STITCH: TRAINING-FREE POSITION CONTROL IN MULTIMODAL DIFFUSION TRANSFORMERS

Jessica Bader[1,2,3,*] **Mateusz Pach**[1,2,3,*]
María A. Bravo[1,2,3] **Serge Belongie**[4] **Zeynep Akata**[1,2,3]
[1]Technical University of Munich [2]Helmholtz Munich [3]Munich Center for Machine Learning
[4]University of Copenhagen *equal contribution jessica.bader@tum.de

ABSTRACT

Text-to-Image (T2I) generation models have advanced rapidly in recent years, but accurately capturing spatial relationships like "above" or "to the right of" poses a persistent challenge. Earlier methods improved spatial relationship following with external position control. However, as architectures evolved to enhance image quality, these techniques became incompatible with modern models. We propose Stitch, a training-free method for incorporating external position control into Multi-Modal Diffusion Transformers (MMDiT) via automatically-generated bounding boxes. Stitch produces images that are both spatially accurate and visually appealing by generating individual objects within designated bounding boxes and seamlessly stitching them together. We find that targeted attention heads capture the information necessary to isolate and cut out individual objects mid-generation, without needing to fully complete the image. We evaluate Stitch on PosEval, our benchmark for position-based T2I generation. Featuring five new tasks that extend the concept of *Position* beyond the basic GenEval task, PosEval demonstrates that even top models still have significant room for improvement in position-based generation. 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 state-of-the-art results with Qwen-Image on PosEval, improving over previous models by 54%, all accomplished while integrating position control into leading models training-free. Code is available at https://github.com/ExplainableML/Stitch.

1 INTRODUCTION

Text-to-Image (T2I) generation models provide a powerful bridge between natural language and visual content. By transforming textual descriptions into images, they expand human creativity, accelerate prototyping, and enable a wide range of applications. But as their popularity grows, so do user demands. These models should be able to generate real-world scenes with uncommon but still physically plausible object arrangements. They should interpret and fulfill requests that are both complex and linguistically nuanced in their descriptions of positioning. All of these needs could be met with strong position understanding, yet even basic spatial concepts such as "above" or "to the right of" remain a persistent challenge for current models (Ghosh et al., 2023), and the difficulty only grows with more complex prompts. Luckily, positional performance can be improved by augmenting off-the-shelf models with bounding boxes generated by Large Language Models (LLMs), but recent architectural upgrades have made traditional position-control methods incompatible with modern models. Nevertheless, these modifications have also enhanced both image quality and speed, traits that could be even more impactful if they were paired with precise positional prompt-following.

To bridge this gap, we propose Stitch: a test-time technique that effectively improves positional prompt generation (Figure 1a) by integrating additional position support via LLM-generated bounding boxes into leading Flow Matching (FM) T2I models based on Multi-Modal Diffusion Transformer (MMDiT) architecture. Stitch achieves controlled generation by independently generating objects for S steps, with each object constrained to its respective bounding box through attention modulation. Next, we extract foreground objects and combine the predictions. In particular, we find

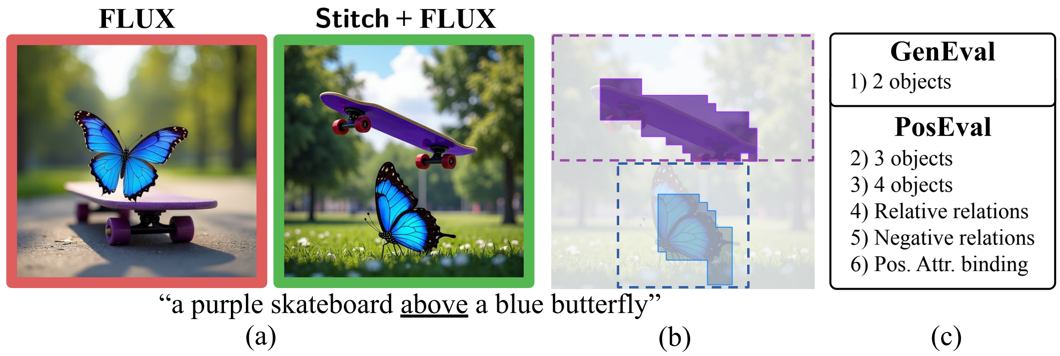


Figure 1: (a) **Stitch** boosts position-aware generation, training-free, (b) by generating objects in LLM-made bounding boxes (dashed lines) and using attention heads for tighter latent segmentation mid-generation (filled). (c) Our **PosEval** benchmark extends GenEval with 5 new positional tasks.

that foreground masks can be inexpensively derived directly from the attention heads (Figure 1b), on-the-fly before completing generation and without requiring an external model. After step S all constraints are lifted, enabling the T2I model to refine the image organically in the remaining steps. Consequently, Stitch facilitates quick and affordable upgrades in the positional performance of the leading T2I models, combining the best image quality with strong positional generation (Figure 2).

