Towards Resolving Positional Control Failures in Diffusion Transformers

Anonymous authors
Paper under double-blind review

Abstract

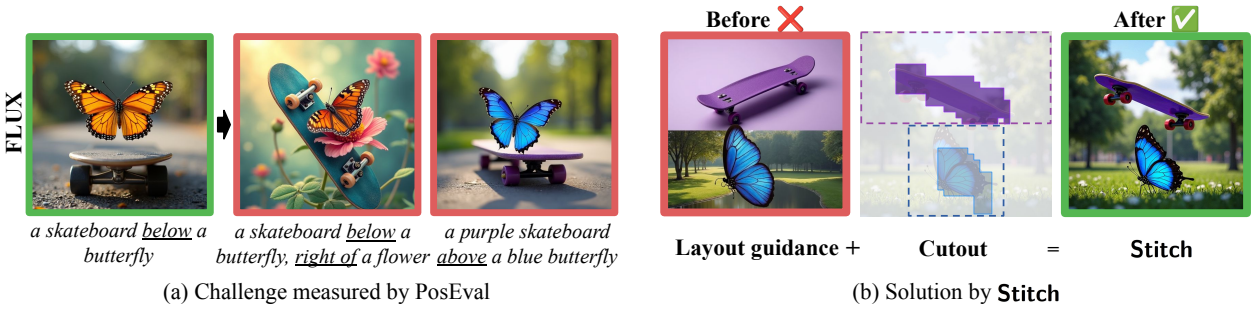
Text-to-image (T2I) generation has advanced rapidly, and the historically difficult Position task now appears largely solved in top models. However, our analysis challenges this view, showing that even state-of-the-art (SOTA) models struggle with more complex position-related tasks. We introduce PosEval, a benchmark comprising five new challenging position-based tasks, designed for in-depth evaluation of existing T2I models and to support the development of next-generation layout guidance methods. Leveraging PosEval for evaluation, we propose a new layout-guidance method compatible with modern architectures. Prior work shows that transferring layout guidance from U-Net to transformer models introduces blending artifacts, a challenge addressed in existing methods by averaging sub-predictions to enforce coherence. We introduce Stitch, which instead uses segmentation to extract and compose foreground objects, yielding a latent the base model can directly transform into a coherent image. Tested on Qwen-Image, FLUX, and SD3.5, Stitch consistently enhances base models, even improving FLUX by 218% on GenEval's *Position* task and by 206% on PosEval. Stitch achieves SOTA results with Qwen-Image on PosEval, improving over previous models by 54%, all while integrating position control into leading models training-free.

1 Introduction

Text-to-Image (T2I) generation models provide a powerful bridge between natural language and visual content. By transforming text descriptions into images, they expand human creativity, accelerate prototyping, and enable a wide range of applications (Chen et al., 2023; Zhou & Lee, 2024). But as their popularity grows, so do user demands. These models should interpret and fulfill requests that are both complex and linguistically nuanced in their descriptions of positioning. They should be able to generate real-world scenes, even if the object arrangements are uncommon. These needs require strong positional understanding—a capability long recognized as challenging (Ghosh et al., 2023; Huang et al., 2023), yet one in which modern models have reported rapid progress (Wu et al., 2025a). For example, BLIP3-o Chen et al. (2025a) achieves 87% on GenEval's *Position* task, suggesting that this important challenge may be nearing resolution.

Our closer examination of positional generation challenges this assumption: with our benchmark, PosEval, we find that even top-performing generation models and layout-guidance methods struggle as task complexity increases. In fact, performance on simple positional tasks does not always translate to stronger results on more demanding prompts, such as those with more objects or complex spatial relationships. This suggests that some state-of-the-art (SOTA) models may be overfitting to existing benchmarks and highlights the need for a more targeted evaluation framework to accurately assess positional generative performance–a gap PosEval is designed to fill. An extension of GenEval featuring 500 new prompts in five categories designed to probe specific failure modes under more challenging conditions, PosEval paves the way for the development of next-generation of T2I models by providing a comprehensive way to benchmark positional capabilities.

We contribute to this development by improving layout guidance, in which the model is conditioned on object-level bounding boxes to adhere to specific spatial layouts. This can be used to improve spatial prompt following by decomposing the prompt with a Large Language Model (LLM) into sub-generations grounded in bounding boxes (Fang et al., 2025; Feng et al., 2023). However, the recent shift from U-Net



(a) Challenge measured by PosEval (b) Solution by Stitch

Figure 1: (a) Existing benchmarks target basic positioning tasks, yet T2I models still struggle with complex spatial reasoning; we introduce PosEval to measure these gaps. (b) Applying U-Net–based layout guidance to transformer-based models causes inadequate blending (left), which Stitch addresses with Cutout (middle): tight extraction of foreground objects enables seamless composition without additional intervention (right).

to transformer-based architectures has introduced challenges in conditioning the model on these bounding boxes; in particular, methods developed on U-Net models produce noticeable blending artifacts when naively applied to transformer-based architectures (Chen et al., 2024a; 2025c) (see Figure 1b left). To mitigate these, existing approaches smooth the transition from the individual sub-generations to the shared one by averaging sub-predictions (Chen et al., 2024a; 2025c). We take another approach, instead composing the sub-generations in one step in a way that base model can directly generate a coherent image without additional intervention. We hypothesize that failure to blend arises from conflicting background information with abrupt color transitions that the model cannot adequately smooth. Therefore, we propose using tight segmentation masks—inexpensively derived directly from attention heads—to isolate and extract foreground objects (see Figure 1b middle), before composing them together on a single background to eliminate discrepancy between backgrounds from sub-generations (see Figure 1b right) and enable effective scaling to prompts of high complexity (see Figure 2). We call our method Stitch and show on PosEval that it both effectively blends sub-regions and generates more accurate images than prior layout guidance approaches.

The primary contributions of this work are: (1) we present PosEval, a GenEval (Ghosh et al., 2023) extension introducing five new position-focused tasks targeting the next level of generation challenges; (2) with PosEval, we show position control remains far from solved in SOTA models; and (3) we introduce Stitch, an alternative approach to blending for layout control in transformer-based models which substantially improves accuracy.

2 Related Works

T2I generation. T2I synthesis rapidly advanced in past years (Podell et al., 2023; OpenAI, 2023; Midjourney, 2025; OpenAI, 2024; DeepMind, 2025), moving from diffusion-based models like Stable Diffusion (Rombach et al., 2022), Imagen (Saharia et al., 2022), and DALL · E (OpenAI, 2023) to powerful FM approaches like FLUX (BlackForest, 2024), SD3.5 (Esser et al., 2024), and HiDream (Cai et al., 2025), advancing fidelity, diversity, and controllability. They enhance creativity, productivity, and communication (Chen et al., 2023; Zhou & Lee, 2024; Zhou et al., 2024). Previous works improved pre-trained models by expanding conditioning beyond prompts (Pach et al., 2026), aligning with human preferences (Karthik et al., 2025; Liang et al., 2024; Bravo, 2025; Sendera et al., 2025), and personalizing (Liu et al., 2023b; Ruiz et al., 2022; Kim et al., 2024). Despite these advances, positional control remains underexplored, in part due to limited benchmarks.

Positional benchmarks. Many benchmarks evaluate image generation, with GenEval (Ghosh et al., 2023) and T2I-CompBench++ (Huang et al., 2023; 2025) among the most widely used; both include basic positional tasks. While some benchmarks target positional understanding, many focus on basic formats lacking task complexity (Gokhale et al., 2022), or shift to related spatial tasks like shape generation (Sim et al., 2024) or 3D spatial reasoning (Wang et al., 2025b). Other general benchmarks include 2D spatial tasks (Li et al., 2025; Wang et al., 2025a) which may extend to 3 or 4 objects (Bakr et al., 2023; Feng et al., 2023), or evaluate prompts that are natural (Feng et al., 2023) or long (Hu et al., 2024), but lack explicit structure or precise positional tasks. Our proposed PosEval advances positional evaluation with a comprehensive, targeted suite of complex tasks, with further discussion of its differences from existing literature provided in Appendix A.4.

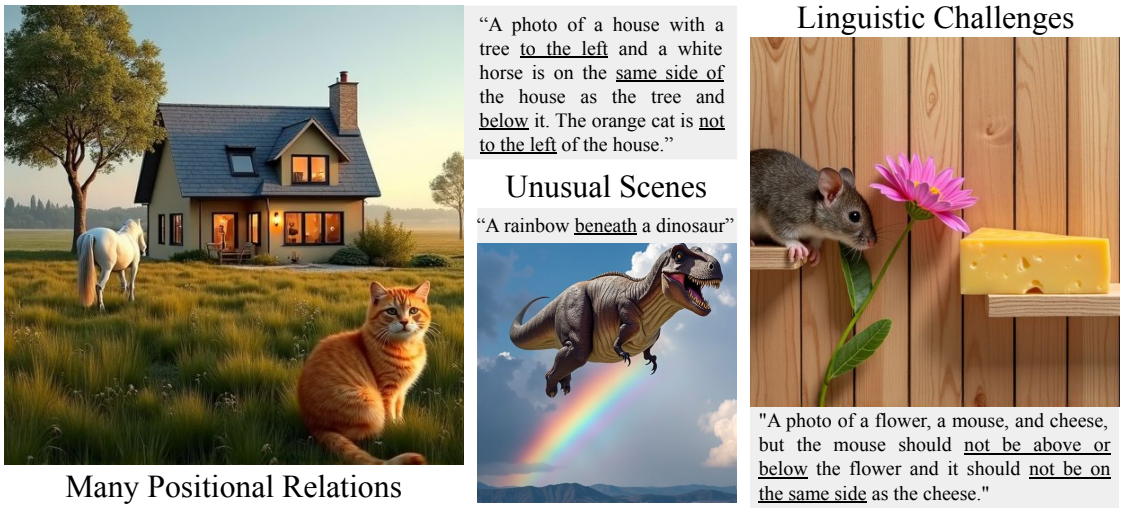


Figure 2: Stitch excels at complex positional prompts.

Position control. Many base models tried to improve spatial limitations on basic tasks during training, e.g., with a joint understanding-generation objective (Wu et al., 2025b; Ma et al., 2025; Chen et al., 2025a;b; Deng et al., 2025) or Multimodal LLM (MLLM)-based prompt embeddings (Wu et al., 2025a; Fang et al., 2025; Wu et al., 2025c;d). Spatial guidance, like bounding boxes, was used post hoc on U-Net-based models to improve positional control (Fang et al., 2025; Feng et al., 2023; Yang et al., 2024; Lian et al., 2024; Zhang et al., 2023b; 2024), and layout-conditioned T2I synthesis (Zhang et al., 2023a; Li et al., 2023; Farshad et al., 2023; Chen et al., 2024b; Mou et al., 2024; Ohanyan et al., 2024; Xie et al., 2023; Mo et al., 2024), which is more constrained by fine-grained user input. With PosEval, we show these improvements remain insufficient.

Moreover, these methods targeted U-Net architectures and caused blending artifacts on transformer models (Chen et al., 2025c; 2024a). This poses a significant limitation, as many recent models are transformer-based (Esser et al., 2024; Wu et al., 2025a; BlackForest, 2024; Cai et al., 2025). In fact, previous methods did not generalize to transformer-based architectures for other tasks as well (Jiao et al., 2025), and latent averaging emerged as a key strategy (Bader et al., 2025; Jiao et al., 2025; Chen et al., 2025c; 2024a). Existing work explored latent averaging for blending spatially-guided generation in transformer-based models (Chen et al., 2025c; 2024a). However, cross-attention can yield segmentation maps (Kim et al., 2025; Kukleva et al., 2025), which we use to propose an alternative blending strategy: Stitch extracts foreground objects, enabling coherent composition without additional intervention and improving accuracy.

3 PosEval: Benchmarking Positional Accuracy in T2I Generation

As models improve on basic *Position* tasks of form $\langle obj_1 \rangle \langle relation_{12} \rangle \langle obj_2 \rangle$, broader benchmarks are necessary to more thoroughly evaluate T2I spatial capabilities. We introduce PosEval, featuring tasks going beyond basic *Position* in difficulty, yet maintaining clear and specific task definitions.

