pro (Seed Team, 2026)	89.92	89.66	85.59	88.69	76.70	87.05	2.97
Open-source generators							
FLUX.1-dev (Black Forest Labs, 2024)	26.53	29.31	30.26	15.33	33.33	30.36	20.08
FLUX.2-klein (Black Forest Labs, 2026)	69.76	55.75	67.15	52.19	45.13	60.90	12.89
FLUX.2-dev (Black Forest Labs, 2025a)	78.51	81.32	76.66	58.03	58.70	71.06	9.55
Hidream-O1-Image (HiDream-ai, 2026)	81.17	66.95	64.55	40.88	36.58	56.54	16.50
Z-Image (Tongyi-MAI, 2025)	72.15	76.15	64.84	71.90	59.59	69.00	9.92
Qwen-Image (Wu et al., 2025)	76.92	74.14	71.18	60.95	62.83	69.37	10.89
Layout-to-Image Methods							
PlanGen (He et al., 2025)	7.16	4.89	8.93	4.74	5.01	6.58	0.00
Ranni (Feng et al., 2024)	3.98	1.44	0.86	6.57	2.95	2.94	0.00
RPG (Yang et al., 2024)	7.69	4.31	3.75	5.11	4.72	6.98	0.00
SCAFFOLDGEN	89.39	87.07	72.91	79.20	74.63	80.80 (↑19.90)	2.79
SCAFFOLDGEN-PFT	83.55	87.64	66.57	71.17	68.14	76.62 (↑15.72)	5.70
SCAFFOLDGEN-Flow	85.41	90.80	87.03	77.74	82.60	85.17 (↑5.95)	2.88

combines the proprietary agent with the fine-tuned FLUX.2-klein generator. SCAFFOLDGEN-PFT replaces the planning agent with the fine-tuned agent built on Qwen3.8-27B Qwen Team (2026) while retaining the same generator. SCAFFOLDGEN-Flow invokes the proprietary planning agent and Nano-Banana-Pro Google (2025a) for generation. Results for other tasks are provided in Appendix B.

5.1.1 OVERALL PERFORMANCE

As shown in Table 1, SCAFFOLDGEN outperforms all evaluated open-source models in overall QAA on SLAB and even surpasses closed-source models such as Nano-Banana-Pro and GPT-Image-1.5. It also matches or exceeds these two closed-source models on all five displayed subtasks. Compared with its baseline, SCAFFOLDGEN improves overall QAA by 19.90 percentage points, from 60.90% to 80.80%, with gains across all ten task categories. These results demonstrate the effectiveness of visual scaffold guidance across diverse spatial tasks.

As an auxiliary diagnostic, we report Δ_{dep}, the gap between raw accuracy and QAA. It measures the fraction of questions that are judged correct in isolation but invalidated by failed spatial prerequisites. Compared with its baseline, SCAFFOLDGEN not only improves QAA but also reduces Δ_{dep} from 12.89 to 2.79 percentage points, indicating that the gains stem from more reliable spatial structure rather than from improved local semantic fidelity alone.

5.1.2 EFFECTIVENESS ACROSS COMPONENT CHOICES

In Table 1, we assess the robustness of SCAFFOLDGEN to the choice of planning agent and generator. All three configurations outperform their respective baselines. These consistent gains indicate that the benefit of SCAFFOLDGEN is not tied to any particular planning agent or generator. Even for generators with strong spatial capability, the visual scaffold still provides structural cues that are difficult to infer from textual prompts alone, thereby strengthening their spatial composition capability.

Table 2: Comparison with other methods on SpatialGenEval. Due to space constraints, we show five representative task categories alongside the overall score. Bold marks the best reported value.