Such positional upgrades are extremely valuable for improving 2D spatial performance. Because while recent successes on *Position* tasks in existing benchmarks such as GenEval (Ghosh et al., 2023) might suggest that the problem is nearing resolution, a closer analysis shows that it is far from solved. We take the next step in evaluating whether T2I models can consistently and robustly generate images that accurately reflect positional prompts. To this end, we introduce PosEval, an extension of the GenEval (Ghosh et al., 2023) benchmark designed for in-depth evaluation of positional abilities in T2I generation, going beyond the traditional *Position* category (Figure 1c). PosEval includes five new tasks, each aimed at probing specific failure modes in T2I models. Using PosEval, we demonstrate that state-of-the-art (SOTA) models continue to struggle with positional tasks, highlighting considerable room for improvement. PosEval additionally provides a comprehensive evaluation platform that rigorously measures Stitch's effectiveness in improving positional generation, showcasing significant performance improvements compared to baseline models.

The primary contributions of this work are: (1) We introduce Stitch, a test-time method that substantially improves MMDiT-based models' capacity to accurately generate images from position-based prompts; (2) We find that certain attention heads encode sufficient information to extract the foreground object from the background within the latent space, long before the image is fully generated; and (3) We present the PosEval benchmark, a GenEval (Ghosh et al., 2023) extension featuring five new targeted, position-focused tasks designed to tackle the next level of generation challenges.

2 Related Works

T2I generation. T2I synthesis has rapidly advanced in the 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 more 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 have improved pre-trained T2I models by enhancing prompt adherence (Eyring et al., 2024; 2025), alignment with human preferences (Karthik et al., 2025; Liang et al., 2024; Bravo, 2025; Sendera et al., 2025), and personalization (Liu et al., 2023b; Ruiz et al., 2022; Kim et al., 2024). Despite these significant advances, positional control remains underexplored, and T2I models still struggle to manage object positions accurately (Bakr et al., 2023; Gokhale et al., 2022).

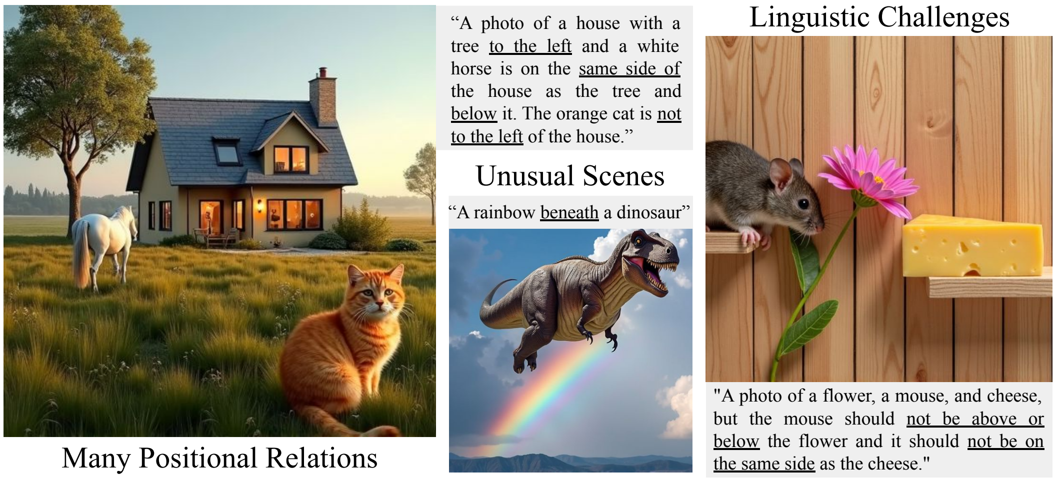


Figure 2: Stitch excels at complex positional prompts.

Position control improvements. Many base models improve spatial limitations during end-to-end training, for example with a joint understanding-generation objective (Wu et al., 2025b; Ma et al., 2025; Chen et al., 2025a;b; Deng et al., 2025) or prompt embeddings from Multimodal LLM (MLLM)-based encoders (Wu et al., 2025a; Fang et al., 2025; Wu et al., 2025c;d). In earlier T2I models, spatial guidance in forms such as bounding boxes could be used 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) often via generating and stitching sub-images. In fact, stitching images dates back to classical methods (Davis, 1998; Efros & Freeman, 2001), though modern approaches use deep networks. Methodologically, these externally position-guided modifications are related to tasks involving layout-conditioned T2I synthesis (Zhang et al., 2023a; Li et al., 2023; Farshad et al., 2023; Chen et al., 2024; Mou et al., 2024; Ohanyan et al., 2024; Xie et al., 2023; Mo et al., 2024), which are more constrained by fine-grained user input. Unfortunately, these methods are implemented on older U-Net or Diffusion Transformer (Peebles & Xie, 2023) architectures. In contrast, many recent models adopt MMDiT (Esser et al., 2024) or related architectures (Wu et al., 2025a; BlackForest, 2024; Cai et al., 2025), where such techniques often fail to generalize effectively (Jiao et al., 2025). Instead, it has been shown that MMDiT-based architectures may be more amenable to generation-level edits like prediction swapping (Jiao et al., 2025; Bader et al., 2025), offering insights into MMDiT's inference behavior that we build upon.

Position control benchmarks. Many T2I benchmarks evaluate image generation capabilities, with GenEval (Ghosh et al., 2023) and T2I-CompBench++ (Huang et al., 2023; 2025) among the most widely used. Both include basic spatial reasoning categories of form $\langle obj_1 \rangle \langle relation_{12} \rangle \langle obj_2 \rangle$. 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.