3.1 Benchmark Composition

To ensure ease of use, PosEval builds upon GenEval, using Mask2Former-based (Cheng et al., 2022) object detection and procedural relation verification extended to the new tasks. We reuse the same set of objects $\langle obj \rangle$ and relations $\langle relation \rangle$, and adopt their protocol of generating 100 prompts per category and four images per prompt. In addition to GenEval's *Position* task, we introduce five new tasks, each designed to isolate and evaluate a specific aspect of positional capability. The new tasks include:

Two Objects (2 Obj): For completeness, we inherit GenEval's *Position* task, unchanged. *E.g., "a photo of an elephant below a surfboard"*

Three Objects (3 Obj): We extend the basic *Position* task to three objects linked by two spatial relations. Objects $\langle obj_1 \rangle, \langle obj_2 \rangle, \langle obj_3 \rangle$ are arranged consecutively in a chain wrapped to fit on a 2×2 grid, ensuring one horizontal and one vertical relation. For a single $\langle relation_{ij} \rangle$ between adjacent $\langle obj_i \rangle$ and $\langle obj_j \rangle$, we use $f(\langle relation_{ij} \rangle) :=$ "The $\langle obj_i \rangle$ is $\langle relation_{ij} \rangle$ the $\langle obj_j \rangle$" or "The $\langle obj_j \rangle$ is $\langle relation_{ji} \rangle$ the $\langle obj_i \rangle$", with equal probability. The task prompts are built in the form: "A photo of a $\langle obj_1 \rangle$, a $\langle obj_2 \rangle$, and a $\langle obj_3 \rangle$. $f(\langle relation_{12} \rangle)$. $f(\langle relation_{23} \rangle)$." Afterwards, objects in the first sentence are shuffled, along with sentences 2 and 3. Objects are listed upfront to avoid ambiguity. The model is tested on correctly interpreting the relations, not the underlying grid. *E.g., "a photo of a baseball glove, a cell phone, and a teddy bear. the baseball glove is above the cell phone. the cell phone is right of the teddy bear."*

Four Objects (4 Obj): We further extend to four objects connected by four spatial relations, otherwise using the same setup as in the *3 Obj* task. The task prompts are built in the form: "A photo of a $\langle obj_1 \rangle$, a $\langle obj_2 \rangle$, a $\langle obj_3 \rangle$, and a $\langle obj_4 \rangle$. $f(\langle relation_{12} \rangle)$. $f(\langle relation_{23} \rangle)$. $f(\langle relation_{34} \rangle)$. $f(\langle relation_{41} \rangle)$." Again, objects in sentence 1 and sentences 2–5 are randomly permuted, and the model is evaluated on correctly interpreting the specified spatial relations. *E.g., "a photo of a pizza, a laptop, an oven, and a vase. the pizza is left of the laptop. the oven is below the pizza. the vase is right of the oven. the vase is below the laptop."*

Positional Attribute Binding (PAB): In the form "a $\langle attr_1 \rangle \langle obj_1 \rangle \langle relation_{12} \rangle$ a $\langle attr_2 \rangle \langle obj_2 \rangle$", PAB extends *Attribute Binding* ("a $\langle attr_1 \rangle \langle obj_1 \rangle$ and a $\langle attr_2 \rangle \langle obj_2 \rangle$") with inter-object positions. *E.g., "a photo of a yellow dining table right of a pink dog"*

Negative Relations (Neg): We use prompts of form, "a photo of a $\langle obj_1 \rangle$ and a $\langle obj_2 \rangle$, a $\langle obj_1 \rangle$ is not $\langle relation_{12} \rangle$ a $\langle obj_2 \rangle$", with objects listed for clarity. It is evaluated that: (1) both objects are present, and (2) $\langle obj_2 \rangle$ appears anywhere *except* in the specified relation to $\langle obj_1 \rangle$. Prompts are derived from GenEval's position prompts by replacing the relation with a negation of the opposite relation, keeping the same target image (e.g. "a photo of a dog right of a teddy bear" → "a photo of a dog and a teddy bear, a dog is not left of a teddy bear"). *E.g., "a photo of a cow and a laptop, a cow is not left of a laptop"*

Relative Relations (Rel): We evaluate understanding of relations *relative to other relations*. Prompts contain three objects and two spatial relations, the first in the form $\langle relation_{ij} \rangle$ and the second defined relative to the first. Relative relations $\langle rel.relation \rangle$ can be same or opposite, each comprising half the prompts. The *same* relation is "on the same side of", while the *opposite* relations include "on the other side of", "on the opposite side of", and "on the contrary side of", for diversity. Prompts take the form, "a photo of a $\langle obj_1 \rangle$ $\langle relation_{12} \rangle$ $\langle obj_2 \rangle$, and a $\langle obj_3 \rangle$ $\langle rel.relation \rangle$ the $\langle obj_2 \rangle$ $\langle prep \rangle$ the $\langle obj_1 \rangle$", with $\langle prep \rangle \in$ {"for", "as"} chosen to be grammatically correct. *E.g., "a photo of an apple left of a computer mouse, and a teddy bear on the same side of the computer mouse as the apple"*

Since PosEval inherits both the objects and detection methodology from GenEval, its overall accuracy remains largely unchanged—a result that has already been thoroughly validated in GenEval. As the only modification involves minor changes to the logical assertions between detections, we refer the reader to Appendix A.2 for validation that PosEval's final composed tasks maintain sufficient accuracy.

3.2 Evaluating Top T2I Models with PosEval

Overall, PosEval shows that while SOTA models perform well on basic *Position* prompts, they struggle as task complexity increases, suggesting that the broader challenge of positional reasoning remains unsolved (Table 2, averaged over three seeds). While models like BLIP3-o (Chen et al., 2025a) and JanusPro (Chen et al., 2025b) previously appeared strong in positional understanding—achieving 87% and 79% on the basic task, respectively—PosEval reveals that their accuracy drops sharply as task difficulty increases. On *Relative Relations*, they score only 8% and 11%, and when scaling to *Four Objects*, their accuracy falls to 2% and 4%. PosEval exposes shortcomings in top models that the basic *Position* task alone cannot capture.

Designing PosEval tasks to also be complementary enables additional insights through comparison. Comparing *2 Obj*, *3 Obj*, and *4 Obj* reveals how models handle positional generation as object count and relationships grow. As shown in Table 2, strong *2 Obj* performance does not guarantee scaling. BLIP3-o (Chen et al., 2025a) scores 87% on *2 Obj* but only 15% and 2% on *3 Obj* and *4 Obj*, while Qwen-Image (Wu et al., 2025a), with lower initial scores, maintains better performance at 40% and 21%. This suggests that some models

Table 1: PosEval reveals that leading T2I models still struggle with complex positional prompts.

Model	2 Obj	3 Obj	4 Obj	Neg	Rel	PAB	Avg.↑
FLUX.1 [Dev] (BlackForest, 2024)	0.22	0.06±0.01	0.02±0.00	0.62±0.01	0.03±0.01	0.15±0.02	0.18
SD3 Medium (Esser et al., 2024)	0.34	0.05±0.01	0.01±0.00	0.62±0.02	0.05±0.01	0.14±0.01	0.20
SD3.5 Large (Esser et al., 2024)	0.34	0.06±0.00	0.02±0.01	0.64±0.03	0.06±0.02	0.16±0.01	0.21
HiDream-I1-Full (Cai et al., 2025)	0.60	0.19±0.01	0.10±0.00	0.66±0.00	0.09±0.01	0.29±0.01	0.32
BAGEL (Deng et al., 2025)	0.64	0.23±0.01	0.16±0.02	**0.73**±0.01	0.07±0.01	0.22±0.02	0.34
OpenUni-B-512 (Wu et al., 2025d)	0.77	0.09±0.01	0.03±0.00	0.56±0.01	0.01±0.00	0.60±0.01	0.34
Janus-Pro (Chen et al., 2025b)	0.79	0.14±0.01	0.04±0.01	0.69±0.02	0.11±0.00	0.46±0.03	0.37
BLIP3-o (Chen et al., 2025a)	**0.87**	0.15±0.00	0.02±0.00	0.55±0.01	0.08±0.01	**0.63**±0.01	0.38
Qwen-Image (Wu et al., 2025a)	0.76	0.40±0.01	0.21±0.02	0.49±0.00	0.10±0.01	0.61±0.03	0.43
FLUX.2 [dev] (Labs, 2025)	0.69	**0.50**±0.02	**0.50**±0.01	0.61±0.01	**0.17**±0.00	0.55±0.00	**0.50**

may overfit to previously seen tasks without capturing the underlying mechanisms required for positional generation. Although *Negative* prompts generate the same images as *2 Obj* prompts, they are easier due to looser criteria (e.g., "not right" vs. "left"). If a model scores higher on *2 Obj* than *Neg* (e.g. Janus-Pro's 79% vs. 69%), it likely reflects prompt interpretation issues rather than image generation. Similarly, both *Rel* and *3 Obj* include three objects and two spatial relations, but *Rel* is linguistically more complex and restricts relations to one axis, while *3 Obj* spans two axes. For example, Qwen-Image scores 40% on *3 Obj* but only 10% on *Rel*, and 76% on *2 Obj* versus 49% on *Neg*, indicating difficulty with complex prompts.

4 Stitch: Advancing Training-Free Position Control

Building on the more detailed evaluation provided by PosEval, we propose targeted methodological improvements focusing on transformer-based models, as many of the leading T2I models fall in this category (BlackForest, 2024; Esser et al., 2024; Wu et al., 2025a). Layout control methods designed for U-Net models introduce blending artifacts when applied to transformer-based models (Chen et al., 2025c; 2024a), which prior work addressed by averaging the latents of sub-prompts; instead, we observe that it appears to arise from the need to reconcile contradictory information in the backgrounds. We therefore improve blending by leveraging precise segmentation maps to isolate foreground objects, removing the difficult-to-blend context.

4.1 Preliminaries

Recent generative models define generation as a time-dependent transformation from Gaussian noise $x_0 \sim \mathcal{N}(0, I)$ to data $x_1 \sim p_{\text{data}}$, expressed as $x_\tau = \alpha_\tau x_0 + \sigma_\tau x_1$, with α_τ decreasing and σ_τ increasing over time τ. FM models (Albergo & Vanden-Eijnden, 2023; Liu et al., 2023a; Lipman et al., 2023; Esser et al., 2024) treat this trajectory as a differential equation, with learned approximation of its conditional vector field. Simulating this equation from noise to data in T steps yields samples.

Many models operate in a VAE latent space (Kingma & Welling, 2022), encoding an image into latent token sequence $\mathcal{Z}_v = \{z_v^i\}_{i=1}^{N_v}$. To condition generation, a text encoder (e.g. T5 (Raffel et al., 2019) or CLIP (Radford et al., 2021)) embeds the prompt as $\mathcal{Z}_t = \{z_t^i\}_{i=1}^{N_t}$. Learnable projections E_v and E_t map both sets into a shared embedding space, yielding sets of *visual and text vectors*:

$$\mathcal{X}_v = \{x_v^i\}_{i=1}^{N_v} = E_v(\mathcal{Z}_v), \quad \mathcal{X}_t = \{x_t^i\}_{i=1}^{N_t} = E_t(\mathcal{Z}_t),$$

where each $x_v^i, x_t^i \in \mathbb{R}^d$ is a vector in the sequence. MMDiT (Esser et al., 2024) processes $\mathcal{X}_v$ and $\mathcal{X}_t$ iteratively with stacked, time-conditioned transformer blocks. Each block applies a pre-attention transformation $(\tilde{\mathcal{X}}_v, \tilde{\mathcal{X}}_t) = F_{\text{pre}}(\mathcal{X}_v, \mathcal{X}_t; \tau)$, including normalization and timestep modulation. Let $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_v \cup \tilde{\mathcal{X}}_t = \{\tilde{x}_i\}_{i=1}^{N}, N = N_v + N_t$ denote all pre-attention vectors. Each $\tilde{x}_i \in \tilde{\mathcal{X}}$ is mapped to queries, keys, and values