Model	Object	Position	Orientation	Comparison	Proximity	Overall
Closed-source methods						
DALL-E-3 (OpenAI, 2023)	51.1	41.5	52.9	28.4	62.4	54.8
GPT-Image (OpenAI, 2025a)	56.3	53.3	58.9	31.4	66.8	60.5
Nano-Banana (Google, 2025b)	58.5	55.5	58.9	31.8	68.7	61.7
Seed Dream 4.0 (Seedream et al., 2025)	59.9	57.2	58.9	**32.1**	68.1	62.7
Open-source methods						
SD-3.5-L (stability.ai, 2024)	52.4	44.7	52.0	25.4	61.3	54.0
FLUX.1-dev (Black Forest Labs, 2024)	51.7	50.0	55.5	28.2	62.9	56.5
FLUX.1-krea (Black Forest Labs, 2025b)	58.0	50.7	55.7	28.3	66.7	58.5
FLUX.2-klein (Black Forest Labs, 2026)	74.7	56.0	57.1	29.2	67.6	61.9
Qwen-Image (Wu et al., 2025)	61.0	55.6	56.7	28.6	67.7	60.6
SCAFFOLDGEN	**78.3**	**57.9**	**59.2**	31.5	**69.2**	**63.8**

5.1.3 COMPARISON WITH LAYOUT-TO-IMAGE METHODS

We further compare SCAFFOLDGEN with three representative layout-to-image methods (He et al., 2025; Feng et al., 2024; Yang et al., 2024) in Table 1. These methods provide an object-centric guidance with a collection of independently anchored instances. Visual scaffolds instead lift control to the level of the composition, where local content is organized by the global structure rather than assembled piece by piece. Accordingly, layout-to-image methods remain notably behind SCAFFOLDGEN across all task categories, suggesting that instance-level layout control is ill-suited to tasks with complex spatial composition requirements.

5.2 PERFORMANCE ON SPATIALGENEVAL

We further evaluate on SpatialGenEval (Wang et al., 2026) as a complementary test. Results are reported with Qwen2.5-VL-Instruct-72B Qwen Team (2025) as the evaluator. SpatialGenEval shares closely related task settings with SLAB. It centers on multi-object scenes with interacting relative-position constraints, probing the spatial fidelity of generated images. Table 2 reports representative categories and overall accuracy, with the remaining results in Appendix B. SCAFFOLDGEN achieves the highest overall accuracy among the reported models and improves over its baseline across all displayed categories. These results confirm that SCAFFOLDGEN effectively strengthens the spatial composition capability of the generator and benefits are not confined to SLAB.

Table 3: Effect of HTML integration

HTML Appended	Visual Scaffold	Generator SFT	QAA (%) ↑
–	–	–	60.90
✓	–	–	57.45
–	✓	–	70.67
–	✓	✓	**80.80**

Table 4: Effect of SFT data conditioning.

Prompt Alone	HTML Appended	Visual Scaffold	QAA (%) ↑
–	–	–	60.90
✓	–	–	70.22
–	✓	–	69.12
–	–	✓	**80.80**

5.3 ABLATION STUDY

We first compare different ways of incorporating HTML plans into image generator. Table 3 considers three strategies: the original prompt alone, the prompt with HTML code appended to the prompt as text, and the prompt paired with the rendered visual scaffold as explicit spatial guidance. Without generator SFT, appending HTML lowers QAA from the prompt-only baseline, suggesting that the generator struggles to effectively understand the spatial information encoded in HTML directly. Rendering the plan into a visual scaffold instead raises QAA score, demonstrating the benefit of providing this information as visual guidance.

To further explore whether the improvement stems from additional task-specific training, we conduct an ablation experiment on the same dataset. Table 4 compares three schemes: the original prompt alone, HTML code appended to the prompt, and prompt with the rendered visual scaffold. Prompt-only SFT raises QAA from the baseline, confirming that adaptation to the task data contributes to the improvement. However, SFT with appended HTML remains worse than prompt-only SFT, suggesting that the generator still fails to effectively use the textual plan after adaptation. Scaffold-conditioned SFT outperforms both alternatives. Together, these results show that both task-specific generator adaptation and explicit visual scaffold guidance contribute to the improvement of SCAFFOLDGEN.