3 Methods

Despite progress in many aspects of image synthesis, position-related prompts continue to pose challenges for T2I generation models. To address this, we introduce Stitch, a flexible method to improve positional generation in models with MMDiT (Esser et al., 2024) or similar architectures.

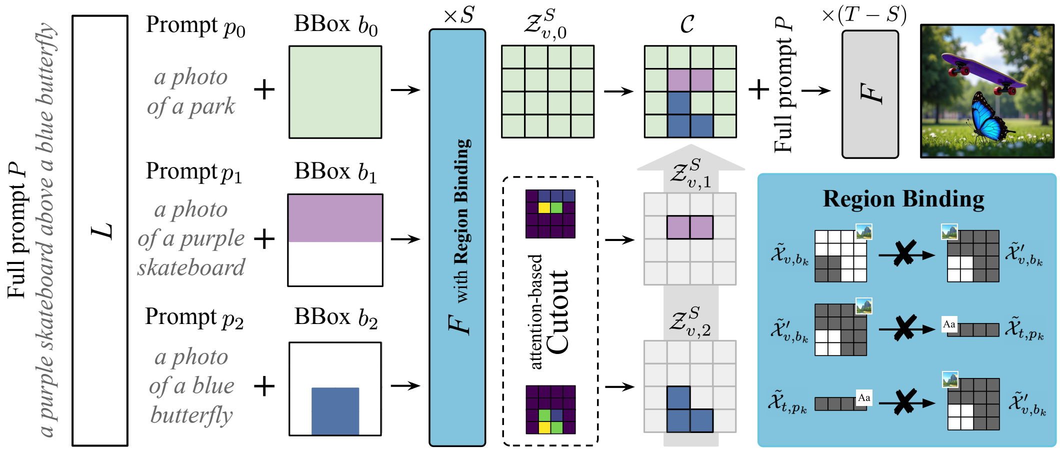


Figure 3: Stitch: Multimodal LLM L splits full prompt P into object prompts p_k and bounding boxes b_k, along with full-image background prompt p_0. MMDiT-based model F seperately sketches objects and background (*butterfly* ■, *skateboard* ■, *park* ■) for S timesteps. With **Region Binding** attention-masking 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.

3.1 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 single vector in the corresponding sequence. MMDiT-like (Esser et al., 2024) architectures process $\mathcal{X}_v$ and $\mathcal{X}_t$ iteratively with a stack of time-conditioned transformer blocks. Each block first applies a pre-attention transformation $(\tilde{\mathcal{X}}_v, \tilde{\mathcal{X}}_t) = F_{\text{pre}}(\mathcal{X}_v, \mathcal{X}_t; \tau)$, which includes 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 via functions Q, K, V derived from linear projections and normalization. Attention outputs are then computed as:

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

where $M : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \{0, -\infty\}$ is an attention mask, and $\mathcal{Z} = \{z(\tilde{x}_i)\}_{i=1}^{N}$ denotes the set of attention outputs. 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, where each head h computes attention independently over $\tilde{\mathcal{X}}$ with its own Q_h, K_h, V_h, and the outputs are combined into $\mathcal{Z}$. Stitch builds upon MMDiT's design, particularly with targeted masking within the multimodal transformer blocks.

3.2 Stitch: Training-Free Position Control for MMDiT

As seen in Figure 3, Stitch begins by using MLLM 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

$b_k = [x_{\min}, x_{\max}, y_{\min}, y_{\max}] \in \{0, \dots, W-1\}^4$. Each pair (p_k, b_k) represents a distinct object in the scene. 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 (detailed below). Once objects are adequately formed, their corresponding foreground latent tokens $\mathcal{Z}^S_{v,k}$ are cut from the resulting latent maps. We do so with our Cutout (discussed later in this section), the necessary information for which is obtained from a model-specific attention head. The extracted foreground latent tokens $\mathcal{Z}^S_{v,k}$ are combined with the common background latent tokens $\mathcal{Z}^S_{v,0}$ 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.

Region Binding To ensure objects are fully generated within their specified bounding boxes and to prevent capturing of partial fragments during foreground extraction, we introduce three attention-masking constraints. These constraints guide the model to focus generation solely within each bounding box, effectively isolating the object from the surrounding context.

Formally, 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 corresponding to the sub-prompt. In every 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.

Cutout To avoid visible seams from mismatched backgrounds, we extract the object latent tokens before constructing the composite latent $\mathcal{C}$. As generation is incomplete after step S, conventional segmentation tools like SAM (Kirillov et al., 2023) cannot be used to find the shape to cut out. However, we observe that in some attention heads, text vectors focus on foreground objects. Given such a head (selection discussed in Section 5.2), 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.

4 POSEVAL: BENCHMARKING POSITIONAL GENERATION CAPABILITIES

As models improve on basic *Position* tasks of form $\langle obj_1\rangle\langle relation_{12}\rangle\langle obj_2\rangle$, broader benchmarks are needed 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. Each task is designed to isolate and evaluate a specific aspect of T2I capabilities. 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. Figure 4 and the Appendix show example prompts. The new tasks include:

Two Objects (2 Obj): For completeness, we inherit GenEval's *Position* task, unchanged.