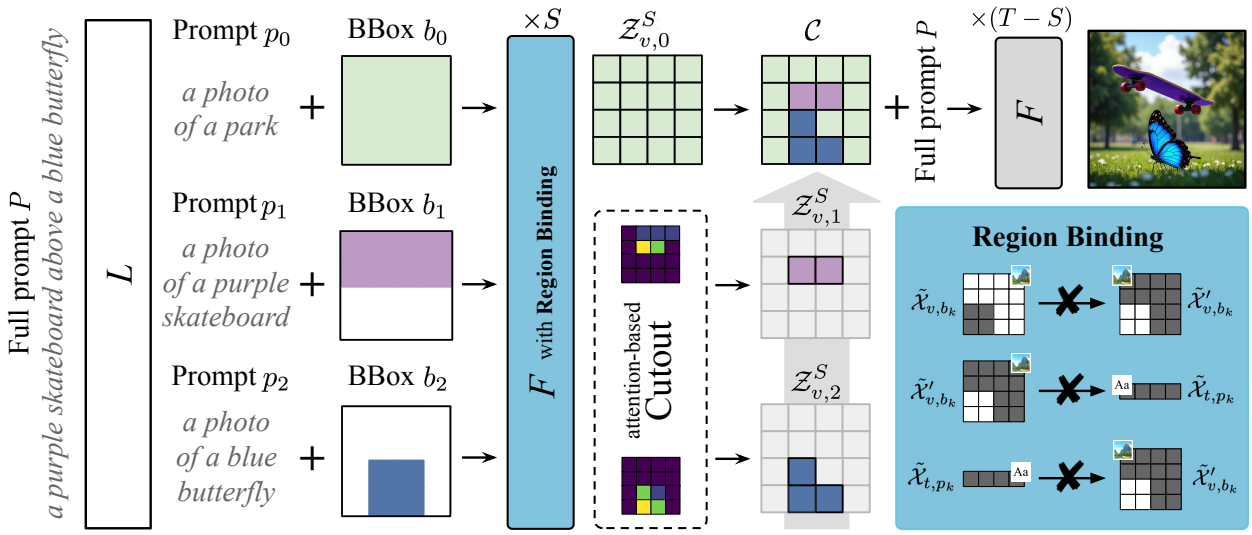


Figure 3: Stitch: LLM L splits full prompt P into object prompts p_k and bounding boxes b_k, along with full-image background prompt p_0. MMDiT-based model F seperately sketches objects and background (*butterfly* ■, *skateboard* ■, *park* ■) for S timesteps. With **Region Binding** attention-masking constraints, F generates each p_k in b_k. In **Cutout**, the highest attention weights in a targeted head select tighter latent regions linked to foreground objects $\mathcal{Z}^S_{v,k}$, which are merged with the background latents $\mathcal{Z}^S_{v,0}$ to form composite latent $\mathcal{C}$. For the remaining steps, the unconstrained F seamlessly stitches the sketches into a coherent image conditioned on the full prompt.

via functions Q, K, V derived from linear projections and normalization. Attention outputs are computed:

$$z(\tilde{x}_i) = \sum_{\tilde{x}_j \in \tilde{\mathcal{X}}} \sigma_j \left(\frac{\langle Q(\tilde{x}_i), K(\tilde{x}_j) \rangle}{\sqrt{d}} + M(\tilde{x}_i, \tilde{x}_j) \right) V(\tilde{x}_j),$$

where $M : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \{0, -\infty\}$ is an attention mask, $\mathcal{Z} = \{z(\tilde{x}_i)\}_{i=1}^N$ the set of attention outputs, and σ a softmax operation. Finally, a post-attention transformation $(\mathcal{X}_v^*, \mathcal{X}_t^*) = F_{\text{post}}(\mathcal{Z}, \mathcal{X}_v, \mathcal{X}_t; \tau)$ updates the representations. Attention is often multi-headed, and each head h computes attention independently over $\tilde{\mathcal{X}}$ with its own Q_h, K_h, V_h. Outputs are combined into $\mathcal{Z}$. Stitch builds upon MMDiT's design, particularly with targeted masking within the multimodal transformer blocks.

4.2 Basics of Bounding Box Integration

As seen in Figure 3, Stitch begins by using LLM L to decompose the full text prompt P into K sub-prompts $\{p_k\}_{k=1}^K$, each associated with a corresponding bounding box $\{b_k\}_{k=1}^K$, where $b_k = [x_{\min}, x_{\max}, y_{\min}, y_{\max}] \in \{0, \ldots, W-1\}^4$. Each pair (p_k, b_k) represents a distinct object. Additionally, L generates a background prompt p_0, with a bounding box b_0 that spans the entire image. For the first S steps, the FM model F separately generates everything within the designated bounding boxes via Region Binding.

Region Binding To ensure objects are fully generated in their bounding boxes, we use three attention-masking constraints, isolating the object from the surrounding context. For each sub-prompt p_k and bounding box b_k, let $\tilde{\mathcal{X}}_{v,b_k} \subseteq \tilde{\mathcal{X}}_v$ denote the visual vectors inside the bounding box, $\tilde{\mathcal{X}}'_{v,b_k} = \tilde{\mathcal{X}}_v \setminus \tilde{\mathcal{X}}_{v,b_k}$ those outside, and $\tilde{\mathcal{X}}_{t,p_k}$ the text vectors for the sub-prompt. In each layer and head we apply the attention mask M with the following constraints: (1) block attention from inside to outside the bounding box, (2) block attention from outside the bounding box to the text, (3) block attention from the text to outside the bounding box:

$$M(\tilde{\mathcal{X}}_{v,b_k}, \tilde{\mathcal{X}}'_{v,b_k}) = -\infty, \quad M(\tilde{\mathcal{X}}'_{v,b_k}, \tilde{\mathcal{X}}_{t,p_k}) = -\infty, \quad M(\tilde{\mathcal{X}}_{t,p_k}, \tilde{\mathcal{X}}'_{v,b_k}) = -\infty$$

The background bounding box b_0 spans the entire image, so $\tilde{\mathcal{X}}_{v,b_0} = \tilde{\mathcal{X}}_v$, imposing no constraints.

Table 2: Evaluated with PosEval, Stitch boosts positional generation on three different MMDiT-based models, achieving SOTA over training-based layout guidance methods while remaining training-free.

Model	2 Obj	3 Obj	4 Obj	Neg	Rel	PAB	Avg.↑
Layout guidance, with training							
GoT (Fang et al., 2025)	0.34	0.02±0.00	0.00±0.00	0.52±0.01	0.02±0.01	0.09±0.01	0.17
LayoutGPT (Feng et al., 2023)	0.41±0.04	0.18±0.03	0.10±0.02	0.37±0.03	0.25±0.02	0.07±0.02	0.23
Layout guidance, training-free							
RPG (Yang et al., 2024)	0.28±0.00	0.07±0.01	0.01±0.00	0.65±0.00	0.06±0.01	0.14±0.01	0.20
RAG (Chen et al., 2025c)	0.50±0.00	0.11±0.00	0.04±0.01	0.61±0.01	0.13±0.01	0.21±0.01	0.27
Reg. Prompting (Chen et al., 2024a)	0.25±0.01	0.17±0.00	0.28±0.03	0.68±0.00	0.30±0.01	0.19±0.01	0.31
LMD (Lian et al., 2024)	0.74±0.01	0.43±0.02	0.23±0.01	0.54±0.01	0.35±0.01	0.46±0.01	0.46
Stitch (ours) + SD3.5 Large	0.53±0.02	0.22±0.03	0.12±0.01	0.79±0.00	0.27±0.02	0.37±0.02	0.38
Gain over SD3.5 Large	+0.19	+0.16	+0.10	+0.15	+0.21	+0.21	+0.17
Stitch (ours) + FLUX.1 [Dev]	0.70±0.02	0.44±0.01	0.38±0.01	0.83±0.02	**0.48**±0.02	0.44±0.00	0.55
Gain over FLUX.1 [Dev]	+0.48	+0.38	+0.36	+0.21	+0.45	+0.29	+0.37
Stitch (ours) + Qwen-Image	**0.87**±0.01	**0.67**±0.01	**0.61**±0.01	**0.93**±0.00	0.43±0.01	**0.77**±0.00	**0.71**
Gain over Qwen-Image	+0.11	+0.27	+0.40	+0.44	+0.33	+0.16	+0.28

4.3 Improving Blending with Cutout

Once objects are adequately formed, we extract only the foreground information from each sub-prompt by cutting their corresponding foreground latent tokens $\mathcal{Z}_{v,k}^{S}$ from the latent maps. We refer to this as Cutout. As generation is incomplete at step S, conventional segmentation tools like SAM (Kirillov et al., 2023) cannot be used to extract the object. However, we observe that in some attention heads, text vectors focus on foreground objects. We outline an automated process for selecting such a head in Section 4.4.

Given such a head, we construct the Cutout mask by selecting visual tokens in descending order of attention weight. The attention weight for each visual token is computed as the average attention that its corresponding visual vector gets from all non-padding text vectors. We select tokens until their attention weights sum to a fraction $\eta \in [0, 1]$ of the total attention assigned to all visual tokens. The mask is then smoothed with 2D max pooling of kernel size κ. The extracted foreground latent tokens $\mathcal{Z}_{v,k}^{S}$ are combined with the common background latent tokens $\mathcal{Z}_{v,0}^{S}$ into a single composite latent $\mathcal{C}$, used exclusively for the remainder of the generation. The model proceeds without constraints and conditioned on full prompt P, allowing F to enhance the overall quality and consistency while completing the image.

4.4 Automated Selection of the Cutout Head

To select a head for a given model, we propose a short automated process. First, we generate 80 images with the prompt "a photo of a ⟨obj⟩" using the 80 GenEval objects. After step S, we save text-to-image attention averaged over non-padding tokens, avoiding unwanted behavior introduced by padding. We test thresholds $\eta \in \{0.75, 0.80, 0.85, 0.90, 0.95, 0.97, 0.99\}$ and extract SAM (Kirillov et al., 2023) masks generated from the final images to use for ground truth, excluding objects with inadequate SAM masks. We evaluate the Intersection over Union (IoU) between masks generated by each attention head across the different η values, choosing the head-threshold combination with the highest IoU. With this procedure, a Cutout head can be selected once per model and applied at test time for all images.

5 Experiments

We evaluate Stitch in terms of accuracy and robustness. In Stitch, $S = 10$ for FLUX and SD3.5 and 6 for Qwen-Image. For all three models, $T = 50$, $\kappa = 5$, and bounding boxes b_k are generated on a $W \times W = 32 \times 32$

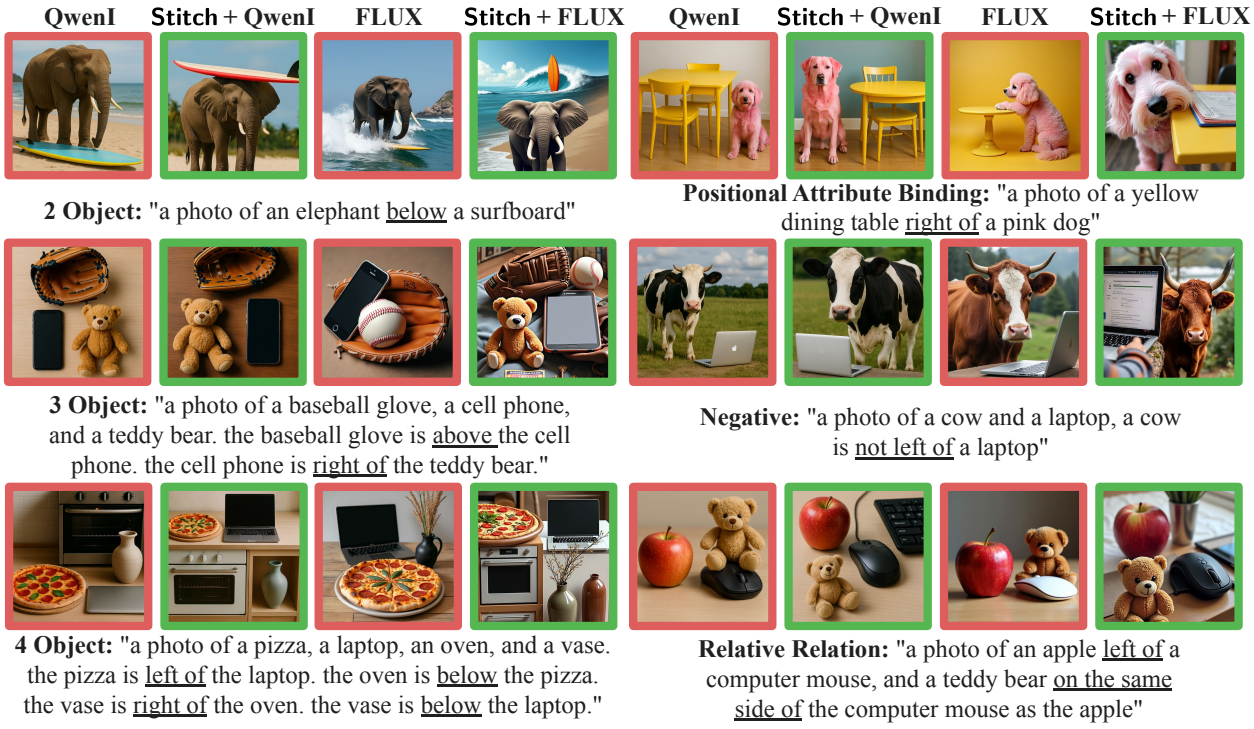


Figure 4: Stitch corrects Qwen-Image (QwenI) and FLUX position on PosEval without quality loss.

grid with prompts p_k from GPT-5 (OpenAI, 2025) and GPT-4 (OpenAI, 2023). $\eta = 0.95$ for SD3.5 and FLUX. Results are on 3 seeds. All other parameters are consistent with the base models.