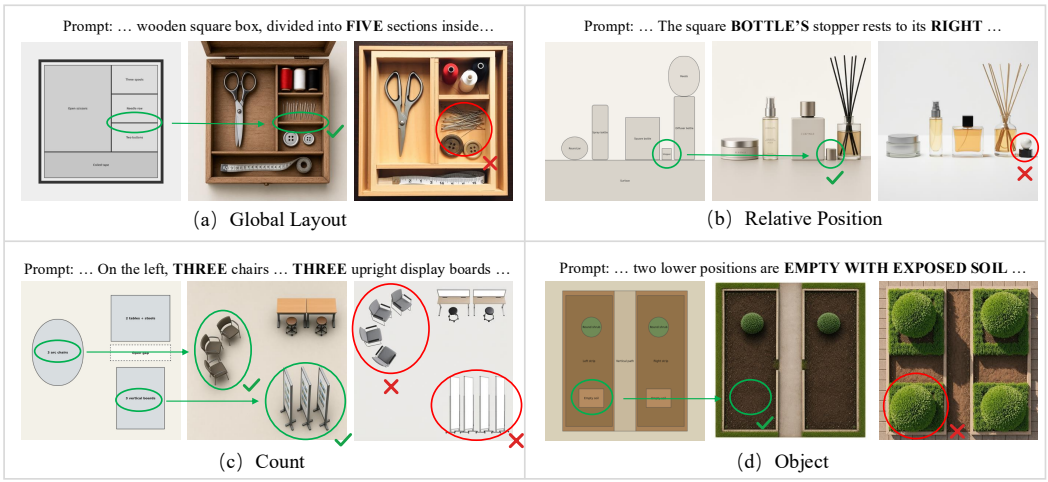


Figure 5: Qualitative examples of visual scaffold guidance. Each case shows a prompt excerpt above the visual scaffold (left), the output of SCAFFOLDGEN (middle), and the baseline output (right). The scaffold-guided outputs correct the highlighted baseline errors in (a) global layout, (b) relative position, (c) object count, and (d) object omission.

5.4 QUALITATIVE EVALUATION

Figure 5 illustrates visual scaffolds as a compact spatial abstraction. Simple geometric primitives delineate region boundaries and positions, while concise labels associate each region with its semantic content. The comparisons between SCAFFOLDGEN and the baseline show how this explicit spatial prior guides the generator toward the intended composition, helping it avoid the structural errors.

6 CONCLUSION

We observe a substantial gap between open-source and proprietary image generators in satisfying complex local semantic and spatial requirements within a prescribed global layout. This diagnosis motivates SCAFFOLDGEN, which harnesses the spatial reasoning capabilities of LLM agents to compensate for limited spatial perception of image generators through a division of labor between spatial planning and visual synthesis. To systematically evaluate performance in these scenarios, we introduce SLAB, a benchmark covering ten categories of spatially demanding tasks with dependency-aware assessment of semantic and spatial constraints. Improvements on SLAB and SpatialGenEval demonstrate the effectiveness of SCAFFOLDGEN across related spatial tasks.

Beyond providing a method for spatially faithful image generation, our study points toward a broader research direction in which advanced LLMs serve as reasoning and planning partners for visual generators, integrating their complementary capabilities into a coordinated generation process. Future work can extend this approach beyond spatial composition, exploring how increasingly capable LLMs can guide visual generation under richer and more diverse requirements.

REFERENCES

Black Forest Labs. Announcing Black Forest Labs. https://bfl.ai/blog/24-08-01-bfl, 2024.

Black Forest Labs. Flux.2 - next generation image generation. https://bfl.ai/models/flux-2, 2025a.

Black Forest Labs. FLUX.1 Krea [dev]: An 'opinionated' text-to-image model. https://bfl.ai/blog/flux-1-krea-dev, 2025b.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, 2026.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2025.