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.

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

Table 1: PosEval reveals that leading T2I models still struggle with complex positional prompts. However, Stitch boosts positional generation on 3 different MMDiT-based models, achieving SOTA.

Model	2 Obj	3 Obj	4 Obj	Neg	Rel	PAB	Avg.↑
No layout guidance							
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**±0.01	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
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
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

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.

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.

Negative Relations (Neg): To evaluate understanding of *Negative Relations*, 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" $\rightarrow$ "a photo of a dog and a teddy bear, a dog is not left of a teddy bear").

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.

5 Main Experiments

We validate our three primary contributions: Stitch, Cutout, and PosEval. In Stitch, $S = 10$ for FLUX and SD3.5 and 6 for Qwen-Image. For all three models, we use $T = 50$, $\kappa = 5$, and bounding boxes b_k are generated on a $W \times W = 32 \times 32$ grid with prompts p_k from GPT-5 (OpenAI, 2025) and GPT-4 (OpenAI, 2023). η is set to 0.95 for SD3.5 and FLUX. Results are on 3 seeds. All other parameters are the same as the base models. Our code is attached in the Supplementary Materials.

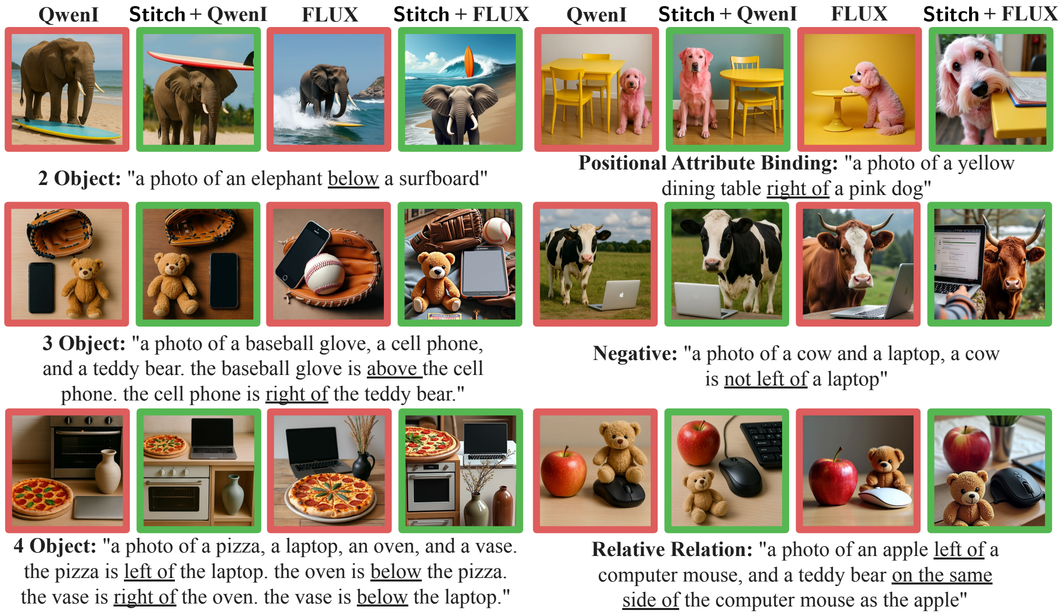


Figure 4: Stitch corrects Qwen-Image (QwenI) and FLUX position on PosEval without quality loss.

5.1 Enhancing Positional Understanding with Stitch

PosEval highlights that while SOTA models handle basic *Position* prompts well, they struggle with complexity, suggesting that the overall positional problem remains unsolved. Table 1 presents PosEval results, covering our five newly introduced tasks and GenEval's *Position* (*2 Obj*). While models like BLIP3-o (Chen et al., 2025a) and JanusPro (Chen et al., 2025b) perform the basic task well (87% and 79%, respectively) their accuracy drops sharply on harder tasks. On *Relative Relations*, they score only 8% and 11%, and when scaling to *Four Objects*, their accuracy falls to 2% and 4%. PosEval reveals greater shortcomings in top models than the basic *Position* task suggests.

Stitch consistently enhances base models on PosEval tasks, 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) and HRS-Bench (Bakr et al., 2023) in the Appendix). Stitch + Qwen-Image, ranks first on 5/6 PosEval tasks and ties with Stitch + FLUX on the last. It has the highest overall accuracy, with a 54% relative gain over the previous best, LMD (Lian et al., 2024).

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 1 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 2 further highlights Stitch's handling of complex scenes on FLUX. On the left, we show how it successfully interprets a prompt that combines all six PosEval task categories. Despite the complexity (featuring four 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.

5.2 Assessing Foreground Segmentation Accuracy of Cutout

We choose per-model attention heads by generating 80 images with the prompt "a photo of a $\langle$obj$\rangle$" 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.

In the Appendix, we present the top five attention heads for each base model, ranked by IoU with SAM maps and displayed with each head's optimal threshold μ. We also report Intersection over Target (IoT), calculated as $\frac{|\text{Prediction} \cap \text{Target}|}{|\text{Target}|}$. For each model, several attention heads exhibit strong IoU and IoT, indicating suitability for Cutout. For example, FLUX block 14's head 20 achieves IoU $= 62\%$ and IoT $= 92\%$. Example masks are presented in Figure 5 for this head, selected for FLUX Cutout. 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's purpose: the border does not degrade the images and helps capture the full object, although even small missing parts can be reconstructed in later generation steps.

