The 28th Meeting on Image Recognition and Understanding

Controlling Unseen Compositions in Diffusion Models by Swapping Positional Embeddings

Peilin Xiong[1,a)] Junwen Chen[1,b)] Honghui Yuan[1,c)] Keiji Yanai[1,d)]

Abstract

Modern text-to-image diffusion models exhibit remarkable fidelity in synthesizing photorealistic imagery yet struggle when tasked with generating novel object compositions beyond their training distribution. This limitation arises from the entanglement of positional semantics and spatial context within standard diffusion processes, causing localized regions to inherit rigid structural biases from their surroundings. We present a training-free framework that redefines compositional generation through dynamic manipulation of twin positional embeddings. By decomposing complex hybrids into sub-components and strategically swapping their positional embeddings across diffusion timesteps, our method enables models to transport visual concepts (e.g., transplanting human heads onto animal bodies) while preserving spatial plausibility. Our core innovation is the temporal control mechanism for encoding interventions: early-phase swaps borrow structural blueprints from donor regions, while late-phase reversion to original embeddings integrates details with the local context. Combined with the FLUX-Fill we can replicate objects in an image and blend them in perfectly with the environment. This work establishes positional embedding as a critical lever for out-of-distribution synthesis without architectural modifications or retraining.

1. Introduction

Despite remarkable progress in synthesizing photorealistic images, generative models remain fundamentally limited in creating *uncommon object compositions* due to training data constraints. Through empirical analysis of FLUX [4], we observe a critical gap: when prompting the model to generate hybrid objects (e.g., a human head on a dog's body), the output frequently collapses into generating familiar patterns (e.g., a standard dog) (Fig. 1), even with explicit textual guidance. This suggests that conventional text-driven control is insufficient for controlling the spatial arrangement of novel concepts unseen in training data.

1 The University of Electro-Communications, Tokyo, JAPAN
a) xiong-p@mm.inf.uec.ac.jp
b) chen-j@mm.inf.uec.ac.jp
c) yuan-h@mm.inf.uec.ac.jp
d) yanai@cs.uec.ac.jp

(a) (b)

Fig. 1: Comparison results. (a) Baseline FLUX: with the prompt "a man and a dog," the model produces only a vigilant dog, missing the intended human–dog composition. (b) Ours — Positional-Embedding Swapping: under the same prompt, our method generates a man crouching on a country path and hugging his dog, faithfully capturing their spatial relationship.

Fig. 2: Using FLUX-Fill and our method to copy objects. The source image (right) contains a paper cup in front of a rock wall. The target image (left) shows the cup copied and seamlessly inserted beside a dog, preserving lighting, shading, and scene consistency.

Our method addresses these challenges through a novel approach to spatial reasoning, by leveraging the often overlooked role of positional embeddings in transformer-based diffusion architectures. Our analysis reveals that standard models like FLUX use img-ids to generate position embedding, which are latent grid coordinates processed through Rotary Position Embedding (RoPE) [2], to help the Transformer better understand the spatial relationships within an image by providing more accurate and flexible encoding of 2D position information. By dynamically swapping these embeddings between user-specified regions during generation, we enable the model to reevaluate spatial con-

texts while preserving its learned visual knowledge. Our method operates through three-phase control:
(1) Spatial alignment rescales donor bbox and recipient regions bbox to match latent grid dimensions;
(2) Timestep-thresholded ID transplantation overrides positional semantics during critical structural formation phases ($t >= \tau$);
(3) Contextual restoration in later denoising stages ($t < \tau$) leverages the model's refinement capabilities to resolve geometric discontinuities.

In Fig. 2, integrating with FLUX-Fill [4]'s masked editing framework extends this approach to multi-image composition, where positional embeddings from source images guide synthesis in target regions while maintaining environmental consistency. Crucially, our method requires neither model retraining nor gradient optimization—swapping operations while without introducing additional parameters.

This theoretical insight, coupled with practical validation across diverse hybrid categories, positions positional embedding manipulation as a foundational tool for controllable out-of-distribution generation. The abrupt semantic collapses observed at τ-threshold boundaries ($\Delta\tau = 0.01$) further highlight opportunities for adaptive temporal control mechanisms in future work.