Zezhong Fan, Xiaohan Li, Luyi Ma, Kai Zhao, Liang Peng, Topojoy Biswas, Evren Korpeoglu, Kaushiki Nag, and Kannan Achan. Layoutagent: A vision-language agent guided compositional diffusion for spatial layout planning. *arXiv preprint arXiv:2509.22720*, 2025.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. In *Advances in neural information processing systems*, volume 36, pp. 18225–18250, 2023.

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4744–4753. IEEE, 2024.

Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. *arXiv preprint arXiv:2310.10640*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Google. Introducing nano banana pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025a.

Google. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025b.

Runze He, Bo Cheng, Yuhang Ma, Qingxiang Jia, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Liebucha Wu, Dawei Leng, and Yuhui Yin. Plangen: Towards unified layout planning and image generation in auto-regressive vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18143–18154, 2025.

HiDream-ai. HiDream-O1-Image. https://github.com/HiDream-ai/HiDream-O1-Image, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *International Conference on Learning Representations*, volume 2026, pp. 86729–86758, 2026.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023.

Ziye Li and Henghui Ding. Occlusionformer: Arranging z-order for layout-grounded image generation. *arXiv preprint arXiv:2605.21343*, 2026.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

OpenAI. Dall-e 3. https://openai.com/index/dall-e-3/, 2023.

OpenAI. Addendum to gpt-4o system card: 4o image generation. `https://openai.com/index/gpt-4o-image-generation-system-card-addendum/`, 2025a.

OpenAI. The new chatgpt images is here. `https://openai.com/index/new-chatgpt-images-is-here/`, 2025b.

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026a.

OpenAI. Introducing chatgpt images 2.0. `https://openai.com/index/introducing-chatgpt-images-2-0/`, 2026b.

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 643–654, 2023.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork. https://qwen.ai/blog?id=qwen3.8, August 2026.

Oindrila Saha, Vojtech Krs, Radomir Mech, Subhransu Maji, Matheus Gadelha, and Kevin Blackburn-Matzen. 3d space as a scratchpad for editable text-to-image generation. *arXiv preprint arXiv:2601.14602*, 2026.

Seed Team. Beyond generation, it understands design — introducing seedream 5.0 pro. https://seed.bytedance.com/en/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro, 2026.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

stability.ai. Introducing stable diffusion 3.5. https://stability.ai/news-updates/introducing-stable-diffusion-3-5, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tongyi-MAI. Z-Image. `https://github.com/Tongyi-MAI/Z-Image`, 2025.

Junsheng Wang, Nieqing Cao, Yan Ding, Mengying Xie, Fuqiang Gu, and Chao Chen. Ske-layout: Spatial knowledge enhanced layout generation with llms. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19414–19423. IEEE, 2025a.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6232–6242. IEEE, 2024.

Zehan Wang, Jiayang Xu, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Genspace: Benchmarking spatially-aware image generation. *Advances in Neural Information Processing Systems*, 38, 2025b.

Zengbin Wang, Xuecai Hu, Yong Wang, Feng Xiong, Man Zhang, and Xiangxiang Chu. Everything in its place: Benchmarking spatial intelligence of text-to-image models. *arXiv preprint arXiv:2601.20354*, 2026.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, Bairui Li, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047, 2024.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7418–7427. IEEE, 2023.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024.

Hui Zhang, Dexiang Hong, Yitong Wang, Jie Shao, Xinglong Wu, Zuxuan Wu, and Yu-Gang Jiang. Creatilayout: Siamese multimodal diffusion transformer for creative layout-to-image generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18487–18497. IEEE, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Yucheng Zhou, Jiahao Yuan, and Qianning Wang. Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation. *arXiv preprint arXiv:2505.24787*, 2025.