5.3 Assuring Reliability of our PosEval Benchmark

PosEval adopts the same evaluation protocol as GenEval (Ghosh et al., 2023), whose components were thoroughly validated by the original authors. We conduct a user study to validate the reliability of the evaluation platform for the newly introduced benchmark categories. 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. Figure 6 shows the evaluation protocol performs consistently (table in Appendix), with all PosEval categories within 10% of the average inter-annotator agreement.

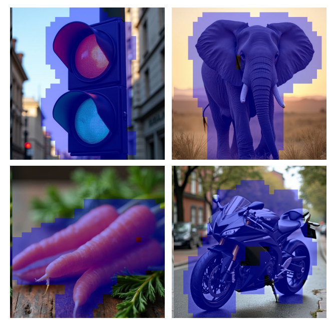


Figure 5: Cutout cleanly extracts objects mid-generation.

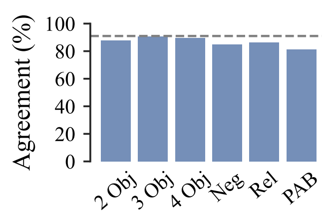


Figure 6: PosEval's evaluation is human-aligned, near average annotator agreement (line - -).

Each PosEval task is challenging yet complementary, allowing 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 1, strong *2 Obj* performance does not guarantee scaling. BLIP3-o (Chen et al., 2025a) scores 87% on *2 Obj* but drops to 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%, respectively. This suggests some models may overfit to previously seen tasks without effectively handling positional generation as complexity grows. Moreover, although *Negative* prompts generate the same images as *2 Obj* prompts, they are easier to satisfy 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%), the gap 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.

6 Additional Analysis and Ablations of Stitch

We investigate Stitch in greater detail. Table 2, presents an ablation study on FLUX to evaluate Region Binding and Cutout's contributions. After choosing a Cutout head, η is raised to 0.95 to widen the object buffer. This choice is also ablated. We test accuracy on PosEval and use a human study to evaluate *Blend* on the basic GenEval *Position* task, *2 Obj*. We define an image having *Blend* as being "visually coherent (and not like many images put side by side)". Figure 7 illustrates this effect: the right image (with Cutout) exhibits *Blend*, while the left image (without Cutout) does not.

Table 2: 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	$0.06_{\pm 0.01}$	$0.02_{\pm 0.00}$	$0.62_{\pm 0.01}$	$0.03_{\pm 0.01}$	$0.15_{\pm 0.02}$	1.0
✓			$0.81_{\pm 0.02}$	$0.58_{\pm 0.03}$	$0.52_{\pm 0.02}$	$0.90_{\pm 0.01}$	$0.63_{\pm 0.01}$	$0.51_{\pm 0.01}$	0.68
✓	✓		$0.51_{\pm 0.01}$	$0.30_{\pm 0.01}$	$0.18_{\pm 0.01}$	$0.71_{\pm 0.03}$	$0.31_{\pm 0.04}$	$0.29_{\pm 0.01}$	0.99
✓	✓	✓	$0.70_{\pm 0.02}$	$0.44_{\pm 0.01}$	$0.38_{\pm 0.01}$	$0.83_{\pm 0.02}$	$0.48_{\pm 0.02}$	$0.44_{\pm 0.00}$	0.95

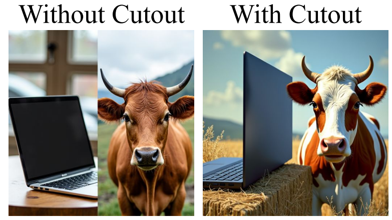
For each setup, three people independently and binarily decide if each of 100 images (one image per prompt) displays *Blend*, and we report the percentage of images for which they answer *yes*.

As shown in Table 2, 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 2, this improves accuracy without significantly affecting *Blend*. This final combination is what allows Stitch to achieve both precise positional control and high image quality.

Preserving base model quality while enhancing positional capabilities is essential. To validate, we compare each base model's quality with Stitch's via Aesthetic Score (Schuhmann & Beaumont, 2022). Across four images per prompt on the 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. This indicates that Stitch does not significantly degrade image quality.

To ensure diversity is preserved, we compute per-prompt mean pairwise distance between samples in DINOv2 (Oquab et al., 2024) embedding space. We find average diversity is preserved or improved: FLUX increases from 0.34 ± 0.15 to 0.38 ± 0.16; SD3.5 remains at 0.43 ± 0.15 from 0.43 ± 0.14; and Qwen-Image rises from 0.16 ± 0.11 to 0.22 ± 0.10.

Finally, we investigate timestep selection S for Region Binding on FLUX. In Figure 8, 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.



"A photo of a laptop left of a cow."

Figure 7: Cutout boosts coherence.

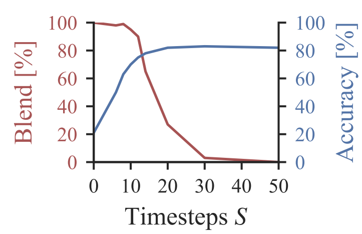


Figure 8: In Stitch, increasing the Region Binding timesteps S improves accuracy but reduces blend.