2. Related Work

2.1 Diffusion-Based Image Editing Methods

Prior techniques for text-driven image manipulation can be broadly categorized into two types: mask-guided editing and attention manipulation. Methods like DiffEdit [1] and Null text inversion [5] require object masks or iterative optimization to restrict edits to localized regions. Prompt-to-Prompt [3] introduced mask-free editing by swapping cross-attention maps between source and target prompts in specific timesteps, based on the observation that early timesteps control global layouts while later timesteps refine details.

2.2 Positional embedding in Vision Models

Positional embedding plays a critical role in spatial reasoning for both language and vision transformers. Rotary Position Embedding (RoPE) [2], originally designed for language models, introduced relative positional information via rotation matrices, enhancing extrapolation capabilities. In diffusion models like FLUX, RoPE is integrated into attention layers to model 2D spatial hierarchies implicitly. Unlike absolute positional embedding, RoPE captures dynamic geometric relationships through relative distances, offering a flexible framework for spatial control.

3. Method

3.1 Positional embedding via Image IDs

In Fig. 3, we show the pipeline of our method on the FLUX model. The FLUX diffusion model uses a structured positional embedding scheme based on *image IDs*

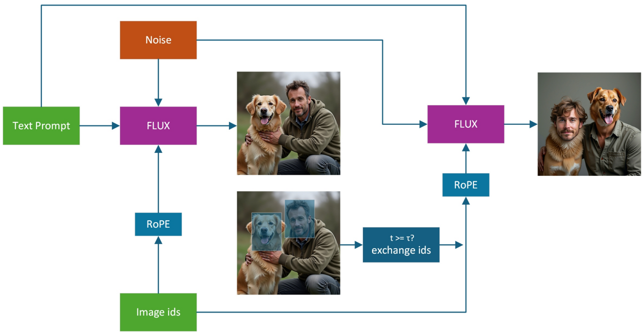


Fig. 3: Overview of our method on FLUX model.

(img_ids) that define spatial coordinates in the latent grid. During the initial denoising step, latent features are partitioned into a grid of size $\left(\frac{h}{2}\right) \times \left(\frac{w}{2}\right)$ for an $h \times w$ image. Each grid cell is assigned a unique *img_id* composed of three parts:
(1) A constant base value (0),
(2) Row indices from 0 to $\frac{h}{2} - 1$,
(3) Column indices from 0 to $\frac{w}{2} - 1$.
Broadcast across the batch dimension, these IDs are processed through RoPE to generate rotation matrices that modulate attention operations.

RoPE injects relative positional biases by rotating query/key vectors using sinusoidal transformations based on *img_ids*. This allows the model to learn spatial relationships through geometric transformations of attention weights. Crucially, the modular design of *img_id* generation enables dynamic manipulation of positional coordinates without altering the underlying diffusion architecture.

3.2 Dynamic Positional ID Swapping

To enable compositional generation of unseen hybrids, our method dynamically swaps *img_ids* between user-specified source and target regions during the diffusion process. Given source bounding box, B_s, (e.g., human head) and target bounding box, B_t, (e.g., dog head), the method performs the following steps:
(1) **Spatial Alignment**: Rescale B_s and B_t to match dimensions in the latent grid, ensuring row/column indices align with the $h//2 \times w//2$ *img_id* structure.
(2) **ID Transplantation**: Replace *img_ids* within B_t with those from B_s, while preserving relative grid offsets for geometric coherence.
(3) **Timestep Thresholding**: Activate swapped *img_ids* only when the normalized timestep $t >= \tau$, with τ governing the balance between structural borrowing and contextual refinement.

This selective swapping forces the model to address conflicting spatial cues: early denoising timesteps utilize the donor region's positional context to initialize coarse structure, while later timesteps revert to original *img_ids* to integrate details with surrounding regions. The τ parameter strikes a balance between novelty and coherence—lower τ offers more structural guidance but increases the risk of geometric artifacts, while higher τ emphasizes local consistency.