5.1 Enhancing Positional Understanding with Stitch

Results on PosEval. Stitch consistently enhances base models, with up to 48 percentage point improvement (218% relative increase) over FLUX on *2 Obj* and 37 percentage point increase on PosEval overall (206% relative increase). This boost is also seen on *Position* tasks in other benchmarks (T2ICompBench (Huang et al., 2023), HRS-Bench (Bakr et al., 2023) in Appendix A.1). Stitch also outperforms prior layout-control methods: Stitch + Qwen-Image performs best on 5 out of 6 PosEval tasks, while Stitch + FLUX achieves the top performance on the final task. This results in Stitch + Qwen-Image achieving the highest overall accuracy, with a 54% relative improvement over the previous best method, LMD (Lian et al., 2024).

Improvements Over Latent Averaging. To understand why Stitch 's Cutout approach outperforms prior methods for improving blending in transformer-based models—namely Reg. Prompting (Chen et al., 2024a) and RAG (Chen et al., 2025c), which rely on averaging latents—we qualitatively analyze the generated images (images included in Appendix A.12). We find that these models are more prone than Stitch to merging multiple objects or omitting them, resulting in lower overall performance.

Qualitative Improvements Over the Base. For Qwen-Image, FLUX, and SD3.5, Stitch refines positional details without degrading visual quality, shown in Figure 4 on all PosEval categories (SD3.5 in Appendix). Even scaling to four objects, it can integrate without awkward or unnatural seams. The boost observed in Table 2 from applying Stitch can be partly attributed to reduced positional errors, such as incorrectly placing or omitting objects. These mistakes are seen when FLUX omits the oven in *Four Object*, and Qwen-Image incorrectly places the cow left of the laptop in *Negative*. Moreover, Stitch demonstrates strong comprehension of linguistically challenging prompts, like those in *Negative* and *Relative Relation*.

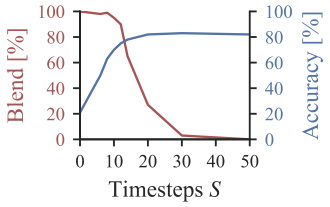


Figure 5: Accuracy vs blend while increasing Region Binding timesteps S.

Table 3: While Region Binding (RB) drives Stitch's positional control, Cutout improves the blend. By increasing Cutout threshold η to 0.95, we recover accuracy without sacrificing blend.

RB	Cutout	$\eta = 0.95$	2 Obj	3 Obj	4 Obj	Neg	Rel	PAB	Blend
			0.22	0.06±0.01	0.02±0.00	0.62±0.01	0.03±0.01	0.15±0.02	1.0
✓			0.81±0.02	0.58±0.03	0.52±0.02	0.90±0.01	0.63±0.01	0.51±0.01	0.68
✓	✓		0.51±0.01	0.30±0.01	0.18±0.01	0.71±0.03	0.31±0.04	0.29±0.01	0.99
✓	✓	✓	0.70±0.02	0.44±0.01	0.38±0.01	0.83±0.02	0.48±0.02	0.44±0.00	0.95

Complex Scenes. Figure 2 highlights Stitch + FLUX on complex scenes. On the left, we show it cleanly interprets a prompt combining all 6 PosEval categories. Despite the complexity (4 distinct objects, multiple attributes, and both negative and relative spatial relations), Stitch generates a coherent and semantically accurate image. In the center, we demonstrate that it enhances FLUX's ability to generate rare and challenging combinations, such as a dinosaur *above* a rainbow. On the right, we show that Stitch's strong language understanding enables it to interpret even confusing text and translate it into coherent visual arrangements.

Timestep Selection. We investigate S selection for Region Binding on FLUX. In Figure 5, we vary the number of timesteps S on the basic *Position* task and compare to *Blend*. Selecting later steps improves *Position* accuracy but reduces *Blend*, with a steep decline occurring between 10 and 20 steps. Offering the best balance, we select timestep 10 for FLUX. Similarly, we choose step 10 for SD3.5 and 6 for Qwen-Image.

Component Ablation. Table 3 presents an ablation on FLUX to evaluate the contributions of Region Binding, Cutout, and raising η to 0.95 to widen the object buffer. For each setup, three people independently and binarily decide if each of 100 images (one image per prompt) displays *Blend*, and we report the percentage of images for which they answer *yes*. As shown in Table 3, the Region Binding is the primary driver of Stitch's position-related performance improvements (e.g. boosting *2 Obj* from 22% to 81%). However, combining full bounding boxes requires using backgrounds from the partially generated Region Binding predictions, which may contain conflicting color information; this can hinder the model's ability to seamlessly stitch components into a coherent image. Consequently, Region Binding alone fails to properly blend 32% of images. Adding Cutout improves the *Blend* success rate to 99% of images, though at the cost of some positional accuracy. Raising the Cutout threshold to $\eta = 0.95$ recovers lost performance. While lower thresholds often aid in selecting heads focused on the foreground, a higher threshold is preferable afterward, as Stitch benefits from higher IoT. As shown in Table 3, this improves accuracy without significantly affecting *Blend*. This final combination is what allows Stitch to achieve both precise positional control and high image quality.

η Impact on Stitch: We examine η's impact on Stitch performance to validate our choice of $\eta = 0.95$. Using Cutout with the best FLUX head, Block 14 Head 20, we evaluate PosEval across different η thresholds, as shown in Table 4. We observe that increasing η improves performance on PosEval, whereas *Blend* performance declines. To maximize performance, we recommend first selecting the Cutout head and then increasing η as much as possible without causing a significant drop in *Blend* performance; we find that $\eta = 0.95$ provides a good balance and generalizes well across other models, namely Qwen-Image and SD3.5.

Impacts on Quality and Diversity. Preserving base model quality while enhancing positional capabilities is essential. We compare each base model's quality with Stitch's via Aesthetic Score (Schuhmann & Beaumont, 2022). Across four images per prompt on six PosEval tasks, Stitch results in only marginal changes: from 6.2 ± 0.8 to 6.1 ± 0.8 with Qwen, 6.3 ± 0.8 to 6.1 ± 0.7 with FLUX, and 5.3 ± 0.8 to 5.1 ± 0.7 with SD3.5. Hence, Stitch does not significantly degrade image quality.

We measure diversity with per-prompt mean pairwise distance between samples in DINOv2 (Oquab et al., 2024) space, finding Stitch preserves or improves average diversity: FLUX increases from 0.34 ± 0.15 to 0.38 ± 0.16; SD3.5 stays at 0.43 ± 0.15 from 0.43 ± 0.14; and Qwen-Image rises from 0.16 ± 0.11 to 0.22 ± 0.10.

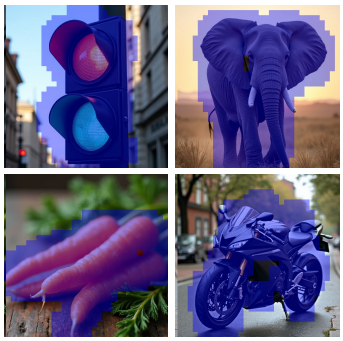


Figure 6: Cutout cleanly extracts objects mid-generation

Table 4: Stitch + FLUX on PosEval across different ηs.

η	**2 Obj**	**3 Obj**	**4 Obj**	**Neg**	**Rel**	**PAB**	Blend
Base	0.22	0.06	0.02	0.62	0.03	0.15	1.00
0.75	0.51	0.30	0.18	0.71	0.31	0.29	0.99
0.80	0.51	0.36	0.22	0.75	0.35	0.35	0.98
0.85	0.59	0.40	0.24	0.80	0.38	0.39	1.00
0.90	0.67	0.42	0.31	0.82	0.47	0.42	1.00
0.95	0.70	0.44	0.38	0.83	0.48	0.44	0.95
0.99	0.79	0.56	0.59	0.89	0.57	0.51	0.80

5.2 Additional Analysis of Cutout

Accuracy of Foreground Extraction. In Appendix A.5, we present each base model's top five heads, ranked by IoU with SAM maps and displayed with each head's optimal threshold μ. We report Intersection over Target (IoT). For each model, several attention heads achieve high IoU, indicating their attention is primarily foreground-focused. IoT is also crucial for Cutout, as it reflects how completely the segmentation captures the object. Together, these highlight which heads are well-suited for Cutout; e.g., FLUX block 14, head 20 achieves an IoU of 62% and an IoT of 92%. We validate this on a second set of 80 objects in Appendix A.9, finding that FLUX's selected head generalizes similarly well to unseen objects.

Qualitative Masks. Example masks are presented in Figure 6 for FLUX's selected Cutout head. We observe that Cutout masks often leave a narrow border around the object, perhaps because low-level details have not yet solidified, helping explain their better IoT compared to IoU. This behavior suits Stitch: the border does not degrade images and helps capture the full object, although even small missing parts can be reconstructed in later generation steps. Additional examples can be found in Appendix 11.

Sensitivity to Head Selection. To assess how sensitive Stitch is to the exact choice of Cutout head, we evaluate Stitch + FLUX on PosEval using each of the top five candidate heads for Cutout, with results shown in Table 12. We find that the top three heads perform similarly well, suggesting that Stitch is not highly sensitive to the exact choice of head and that multiple Cutout heads are suitable.

6 Conclusion

We introduce PosEval, a more challenging position-focused benchmark. PosEval reveals that even the strongest T2I models struggle with position-related tasks, despite performing well on basic ones. This evaluation enables us to propose Stitch, a novel approach to address improper blending in transformer-based models under layout control by using foreground extraction to remove conflicting background information. Stitch significantly boosts Qwen-Image, FLUX, SD3.5, achieving SOTA results while remaining training-free.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *ICLR*, 2023.

Jessica Bader, Leander Girrbach, Stephan Alaniz, and Zeynep Akata. Sub: Benchmarking cbm generalization via synthetic attribute substitutions. *ICCV*, 2025.

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. *ICCV*, 2023.

BlackForest. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Maria A. Bravo. *Advancing vision-language models for open-vocabulary recognition and generative evaluation.* PhD thesis, University of Freiburg, 2025. Chapter 5. Text-Image Concept Human Alignment `https://freidok.uni-freiburg.de/data/269420`.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv:2505.22705*, 2025.

Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv:2411.02395*, 2024a.

Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Controlstyle: Text-driven stylized image generation using diffusion priors. *ACM*, 2023.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *WACV*, 2024b.

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:2501.17811*, 2025b.

Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *ICCV*, 2025c.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *CVPR*, 2022.

Google DeepMind. Imagen 4. https://deepmind.google/models/imagen/, 2025.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *ICML*, 2024.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *NeurIPS*, 2025.

Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Böjrn Ommer, and Nassir Navab. Scenegenie: Scene graph guided diffusion models for image synthesis. *ICCV*, 2023.

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. *NeurIPS*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv:2212.10015*, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *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. *NeurIPS*, 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 . *TPAMI*, 2025.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. Uniedit-flow: Unleashing inversion and editing in the era of flow models. *arXiv:2504.13109*, 2025.

Shyamgopal Karthik, Huseyin Coskun, Zeynep Akata, S. Tulyakov, Jian Ren, and Anil Kag. Scalable ranked preference optimization for text-to-image generation. *ICCV*, 2025.

Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. *NeurIPS*, 2025.

Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. *ECCV*, 2024.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Anna Kukleva, Enis Simsar, Alessio Tonioni, Muhammad Ferjad Naeem, Federico Tombari, Jan Eric Lenssen, and Bernt Schiele. Refam: Attention magnets for zero-shot referral segmentation. *arXiv:2509.22650*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence, 2025.

Yi Li, Haonan Wang, Qixiang Zhang, Boyu Xiao, Chenchang Hu, Hualiang Wang, and Xiaomeng Li. Unieval: Unified holistic evaluation for unified multimodal understanding and generation. *arXiv:2505.10483*, 2025.

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. *CVPR*, 2023.