7 Conclusion

We introduce Stitch, a training-free method for enhancing position-related T2I generation in MMDiT-based models. Stitch generates coherent, position-accurate images by first generating individual objects constrained to LLM-generated bounding boxes with our novel Region Binding constraints. With our Cutout, it then employs a targeted attention head to extract and combine these objects, before completing the image with the full prompt, unconstrained. We evaluate Stitch on PosEval, our new benchmark extending GenEval with five challenging, position-focused tasks. Evaluating top models on PosEval reveals that even the strongest T2I models still struggle with position-related tasks, despite performing well on basic *Position* tasks. Yet Stitch significantly boosts Qwen-Image, FLUX, SD3.5, achieving SOTA results while remaining entirely training-free.

Reproducibility Statement

We ensure the reproducibility of our experiments by providing detailed descriptions in Sections 5, 6, and the Appendix. The Supplementary Materials contain the implementation code, the prompts used in our benchmark, the bounding boxes generated by our method, as well as a README.md file with detailed reproduction instructions.

Acknowledgements

This work was partially funded by the ERC (853489 - DEXIM) and the Alfried Krupp von Bohlen und Halbach Foundation, which we thank for their generous support. We are also grateful for partial support from the Pioneer Centre for AI, DNRF grant number P1. Mateusz Pach would like to thank the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program for support. The authors gratefully acknowledge the scientific support and resources of the AI service infrastructure *LRZ AI Systems* provided by the Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities (BAdW), funded by Bayerisches Staatsministerium für Wissenschaft und Kunst (StMWK). We would like to thank Jakub Pach for his assistance in conducting the user study.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *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. In *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 preprint arXiv:2505.22705*, 2025.

Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Controlstyle: Text-driven stylized image generation using diffusion priors. In *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. In *WACV*, 2024.

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

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

James Davis. Mosaics of scenes with moving objects. *CVPR*, 1998.

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 preprint arXiv:2505.14683*, 2025.

Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *SIGGRAPH*, 2001.

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. In *ICML*, 2024.

Luca Vincent Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *NeurIPS*, 2024.

Luca Vincent Eyring, Shyamgopal Karthik, Alexey Dosovitskiy, Nataniel Ruiz, and Zeynep Akata. Noise hypernetworks: Amortizing test-time compute in diffusion models. 2025. URL `https://api.semanticscholar.org/CorpusID:280641933`.

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. *arXiv preprint arXiv:2503.10639*, 2025.

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

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. In *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.

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 preprint 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. In *CVPR*, 2024.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *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. In *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. In *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. In *AAAI*, 2024.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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*, 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. In *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. *J. Mach. Learn. Res.*, 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. In *NeurIPS*, 2022.

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

Marcin Sendera, Łukasz Struski, Kamil Ksiażek, Kryspin Musiol, Jacek Tabor, and Dawid Rymarczyk. Semu: Singular value decomposition for efficient machine unlearning. *arXiv preprint 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 preprint 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 preprint 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 preprint 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 preprint 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 preprint 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 preprint 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. In *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. In *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 preprint arXiv:2412.03859*, 2024.

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

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

CONTENTS OF APPENDIX

A Use of LLMs

This work used LLMs to improve sentence clarity and flow.

B Stitch Results on Existing Benchmarks

Table 3: 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.

C PosEval User Study Results

We include the numerical results of our user study verifying PosEval's automated evaluation in Table 4. As mentioned in the main paper, all results are within 10% of the inter-annotator alignment (91%).

Table 4: 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%

D Additional Qualitative Examples

Figure 9 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 11, and for Stitch with Qwen-Image in Figure 10, also on two prompts per task.

Figure 9: Qualitative examples for Stitch + SD3.5.