3.3 Timestep-Controlled Denoising Process

FLUX's generation follows a reverse-time denoising schedule $t \in [1.0 \rightarrow 0.0]$, where early steps ($t \approx 1.0$) synthesize coarse layouts, while later steps ($t \rightarrow 0.0$) refine details. our method partitions this process into two critical phases:

(1) **Early-Phase Structural Override ($t >= \tau$)**:
Swapped *img_ids* override FLUX's inherent position-semantic biases (e.g., defaulting to dog anatomy despite "human head" prompts). By disrupting spatial priors learned from training data, the model is forced to synthesize structures consistent with the textual prompt and guided by the donor region's positional context.

(2) **Late-Phase Contextual Anchoring ($t < \tau$)**:
Restoring original *img_ids* grounds the synthesized structure in its local environment (e.g., blending a human head's features with a dog's fur texture). This phase utilizes FLUX's refinement capabilities to resolve inconsistencies while preserving the edited layout.

Empirical analysis identifies $\tau = 0.9$ as optimal for most hybrid objects, striking a balance between overriding default spatial biases early enough, and enabling smooth integration by being late enough. Higher $\tau(e.g., 0.95)$ shortens the override window, yielding minimal changes, while lower $\tau(e.g., 0.6)$ causes geometric distortions due to extended positional conflicts.

(a) prompt: "man chasing dog" (b) using our method

Fig. 4: Because the prompt did not specify a location, the position of the target object in the image shifted using our method.

(a) prompt: "man with dog, The man is on the right and the dog is on the left" (b) using our method

Fig. 5: Explicit Prompting

3.4 FLUX-Fill Integration for Mask Guided Editing

To facilitate localized compositional edits, our method utilizes FLUX-Fill, a masked variant of FLUX that restricts modifications to user-specified regions. Given an input image and a binary mask, M, (white = editable region, black = preserved), the method operates as follows:

(1) **Target Region Isolation**:
The mask, M, (e.g., a dog's head) defines the editable area. Only *img_ids* within M are modified, while unmasked regions maintain their original positional embeddings.

(2) **Positional ID Injection**:
During denoising, *img_ids* from the source region (e.g., a human head in a reference image) replace those in M. This forces FLUX-Fill to reinterpret the masked area's spatial context using the donor's positional biases.

(3) **Timestep-Aware Blending**:
Swapped *img_ids* are activated only when $t >= \tau$ (e.g., $\tau = 0.85$), allowing structural borrowing from the source. At $t < \tau$, the original *img_ids* are restored, allowing FLUX-Fill to refine details using the surrounding context (e.g., matching fur texture around the dog's neck).

This approach addresses FLUX-Fill's limitation in synthesizing unseen hybrids by separating structural guidance (positional swaps) from contextual refinement. Crucially, the mask ensures that edits remain localized, preventing global inconsistencies. Experiments show $\tau = 0.85$ achieves an optimal balance between fidelity to the source structure (human head) and seamless integration with the target environment (dog body).

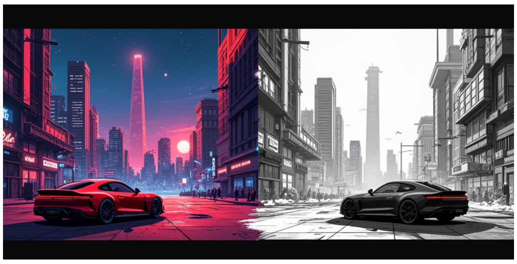
4. Experiments

4.1 Evaluation on FLUX Base Model

Problem Statement:

When generating novel object compositions (e.g., "human head on a dog body") through positional embedding (PE) swapping, FLUX exhibits spatial misalignment: the swapped regions (e.g., head bounding boxes) Fig.4 often deviate from their intended positions due to conflicting positional semantics inherited from the latent space.

Mitigation via Explicit Prompting: Explicit spatial prompts (e.g., "man with dog, The man is on the right and the dog is on the left") Fig.5 partially mitigate the misalignment by reinforcing positional relationships with precise prompts.