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. *TMLR*, 2024.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. *CVPR*, 2024.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ECCV*, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*, 2023a.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *NeurIPS*, 2023b.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *CVPR*, 2025.

Midjourney. midjourney v7, 2025. URL `https://github.com/midjourney`.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. *CVPR*, 2024.

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. *AAAI*, 2024.

Marianna Ohanyan, Hayk Manukyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Zero-painter: Training-free layout control for text-to-image synthesis. *CVPR*, 2024.

OpenAI. Dall-e 3. https://cdn.openai.com/papers/dall-e-3.pdf, 2023.

OpenAI. DALL · E 3. https://openai.com/research/dall-e-3, September 2023.

OpenAI. Gpt-4 technical report, 2023.

OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

OpenAI. Gpt-5, 2025. URL `https://openai.com/index/introducing-gpt-5/`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Mateusz Pach, Jessica Bader, Quentin Bouniot, Serge Belongie, and Zeynep Akata. The latent color subspace: Emergent order in high-dimensional chaos. *ICML*, 2026.

Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *CVPR*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

Christoph Schuhmann and Romain Beaumont. Laion-aesthetics. `https://laion.ai/blog/laion-aesthetics/`, 2022.

Marcin Sendera, Lukasz Struski, Kamil Ksiazek, Kryspin Musiol, Jacek Tabor, and Dawid Rymarczyk. Semu: Singular value decomposition for efficient machine unlearning. *arXiv:2502.07587*, 2025.

Shang Hong Sim, Clarence Lee, Alvin Tan, and Cheston Tan. Evaluating the generation of spatial relations in text and image generative models. *arXiv:2411.07664*, 2024.

Jiarui Wang, Huiyu Duan, Yu Zhao, Juntong Wang, Guangtao Zhai, and Xiongkuo Min. Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms. *arXiv:2504.08358*, 2025a.

Zehan Wang, Jiayang Xu, Ziang Zhang, Tianyu Pan, Chao Du, Hengshuang Zhao, and Zhou Zhao. Genspace: Benchmarking spatially-aware image generation. *arXiv:2505.24870*, 2025b.

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:2508.02324*, 2025a.

Chengyue Wu, Xi aokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *CVPR*, 2025b.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv:2506.18871*, 2025c.

Size Wu, Zhonghua Wu, Zerui Gong, Qingyi Tao, Sheng Jin, Qinyue Li, Wei Li, and Chen Change Loy. Openuni: A simple baseline for unified multimodal understanding and generation. *arXiv:2505.23661*, 2025d.

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. *ICCV*, 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. *ICML*, 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. *arXiv:2412.03859*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *ICCV*, 2023a.

Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv:2305.18583*, 2023b.

Eric Zhou and Dokyun Lee. Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *NeurIPS*, 2024.

A Appendix

A.1 Stitch Results on Existing Benchmarks

Table 5: Results on tasks from existing benchmarks

Model	GenEval	T2ICompBench	HRS-Bench (spatial)		
	Position	Spatial	Easy	Medium	Hard
SD3.5 Large	0.34	0.22	0.46	0.19	0.08
Stitch (ours) + SD3.5 Large	0.53	0.28	0.54	0.26	0.08
Gain over SD3.5 Large	+0.19	+0.06	+0.08	+0.07	+0.00
FLUX.1 [Dev]	0.22	0.25	0.50	0.23	0.08
Stitch (ours) + FLUX.1 [Dev]	0.70	0.44	0.82	0.52	0.23
Gain over FLUX.1 [Dev]	+0.48	+0.19	+0.32	+0.29	+0.15
Qwen-Image	0.76	0.36	0.73	0.45	0.25
Stitch (ours) + Qwen-Image	0.85	0.50	0.88	0.63	0.40
Gain over Qwen-Image	+0.09	+0.14	+0.15	+0.18	+0.15

We present results on several *Position*-related tasks in existing benchmarks: GenEval's *Position* task (Ghosh et al., 2023), T2I CompBench's *Spatial* task (Huang et al., 2023) (referred to as *2D-Spatial* in T2ICompBench++ (Huang et al., 2025)), and HRS-Bench's *Spatial* task (Bakr et al., 2023). T2I CompBench's *Spatial* task includes two objects and a single relation. HRS Bench's *Spatial* task is divided into three sub-categories: *easy* prompts include two objects and one relation; *medium* prompts include three objects and one or two relations; and *hard* prompts include four objects and one or two relations.

With Stitch, we improve all three base models across all tasks, except HRS-Bench *Hard* with SD3.5, where performance remains stable. This indicates that Stitch consistently enhances positional information across *Position*-related tasks from multiple benchmarks.

A.2 PosEval User Study Results

We examine the accuracy of the evaluation system used in PosEval, which directly inherits the framework introduced in GenEval. This system employs Mask2Former (Cheng et al., 2022), trained on MS COCO (Lin et al., 2014), to detect objects. The detected objects are then programmatically analyzed to assess whether the prompt has been correctly generated—for example, by verifying the presence of required objects and the relative positions of their bounding boxes. Color attributes are classified using CLIP (Radford et al., 2021).

Extending this evaluation to PosEval requires only minor modifications to the logical comparisons between detections and prompts, without altering the underlying detection mechanisms that primarily drive failures. To ensure that these adjustments do not meaningfully degrade accuracy on the new tasks, we conduct a user study. For this study, we create a candidate image pool with 4 images per prompt from Qwen-Image, Stitch + Qwen-Image, FLUX, Stitch + FLUX, SD3.5, and Stitch + SD3.5. We divide this pool into images labeled as *correct* and *incorrect* by the automated evaluation, then randomly select 50 images from each group. Three annotators independently evaluate whether each image "correctly follows the prompt", and we report their level of agreement with the automatic evaluation.

Table 6 shows that the evaluation protocol remains consistent, with all PosEval categories falling within 10% of the average inter-annotator agreement of 91%. Using the two-object task from GenEval as a baseline, we observe that accuracy on the new tasks does not significantly decline, with only minor fluctuations—particularly on more challenging tasks such as PAB, as expected.

Table 6: PosEval alignment with human annotators. Average inter-annotator alignment is 91%.

2 Obj	**3 Obj**	**4 Obj**	**Neg**	**Rel**	**PAB**
87.7%	90.7%	89.7%	84.7%	86.3%	81.3%

A.3 Additional Qualitative Examples

Figure 7 presents qualitative examples for Stitch combined with SD3.5, showcasing two examples from each PosEval benchmark task. Additional qualitative results are provided for Stitch with FLUX in Figure 9, and for Stitch with Qwen-Image in Figure 8, also on two prompts per task.

Figure 7: Qualitative examples for Stitch + SD3.5.

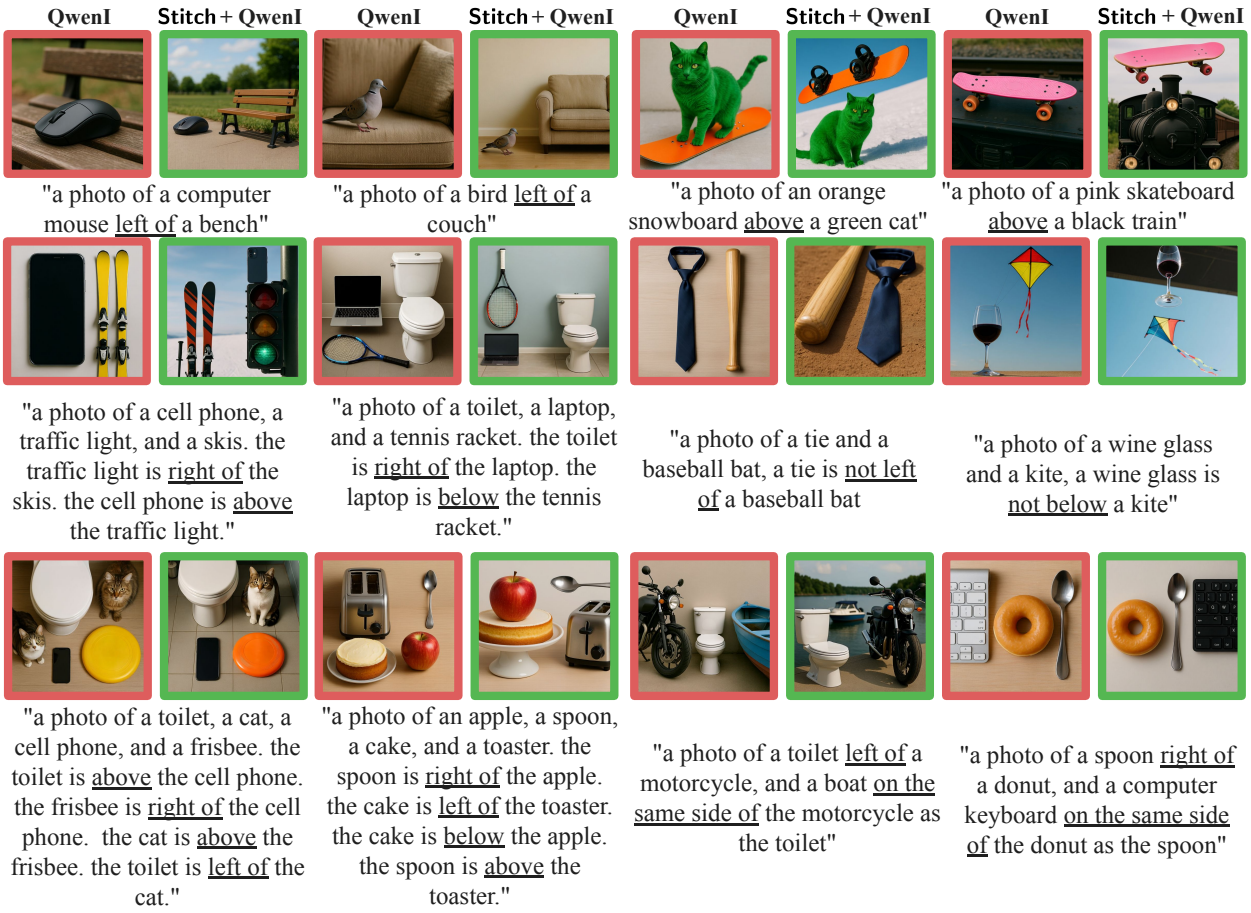


Figure 8: Additional qualitative examples for Stitch + Qwen-Image (QwenI).

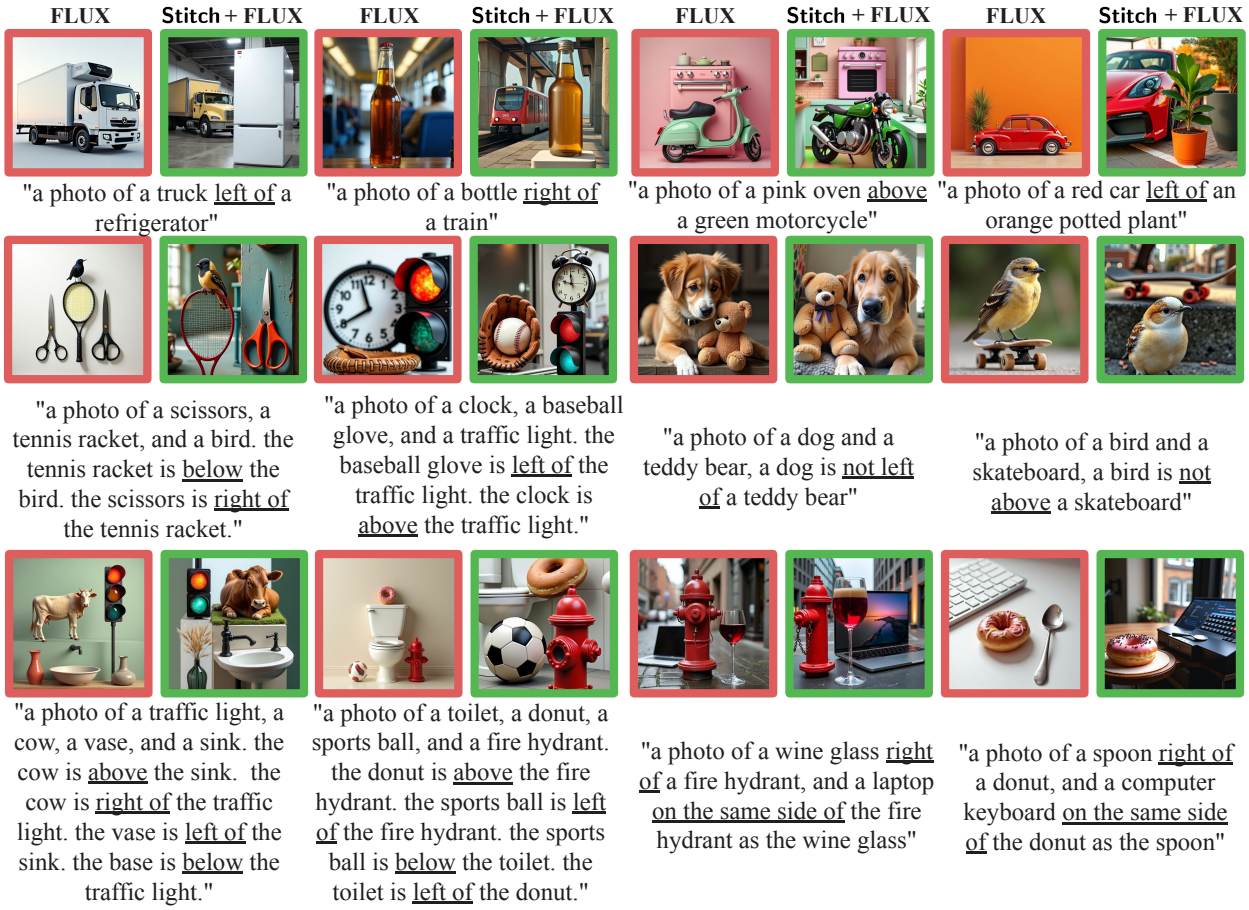


Figure 9: Additional qualitative examples for Stitch + FLUX.