Le Zhuo, Songhao Han, Yuandong Pu, Boxiang Qiu, Sayak Paul, Yue Liao, Yihao Liu, Jie Shao, Xi Chen, Si Liu, et al. Factuality matters: When image generation and editing meet structured visuals. In *International Conference on Learning Representations*, volume 2026, pp. 28249–28275, 2026.

Kai Zou, Hongbo Liu, Dian Zheng, Jianxiong Gao, Zhiwei Zhao, and Bin Liu. Echogen: Cycle-consistent learning for unified layout-image generation and understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14068–14076, 2026.

A AI ASSISTANCE

We used generative AI tools to generate synthetic training data for SCAFFOLDGEN and benchmark prompts for SLAB. We did not use these tools for other tasks requiring disclosure; mathematical proof-related tasks are not applicable to this work. Additionally, we used generative AI tools to draft sections of the paper and polish the writing. We reviewed all AI-assisted work, including the generated data, and checked the accuracy of facts and citations in AI-assisted text. We take responsibility for the final content of this work, including all AI-assisted text, claims, and artifacts.

B ADDITIONAL BENCHMARK RESULTS

Tables 5 and 6 report results on the five remaining task categories of SLAB and SpatialGenEval, respectively, complementing Tables 1 and 2 in the main text.

Table 5: Results on the five SLAB task categories omitted from Table 1.

Model	Axial Sequence	Bilateral Layout	Main-Detail Layout	Nested Layout	Partition Layout
Closed-source generators					
Nano-Banana-Pro (Google, 2025a)	87.05	79.70	78.68	85.20	93.33
GPT-Image-1.5 (OpenAI, 2025b)	84.94	66.54	70.79	80.92	70.91
GPT-Image-2 (OpenAI, 2026b)	94.58	86.09	84.74	92.43	94.85
Seedream-5-Pro (Seed Team, 2026)	88.55	83.83	82.11	95.39	90.91
Open-source generators					
FLUX.1-dev (Black Forest Labs, 2024)	42.17	37.97	26.05	42.43	21.21
FLUX.2-klein (Black Forest Labs, 2026)	75.00	59.77	58.42	76.64	48.18
FLUX.2-dev (Black Forest Labs, 2025a)	75.60	58.27	73.95	74.01	69.09
Hidream-O1-Image (HiDream-ai, 2026)	58.43	61.65	30.26	71.05	53.33
Z-Image (Tongyi-MAI, 2025)	77.71	69.55	64.74	80.59	54.55
Qwen-Image (Wu et al., 2025)	80.12	70.30	67.11	68.09	59.70
SCAFFOLDGEN	84.04	77.44	71.32	84.87	86.97
SCAFFOLDGEN-PFT	84.34	77.44	64.47	76.97	86.06
SCAFFOLDGEN-Flow	93.07	82.33	80.53	88.16	82.73

Table 6: Results on the five SpatialGenEval task categories omitted from Table 2.

Model	Attribute	Layout	Occlusion	Motion	Causal
Closed-source methods					
DALL-E-3 (OpenAI, 2023)	67.9	63.3	28.0	75.2	77.4
GPT-Image (OpenAI, 2025a)	74.1	70.4	30.2	80.9	82.2
Nano-Banana (Google, 2025b)	75.3	**70.9**	33.5	81.4	82.2
Seed Dream 4.0 (Seedream et al., 2025)	80.2	70.1	**33.8**	**83.0**	**83.8**
Open-source methods					
SD-3.5-L (stability.ai, 2024)	72.0	62.7	27.4	69.4	72.6
FLUX.1-dev (Black Forest Labs, 2024)	73.8	66.7	28.9	73.1	73.8
FLUX.1-krea (Black Forest Labs, 2025b)	75.4	67.1	28.0	76.0	78.8
FLUX.2-klein (Black Forest Labs, 2026)	78.5	70.2	30.8	75.9	79.1
Qwen-Image (Wu et al., 2025)	77.2	69.7	30.8	78.1	80.2
SCAFFOLDGEN	**80.7**	69.3	31.4	79.6	81.2