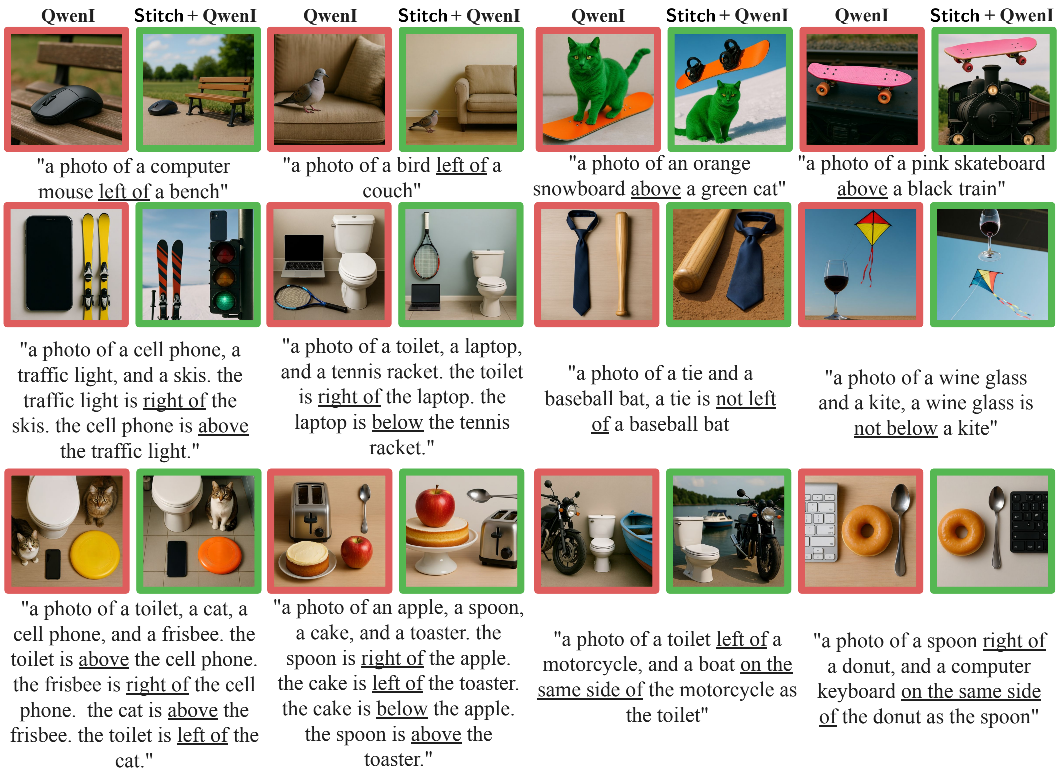


Figure 10: Additional qualitative examples for Stitch + Qwen-Image (QwenI).

Figure 11: Additional qualitative examples for Stitch + FLUX.

E 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 5. 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 5: 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)	✓	✓	✓	✓	✓	✓

F Top Cutout Segmentation Heads for each Model

We present the results of the top attention heads for FLUX (Table 6), Qwen-Image (Table 8), and SD3.5 (Table 7). 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 13. 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 12, 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 6: 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 7: 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 8: 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

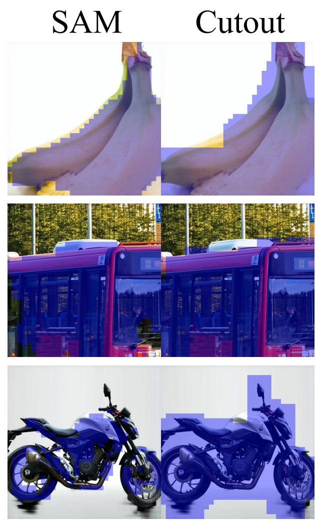


Figure 12: Segmentation head maps for SD3.5.

Figure 13: Attention weights and Cutout masks from head 20 in block 14 of FLUX, at step 10 of 50.