A.4 Comparison of PosEval with other Benchmarks

We include an explicit comparison between PosEval and previously existing benchmark tasks which include position-based information in Table 7. We highlight several key aspects, focused on what is included in the benchmark and how specifically it is evaluated. First, we look at the number of objects present in positional tasks, starting with if they distinguish by the number of objects (*Dist. # Obj.*). For some benchmarks, such as the NSR-1k natural task (Feng et al., 2023), benchmark settings are not explicit about the number of objects present in each image. While all included benchmarks include 2-object prompts, we also specify if they include 3-object (*3 Obj.*) and 4-object (*4 Obj.*) prompts. Similar to *Dist. # Obj.*, we indicate if the tasks are generally focused (*Focused tasks*), meaning that they include settings which separate the positional problem into specific sub-tasks. An example of non-focused tasks is DPG (Hu et al., 2024) which targets the understanding of long prompts, but as a result does not sub-divide if prompts also test attribute binding, what types of positions are used (e.g. 2D vs. 3D), etc. In *Attr.*, we indicate if attributes (e.g. color) are present. Finally, *Neg. Pos.* shows the presence of *negative* in the sense of *position* (e.g. being "not left").

First, we observe that no existing benchmark incorporates the concept of *negative position*. Additionally, none of the prior benchmarks simultaneously include both attributes and focused tasks. However, focused tasks are essential for thoroughly evaluating and advancing positional generation. They enable developers to pinpoint where models fail and clearly compare capabilities. This is a key advantage of our PosEval, which addresses a gap left by previous benchmarks. In particular, our *Positional Attribute Binding* task introduces a uniquely challenging and previously unexplored dimension. Our *Relative Relation* task is also the first to specifically explore these types of more challenging *relative* relations, in an isolated and focused way. To ensure consistency with prior work, we also scale the number of objects to four. Overall, PosEval fills a critical gap in the literature by offering focused positional tasks while appropriately scaling the difficulty.

Table 7: PosEval comparison with other existing benchmarks

	Dist. # Obj.	3 Obj.	4 Obj.	Focused tasks	Attr.	Neg. Pos.
HRS-Bench (Bakr et al., 2023)	✓	✓	✓	✓	✗	✗
NSR-1K - template (Feng et al., 2023)	✓	✗	✗	✓	✗	✗
NSR-1K - natural (Feng et al., 2023)	✗	✓	✓	✗	✓	✗
VISOR (Gokhale et al., 2022)	✓	✗	✗	✓	✗	✗
DPG (Hu et al., 2024)	✗	✓	✓	✗	✓	✗
T2I CompBench++ (Huang et al., 2025)	✓	✗	✗	✓	✗	✗
GenEval (Ghosh et al., 2023)	✓	✗	✗	✓	✗	✗
PosEval (ours)	✓	✓	✓	✓	✓	✓

A.5 Top Cutout Segmentation Heads for each Model

We present the results of the top attention heads for FLUX (Table 8), Qwen-Image (Table 10), and SD3.5 (Table 9). The bold results are those selected for Cutout for each model. This is block 14 head 20 for FLUX, block 14 head 34 for SD3.5, and block 25 head 1 for Qwen-Image. Interestingly, we find that the best segmentation heads typically fall in the middle blocks (between 10 and 20 for FLUX and SD3.5, and between 15 and 30 for Qwen-Image, which has more blocks).

We include additional visualizations of the Cutout masks in FLUX in Figure 11. For six example objects, we present: (1) the final image, (2) the attention map from the Cutout head (block 14, head 20 in FLUX) at step $S = 10$, and (3) the final Cutout masks thresholded at 95%. By comparing the attention maps to the final images, we observe that the selected head already encodes information about the object's location in the latent space, even in the early stages of generation. Furthermore, the Cutout masks accurately capture the object's position, with a slight margin from the 95% threshold. This buffer proves beneficial to Stitch, as confirmed by our ablation study.

In Figure 10, we also include segmentation maps for SD3.5, similar to those provided for FLUX in the main paper. These are compared to the SAM maps, for reference.

Table 8: Best segmentation heads for FLUX.

Block	Head	η	IoU	IoT
14	**20**	0.75	**0.62**	**0.92**
17	23	0.75	0.56	0.82
13	21	0.75	0.54	0.87
14	15	0.75	0.54	0.87
10	1	0.75	0.53	0.86

Table 9: Best segmentation heads for SD3.5.

Block	Head	η	IOU	IOT
14	**34**	0.75	**0.56**	**0.94**
17	13	0.75	0.56	0.96
15	31	0.75	0.55	0.94
18	36	0.75	0.55	0.96
14	6	0.80	0.55	0.90

Table 10: Best segmentation heads for Qwen-Image.

Block	Head	η	IOU	IOT
25	**1**	**0.9**	**0.55**	**0.86**
27	17	0.97	0.53	0.89
26	4	0.85	0.53	0.82
19	12	0.97	0.53	0.90
18	11	0.85	0.52	0.84

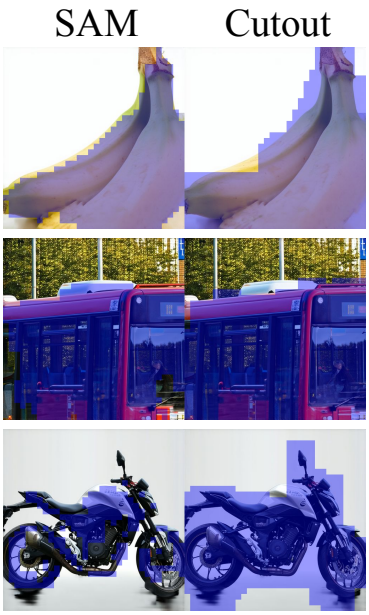


Figure 10: Segmentation head maps for SD3.5.

Figure 11: Attention weights and Cutout masks from head 20 in block 14 of FLUX, at step 10 of 50.

A.6 Bounding Boxes and Sub-prompts Generation

To generate object bounding boxes b_k and prompts p_k, $\forall k > 0$, we utilize GPT-5 (OpenAI, 2025) model with reasoning effort set to minimal and system prompt:

```
You have a canvas of size {W}x{W}. Decompose the given description
into single objects. Do not merge multiple objects into one box.
Fully cover the canvas with the object boxes. Avoid overlaps.
Do not leave space for background. Write 3 sentence justification
and then output just valid JSON in the format:
[
    {"prompt": "<object with properties only mentioned in the
    description>", "x_min": , "y_min": , "x_max": , "y_max": },
    ...
]
```

Each prompt P is processed with user prompt:

```
Description: {P}
```

To generate background prompt p_0 we use GPT-4 (OpenAI, 2023) with system prompt:

```
Provide a simple fitting background description for this scene.
The background must not mention any of the specific objects or
elements from the prompt. The background must not mention any
other specific objects or people. Return only ONE word of the
background text, nothing else.
```

Given a prompt P the user prompt again takes form:

```
Description: {P}
```

A.7 User Study Interface

In Figures 12 and 13, we include and example of the user interface for our user studies to measure PosEval's quality and *Blend* of the images.

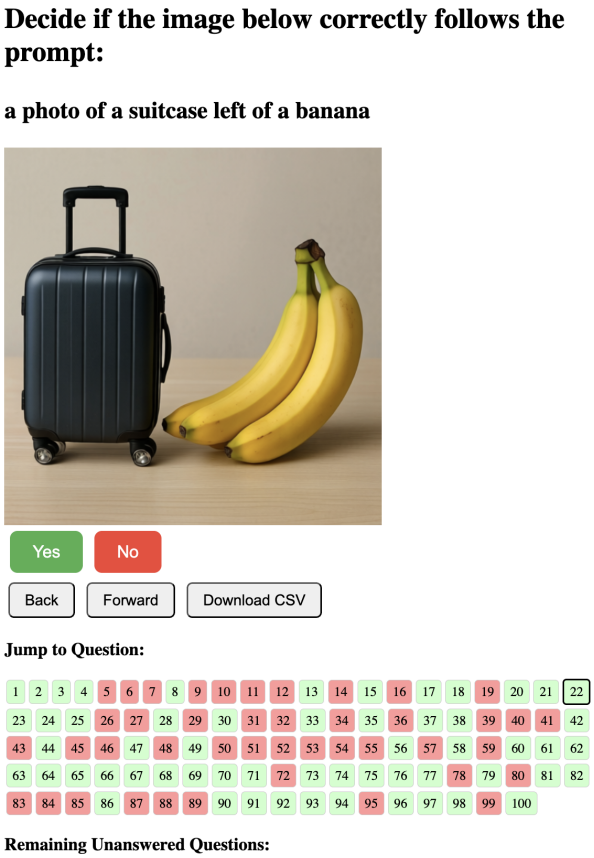


Figure 12: *PosEval* user study.