(a) Direct generation

(b) Using our method

Fig. 6: **Mitigation via Dual-Halves Explicit Prompting.** (a) Split-screen result generated directly; a car appears in the monochrome (right) half, but it does not match the sleek *red* car requested by the prompt. (b) After copying the car's positional encoding and appearance from the left to the right, the identical red car is reproduced in the correct color.

4.2 FLUX-Fill with our method

Approach:

To address spatial misalignment, we integrate our method with FLUX-Fill, a masked editing variant of FLUX. Given an input image and a binary mask M, FLUX-Fill restricts PE swapping to M, preserving unmasked regions.

Implementation:

(1) **Input Pairing**: Combine a source image (e.g., human portrait) and a target image (e.g., dog) into a concatenated latent tensor.

(2) **Mask-Guided Swapping**: Define M on the target image (e.g., dog's head region) and inject source positional embeddings (human head) into M.

(3) **Timestep Control**: Activate swapped embeddings only when $t >= \tau(\tau = 0.85)$ to balance structural borrowing and contextual blending.

4.3 Multi-Image Composition via Masked Fusion

Approach: To enable cross-image content transplantation (e.g., inserting a cup from Image A into Image B), we:

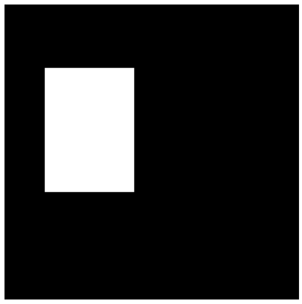
(1) **Input Concatenation**: Stitch Image A (source) and Image B (target) into a unified latent tensor.

(2) **Dynamic Masking**: Define a binary mask M on Image B's target region .

(3) **Timestep-Controlled Swapping**:

- Inject positional embeddings from Image A's source region into M **only when** $t >= \tau$.

(a) Generated image



(b) mask

Fig. 7: using FLUX-Fill

(a)

(b)

(c)

Fig. 8: Comparison of different τ values. From left to right: $\tau = 0.95$, $\tau = 0.96$, and the original image.

- At $t < \tau$, restore Image B's original embeddings to harmonize details.

4.4 Threshold Effect of Timestep τ

(1) **Stepwise Collapse**:

- At $\tau = 0.96$, swapped regions retain source structure.
- At $\tau = 0.95$, source features abruptly collapse, replaced by target contextual attributes Fig.7.

(2) **Nonlinear Transition**: The model exhibits **binary behavior**—swapped content either fully persists or vanishes between adjacent τ values ($\Delta\tau = 0.01$) (Fig.8).

5. Conclusion

We introduce a training-free method that creates unseen object compositions by dynamically swapping positional embeddings in diffusion models. With *FLUX-Fill*'s mask-guided editing, the approach transplants local content (e.g., a human head onto an animal body) with precise spatial control—without any model retraining. Key limitations remain:

(1) The timestep threshold τ still requires manual tuning;

(2) Even a small change ($\Delta\tau = 0.01$) can fail when the swapped regions are semantically incompatible;

(3) Abrupt transitions between denoising phases can leave discontinuous hybrid structures.

Future work should automatically adapt τ based on object semantics and support exchanging bounding boxes of different sizes—ideally by compressing or expanding latent tokens rather than distorting the image aspect ratio.

References

[1] Couairon, G., Verbeek, J., Schwenk, H. and Cord, M.: DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance, *ICLR* (2023).

[2] Halacheva, A.-M., Nayyeri, M. and Staab, S.: Expanding Expressivity in Transformer Models with MöbiusAttention, *CoRR* (2024).

[3] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y. and Cohen-Or, D.: Prompt-to-Prompt Image Editing with Cross Attention Control, *ICLR* (2023).

[4] Labs, B. F.: FLUX, `https://github.com/black-forest-labs/flux` (2024).

[5] Mokady, R., Hertz, A., Aberman, K., Pritch, Y. and Cohen-Or, D.: Null-text Inversion for Editing Real Images Using Guided Diffusion Models, *CVPR* (2023).