G 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}
```

H USER STUDY INTERFACE

In Figures 14 and 15, we include and example of the user interface for our user studies to measure PosEval's quality and *Blend* of the images.

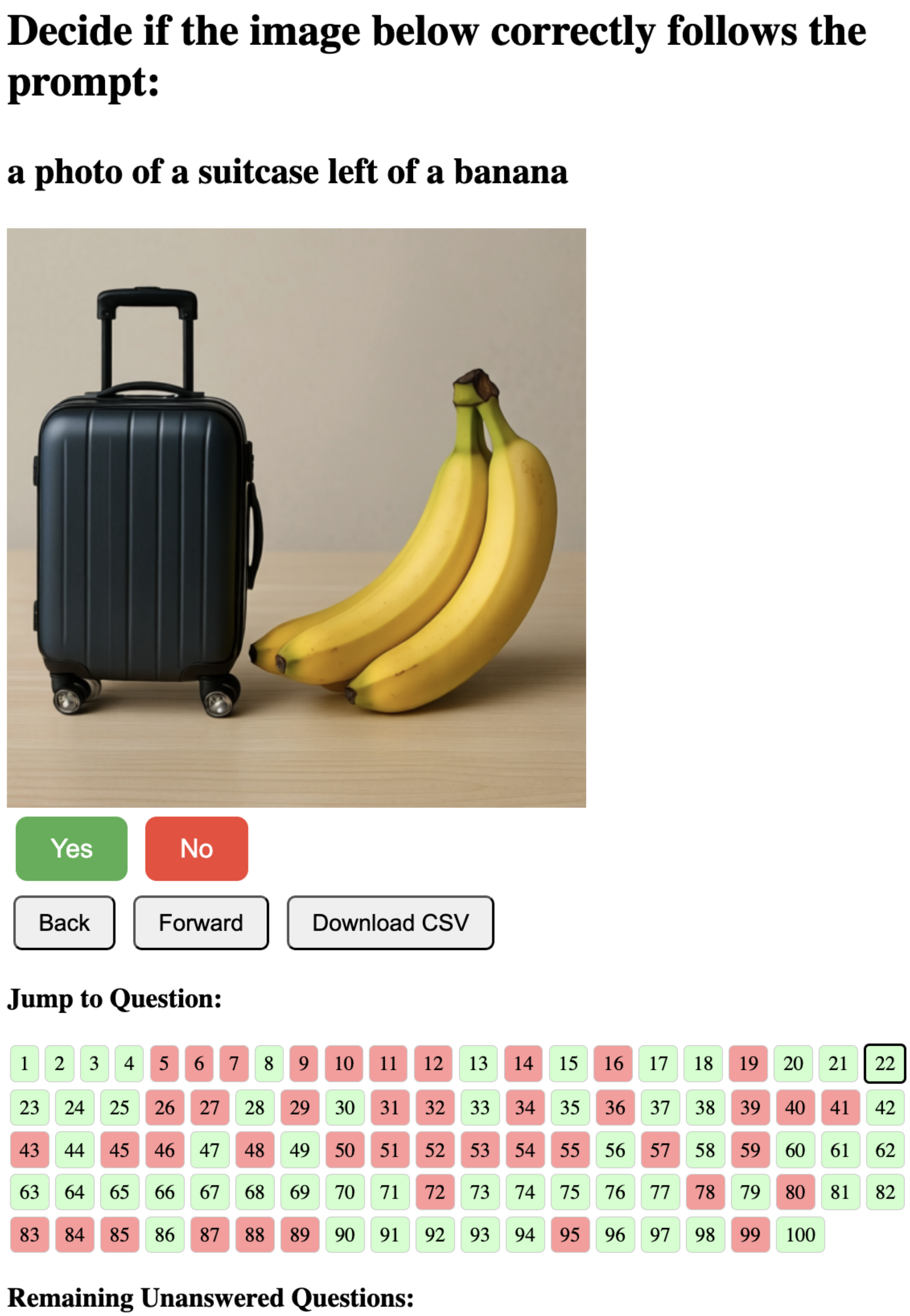


Figure 14: *PosEval* user study.

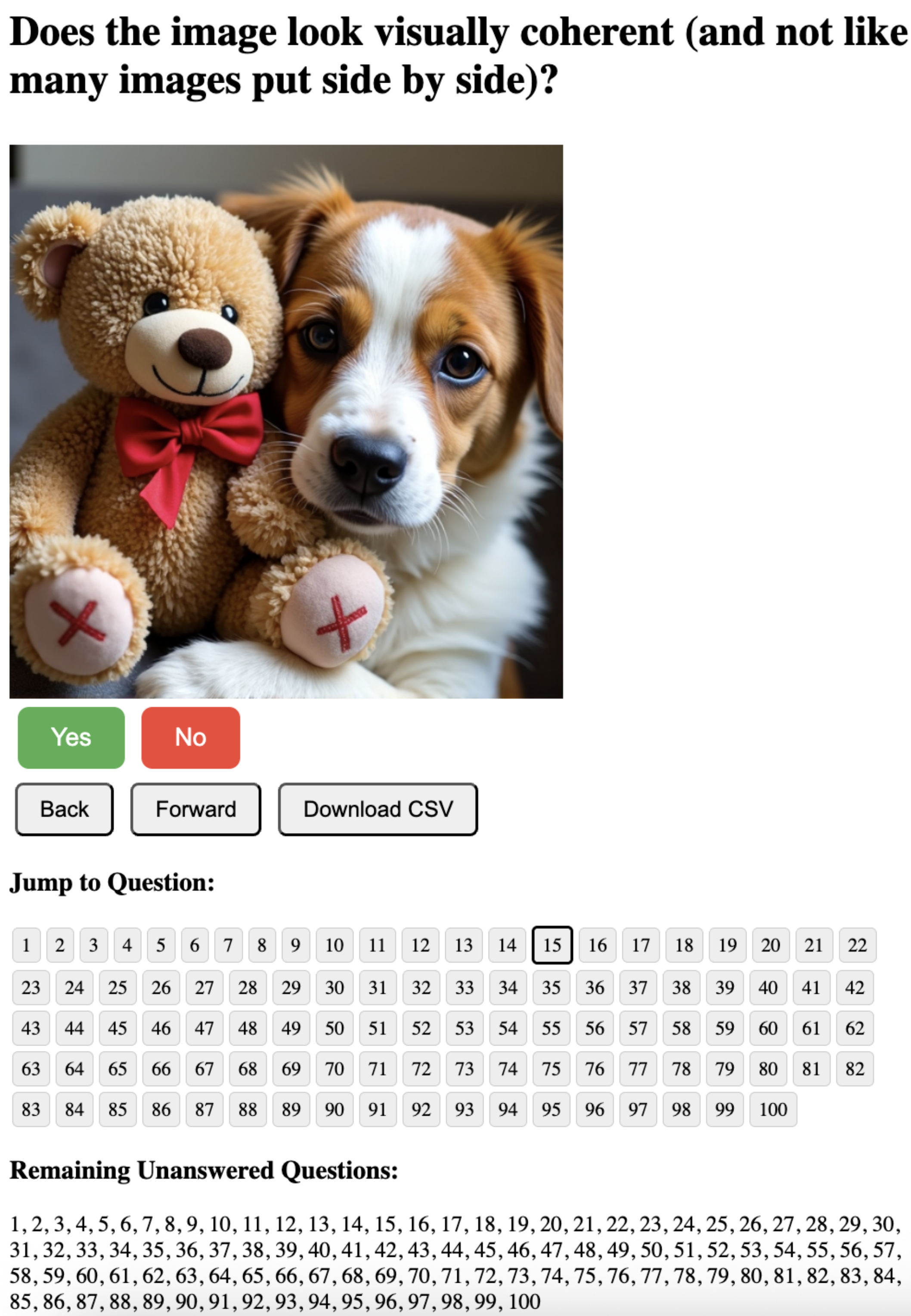


Figure 15: *Blend* user study.