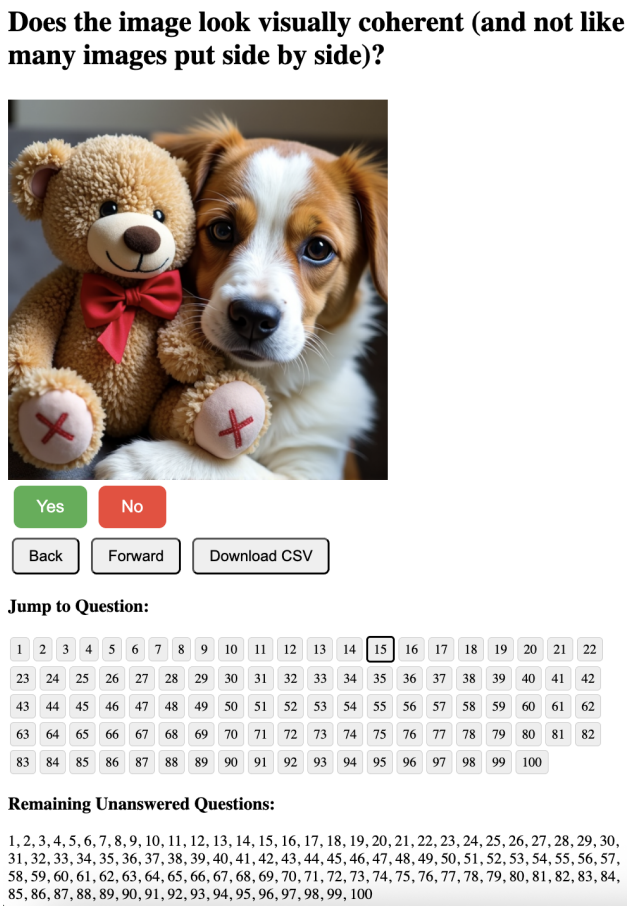
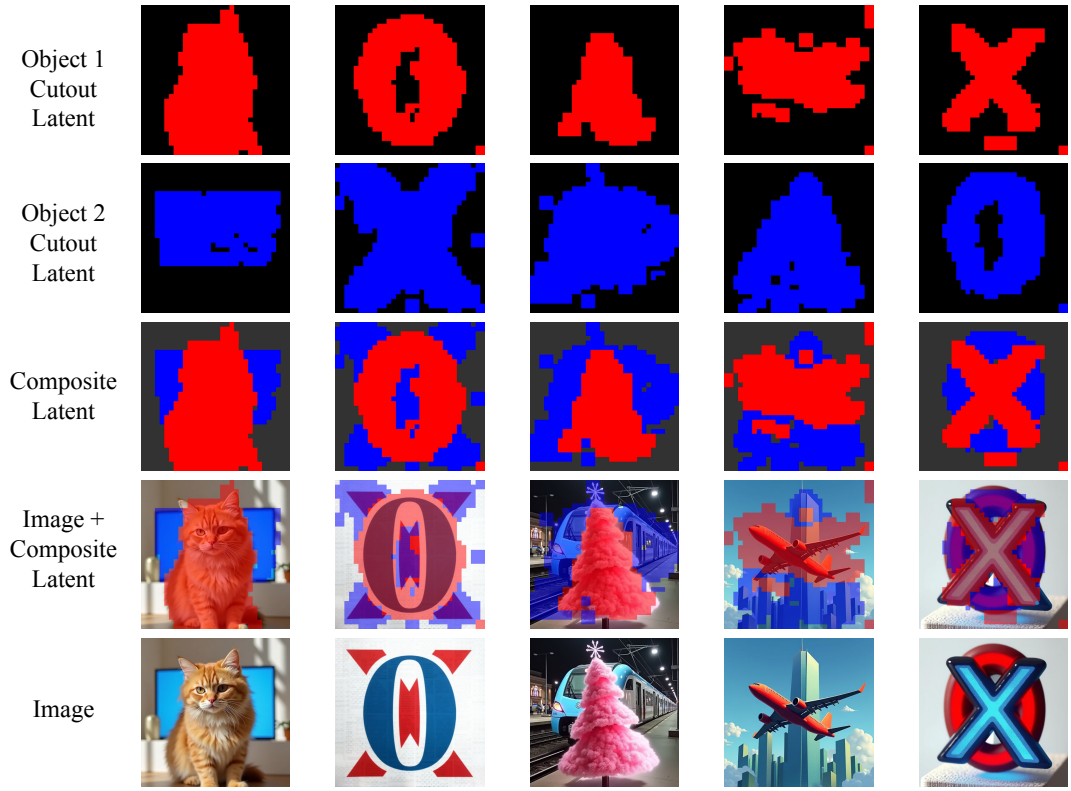


Figure 13: *Blend* user study.

A.8 Overlapping Regions

In Figure 14, we illustrate how Cutout enables Stitch to handle overlapping bounding boxes, allowing it to generate objects *in front of* or *behind* one another. This also equips Stitch with the flexibility to handle various MLLM-generated bounding boxes, including those that may overlap.



Main Prompt	a cat in front of a tv	letter O overlapping letter X	a christmas tree in front of a train	a plane in front of a skyscraper	letter X overlapping letter O
Object 1 Sub-prompt	a small sitting orange cat	thin red capital letter O	a pink christmas tree	a long red plane in the sky	thin blue capital letter X
Object 2 Sub-prompt	a tv with blue screen	thin blue capital letter X	a blue train standing at the station	a tall green skyscraper	thin red capital letter O
Background Sub-prompt	empty room	white wall	a station	sky	white wall

Figure 14: Stitch's use of Cutout enables it to handle overlapping bounding boxes. In an extreme setting, we task Stitch with generating pairs of objects that share identical bounding boxes spanning the entire image. By stacking the Cutout object latents in a controlled order (allowing partial occlusion), Stitch can render both objects while preserving the correct relational structure. Notably, the generated objects appear within regions specified by the cutout masks. Moreover, objects can retain underlined attributes from their sub-prompts even when those attributes are absent from the main prompt, providing further evidence that Stitch successfully merges overlapping sub-generations. This is not achievable with earlier methods operating strictly on disjoint rectangular bounding boxes.

A.9 Validation of Cutout Head Selection

To validate the selected Cutout head's IoU, we generate a second set of 80 objects. To ensure broad and balanced coverage of real-world concepts, we define eight high-level categories and sample ten representative words from each, curated by ChatGPT (OpenAI, 2025) and non-overlapping with the GenEval objects. Prompts are made as, "A photo of a {obj}". Each prompt is repeated for 3 seeds.

Table 11: Validation of FLUX's best Cutout head performed on 80 prompts across 10 categories. Results are consistent across categories and closely match those observed during head selection.

	IoU	IoT
Animals	0.72	0.91
Clothes	0.71	0.87
Electronics	0.58	0.90
Food	0.65	0.93
Household	0.66	0.90
Nature	0.59	0.89
Plants	0.56	0.89
Transportation	0.55	0.89
All	$0.63_{\pm 0.06}$	$0.90_{\pm 0.02}$

We evaluate our selected FLUX head (Block 14, Head 20) using its best-performing value of $\eta = 0.75$ on the new prompts. The results, shown in Table 11, yield 0.63 IoU and 0.90 IoT, closely matching the metrics observed during head selection (0.62 IoU and 0.92 IoT). The standard deviation across categories is also small, with the IoT being particularly consistent at only ± 0.02.

The words chosen are:

Animals: hamster, monkey, turtle, lion, tiger, kangaroo, panda, gorilla, hippopotamus, penguin
Clothes: shirt, jacket, coat, dress, skirt, jeans, sock, shoe, hat, scarf
Food: strawberry, grapes, potato, tomato, loaf of bread, slice of cheese, milk, egg, yogurt, pasta
Household Objects: pillow, blanket, mug, coffee maker, trash, toy, candle, basket, kettle, plate
Nature: rock, leaf, branch, shell, feather, pinecone, volcano, mountain, acorn, iceberg
Plants: sunflower, cactus, pine, bamboo, oak tree, wheat, rose, daffodil, tomato plant, tulip
Technology: tablet, smartwatch, computer mouse, earbuds, camera, printer, camcorder, scanner, gaming console, speaker
Transportation: van, tram, helicopter, sailboat, cruise ship, tractor, scooter, jet, hot air balloon, roller blades

A.10 Impact of Cutout Head Selection on Stitch

To better understand how the chosen attention head for Cutout influences Stitch, we examine three questions: (1) what the distribution of IoU scores looks like across attention heads, (2) how well our proposed head-scoring method correlates with Stitch performance, and (3) how sensitive Stitch is to the specific Cutout head used. Our results show that although many heads are unsuitable for Cutout, several can achieve similarly high performance, and that our proposed SAM IoU-based method can effectively select these heads. While selecting the Cutout head with our IoU method only achieves most of the optimal performance, we also propose a validation-based method that can further boost performance.

Distribution of IoU scores: We analyze the distribution of IoU values across all FLUX heads, shown in the histogram in Figure 15. Most heads achieve similarly low IoU, with only a small number standing out as above average. This is intuitive: only heads that meaningfully attend to the foreground object are expected to yield higher IoU, and relatively few heads in the model appear to play this role.

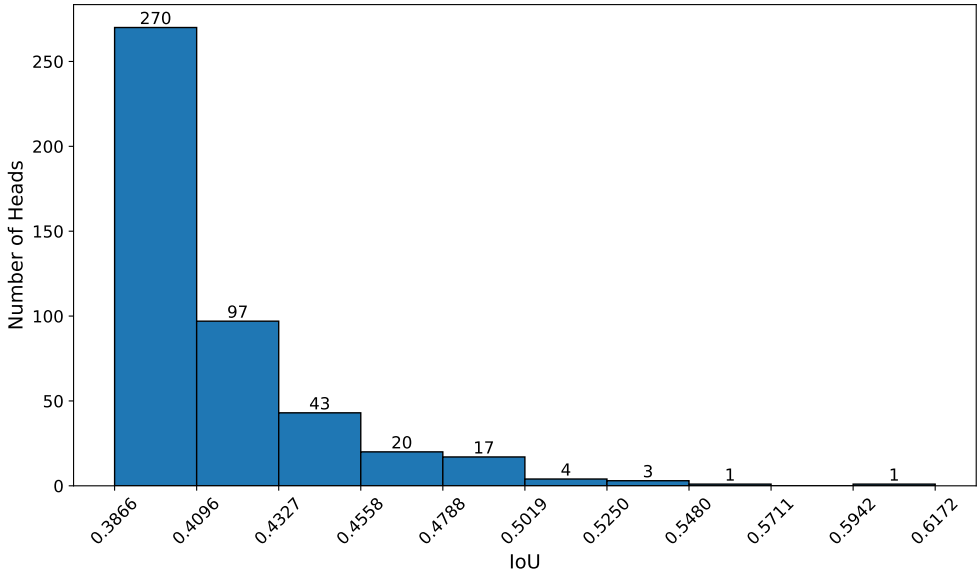


Figure 15: Histogram of the IoU of all FLUX heads when used for Cutout. Multiple attention heads are suitable for Cutout, but most are not.

Correlation with Stitch performance: Next, we verify that the IoU between attention maps and SAM masks is meaningfully correlated with Stitch's performance, specifically with *Blend*. We evaluate Stitch + FLUX on PosEval using Cutout derived from several different attention heads. Testing all 19×24 attention heads in FLUX would be infeasible, so we select a representative subset. To do this, we first rank all attention heads by their highest IoU with SAM masks across $\eta \in 0.75, 0.80, 0.85, 0.90, 0.95, 0.97, 0.99$. From this ranking, we select the top five heads as alternatives to the best-performing head used in Stitch. For additional comparison, we also include the three heads with median performance and the three lowest-performing heads.

Table 12 shows that selecting heads based on their IoU with SAM masks is effective: only the top heads achieve both high accuracy and high *Blend*. Heads with low IoU do not reliably focus on foreground objects. As a result, they may either retain enough of the object to preserve accuracy while keeping too much background and hurting *Blend*, or they may remove too much of the object, reducing accuracy altogether (e.g., Block 13 Head 10).

Sensitivity of Stitch to Cutout head: Next, we examine how sensitive Stitch is to the specific Cutout head used. The results in the *Middle* and *Bottom Head* groups show that randomly choosing a head is

Table 12: Stitch + FLUX performance on PosEval for by Cutout head. Heads are ranked by their best IoU across a range of thresholds, and we evaluate the top five, middle three, and bottom three. Only the top heads achieve both high accuracy and high *Blend*, demonstrating that our selection strategy is closely aligned with Stitch's performance. Several top heads perform similarly, meaning Stitch's effectiveness does not hinge on a single head. Red highlights poor performance.

Rank	Block	Head	IoU	**2 Obj**	**3 Obj**	**4 Obj**	**Neg**	**Rel**	**PAB**	Blend
Without Cutout										
				0.81	0.58	0.52	0.90	0.63	0.51	0.68
Top Heads										
1	14	20	0.62	0.70	0.44	0.38	0.83	0.48	0.44	0.95
2	17	23	0.56	0.69	0.47	0.41	0.84	0.51	0.47	0.95
3	13	21	0.54	0.72	0.51	0.48	0.86	0.57	0.51	0.98
4	14	15	0.54	0.74	0.52	0.40	0.86	0.53	0.49	0.88
5	10	1	0.53	0.79	0.51	0.50	0.86	0.57	0.47	0.90
Middle Heads										
227	15	19	0.40	0.74	0.52	0.46	0.84	0.53	0.47	0.76
228	12	14	0.40	0.77	0.56	0.52	0.87	0.58	0.51	0.73
229	6	21	0.40	0.76	0.56	0.52	0.88	0.52	0.51	0.83
Bottom Heads										
454	3	17	0.39	0.78	0.52	0.51	0.87	0.55	0.48	0.68
455	1	11	0.39	0.70	0.48	0.51	0.87	0.36	0.41	0.83
456	13	10	0.39	0.34	0.18	0.13	0.62	0.20	0.15	0.90

ineffective, as most heads are not well suited for this task. However, within the *Top Heads* group, several options perform similarly. In particular, the three highest-ranked heads (Block 14 Head 20, Block 17 Head 23, and Block 13 Head 21) achieve comparable accuracy and *Blend* scores.

Advanced head selection strategy Because both the SAM masks used for head selection and the generation process itself introduce noise, the optimal head may lie within the top k candidates rather than being the single highest-ranked one. Indeed, we observe that the third-ranked head, Block 13 Head 21, performs slightly better than the top two. If marginally higher accuracy is desired, a small validation set of Stitch prompts can be used to choose the best head from among the top k. Nevertheless, we emphasize that selecting just the top head ($k = 1$) already yields strong performance. As Figure 15 shows that very few heads achieve high IoU, even a very small k is sufficient.

A.11 Alternatives to head selection based on SAM

To select a head for Cutout, we primarily recommend comparing the potential Cutout masks from each attention head with SAM maps, as described in Section 5.2. However, there are also several potential methods for selecting a head that do not require any external models.

Figure 16: Masks generated by classifying each pixel based on whether its color is closer to that of the central pixel or to the average color of pixels along the side edges. With enough prompts featuring objects in diverse and clearly distinct colors, this method could serve as an alternative to SAM for finding ground truth to select the appropriate cutout head.

The first alternative is most similar to our original method, but the masks are generated without an external model. Images are first produced with distinct colors differentiating foreground and background objects, as illustrated in the top row of Figure 16. Masks are then created using a simple heuristic: each pixel is classified based on whether its color is closer to the central pixel (assumed to belong to the foreground) or to the average color of the side edge pixels. The bottom row of Figure 16 illustrates examples of these masks. Masks generated in this way replace the SAM masks described in Section 5.2, while the remainder of the selection process remains unchanged.

Our second head selection alternative takes a different approach. We begin with the assumption that a head focusing on foreground objects will primarily attend to the center of the image. To capture this, we model the expected average attention of a foreground-focused head as a 2D Gaussian centered in the latent space, with width $\sigma = \frac{L_s}{6}$, where $L_s = 32$ is the latent width. For each attention head, we compute its average attention map over the same 80 prompts used in our original selection strategy based on GenEval objects. We then compare these average attention maps with the described Gaussian using Jensen-Shannon (JS) divergence. Ranking heads by JS divergence reveals that those highly ranked by our original metric are also ranked highly by this method, suggesting that JS divergence could serve as a practical, model-free alternative to IoU. Figure 17 shows a histogram of the JS divergences of FLUX's attention heads, along with examples of the average attention maps for selected heads at different JS divergence values.

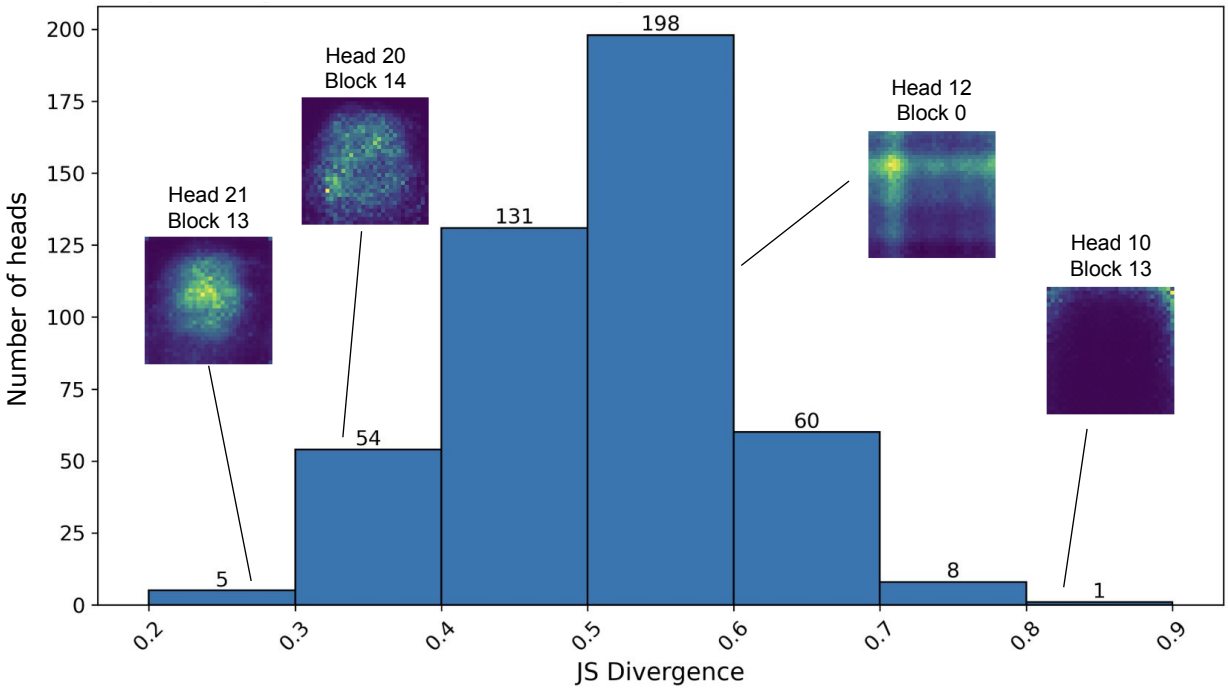


Figure 17: A histogram of Jensen-Shannon (JS) divergences between each head's average attention-map pattern and a centered 2D Gaussian with $\sigma \approx L_S/6$, where L_s is the attention-map size. Heads with the lowest divergences (e.g. Head 21 in Block 13 and Head 20 in Block 14) closely match the expected Gaussian-like pattern and correspond to the top-performing heads based on IoU with SAM-derived ground truths. In contrast, heads with the highest divergences (e.g. Head 12 in Block 0 and Head 10 in Block 13) display off-center attention, indicating that they fail to concentrate on the foreground object.

A.12 Semantic Leakage in Other Methods

RAG (Chen et al., 2025c) and Regional Prompting (Chen et al., 2024a) similarly introduce test-time modifications to improve the blending of sub-prompts for layout-constrained generation in transformer-based models. However, these methods rely on averaging sub-prompt latents, whereas Stitch instead extracts foreground objects from their backgrounds to enable smoother merging.

As shown in Table 2, Stitch outperforms these methods on PosEval. To investigate why, we present several qualitative examples in Figure 19. We observe that both Regional Prompting and RAG are more prone to merging multiple objects (e.g., "a photo of an elephant below a horse") or omitting objects entirely (e.g., "a photo of a bus above a boat"), which may explain their lower performance.

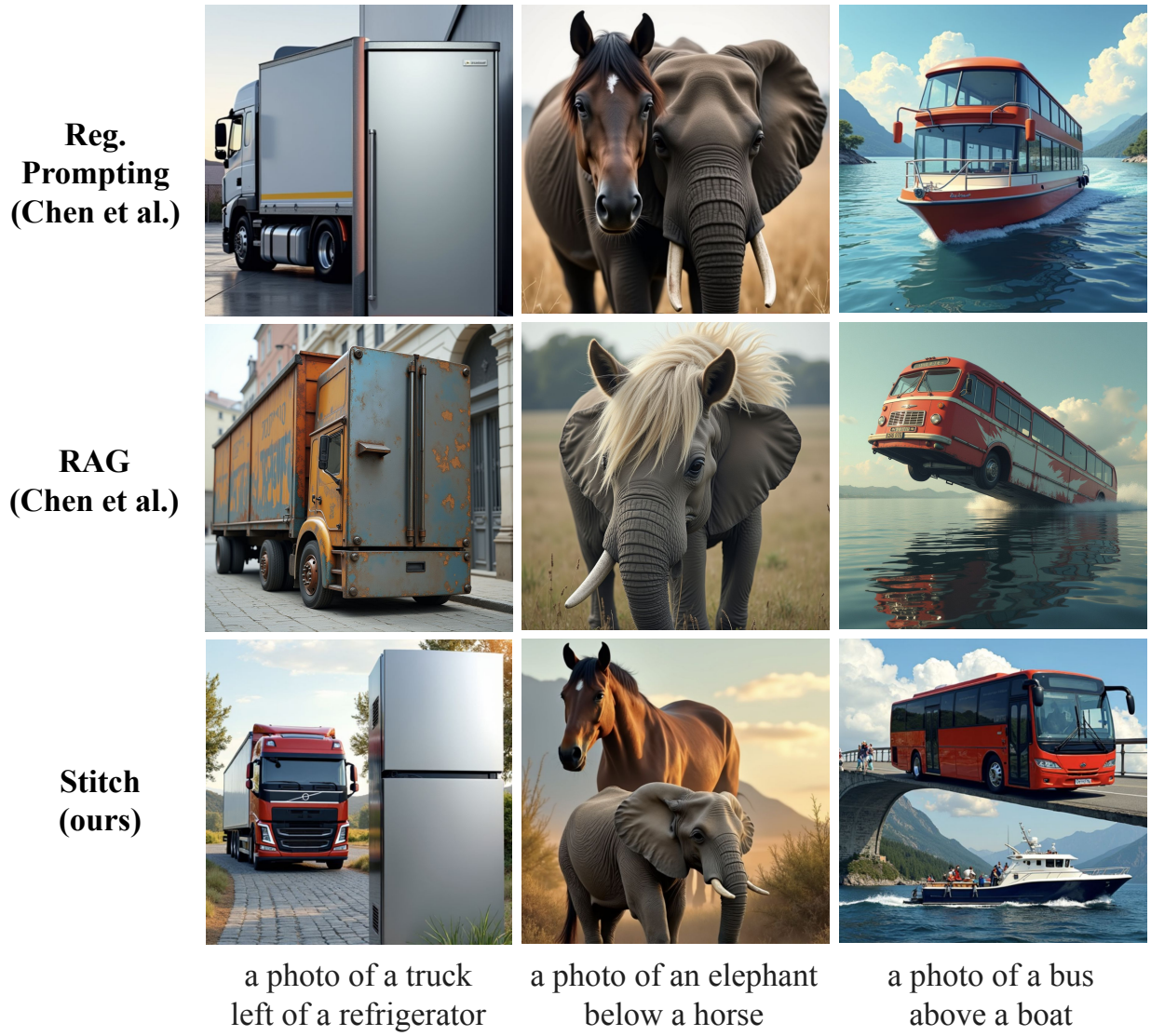


Figure 18: Reg. Prompting (Chen et al., 2024a) and RAG (Chen et al., 2025c) achieve lower scores than Stitch in Table 2, partly because their strategies for avoiding blending artifacts while constraining generation can lead to semantic leakage between regions. We illustrate such behavior across three different prompts, with all methods applied to FLUX.

A.13 Sensitivity of Stitch to the LLM-generated Bounding Boxes

Because Stitch relies heavily on MLLM-generated bounding boxes, we evaluate its robustness to variations in these boxes. To do so, we conduct two experiments: perturbing the bounding boxes and shuffling them. We compare Stitch 's performance under these modifications to its performance with the original boxes, finding that Stitch is robust to variations in the bounding boxes, but strongly reliant on their accuracy. We use FLUX as the base model and evaluate on PosEval.

In the perturbation experiment, we examine how sensitive Stitch is to small, relation-preserving fluctuations in the bounding boxes. We perturb the boxes from the Original evaluation by randomly shifting their edges by up to 10% of their initial width, ensuring that the underlying spatial relations remain correct. As shown in Table 13, performance under perturbation closely matches that of the original boxes, indicating that precise box placement is not required—small deviations do not meaningfully impact accuracy.

In the shuffling experiment, we assess how Stitch responds to incorrect bounding boxes. We keep the box locations fixed but randomly reassign the sub-prompts among them, producing many incorrect layouts. As shown in Table 13, this substantially degrades performance, indicating that Stitch relies strongly on the information encoded in the bounding boxes and is sensitive to incorrect assignments. Fortunately, LLMs are reliable at producing layout-aware sub-prompts that preserve these relations.

Table 13: Stitch + FLUX's PosEval performance under three conditions: the *Original* bounding boxes, *Perturbed* bounding boxes (always preserving correctness) and *Shuffled* sub-prompt/bounding box assignment (sometimes breaking correctness). Relative to the *Original* bounding boxes, Stitch maintains high accuracy with *Perturbed* bounding boxes, indicating robustness to small variations. In contrast, shuffling significantly impacts performance, showing that Stitch prioritizes the information encoded in the bounding boxes.

	2 Obj	**3 Obj**	**4 Obj**	**Neg**	**Rel**	**PAB**
Original	0.70	0.44	0.38	0.83	0.48	0.44
Perturbed	0.71	0.42	0.40	0.82	0.45	0.44
Shuffled	0.39	0.07	0.02	0.47	0.12	0.25

We also explore qualitatively how Stitch behaves when the bounding boxes are incorrect. We identify one failure case for "a photo of a dining table above a suitcase", shown in Figure 19: the MLLM mistakenly includes both objects in both sub-prompts instead of assigning one object to each. As a result, the table appears twice in the final image—once in the top region and once in the bottom.

Figure 19: Erroneous MLLM-generated bounding boxes place both objects in each sub-prompt (right), resulting in the table being rendered twice (left).