Controlling Unseen Compositions in Diffusion Models by Swapping Positional Embeddings

OS2B-10

Peilin Xiong Junwen Chen Honghui Yuan Keiji Yanai
The University of Electro-Communications, Tokyo, JAPAN

Abstract

- Modern diffusion models struggle to synthesize unseen compositions (e.g., "a human head on a dog body").
- We propose a **training-free** method that dynamically **swaps positional embeddings** to guide structure generation during early denoising and restore local details later.
- Our approach works with **FLUX.1-dev** and its editing variant **FLUX.1-Fill**, without retraining or modifying model parameters.

FLUX.1-dev

SD3.5 large

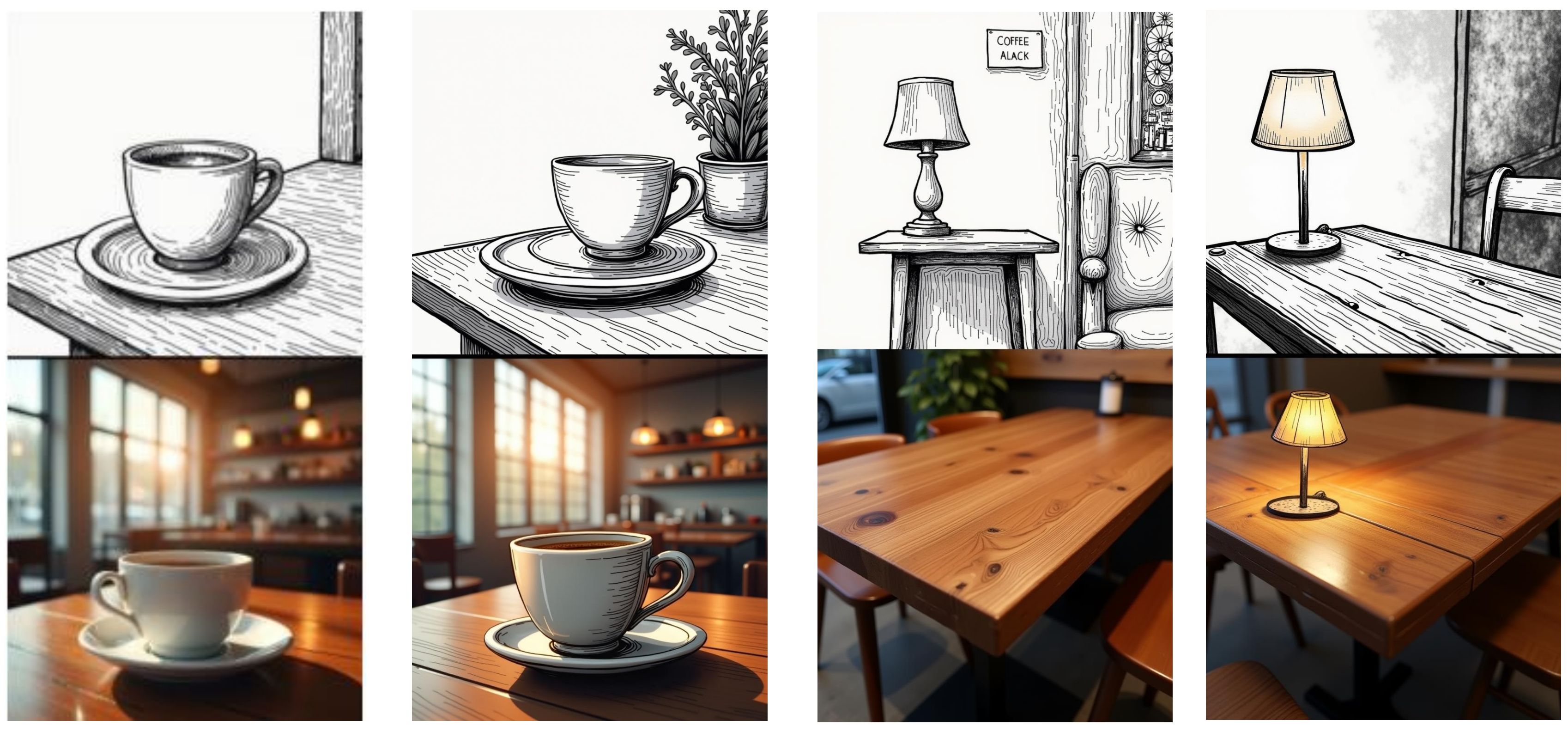
HiDream-I1-Dev

Goal

- Enable controllable spatial manipulation by injecting donor region's positional semantics in early denoising steps.
- Restore target region's embeddings in later steps for coherent blending.

Results

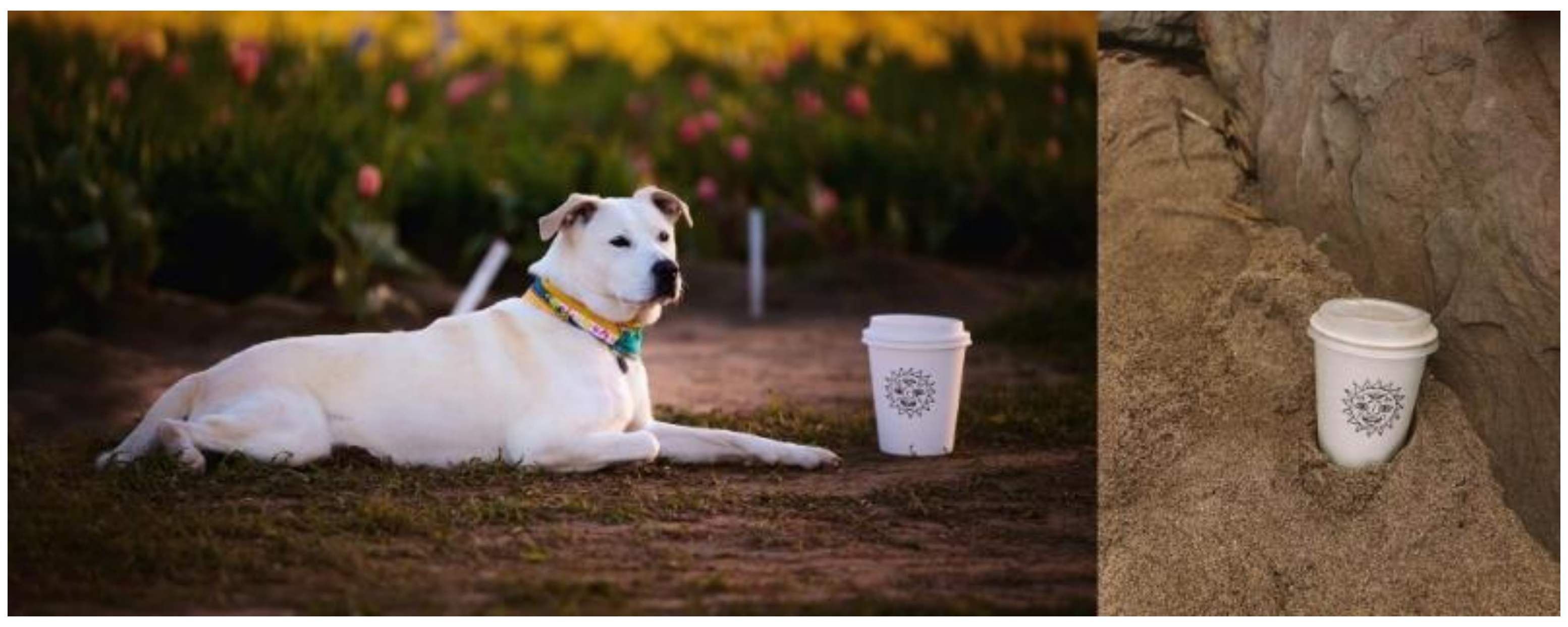
Split-Screen Composition

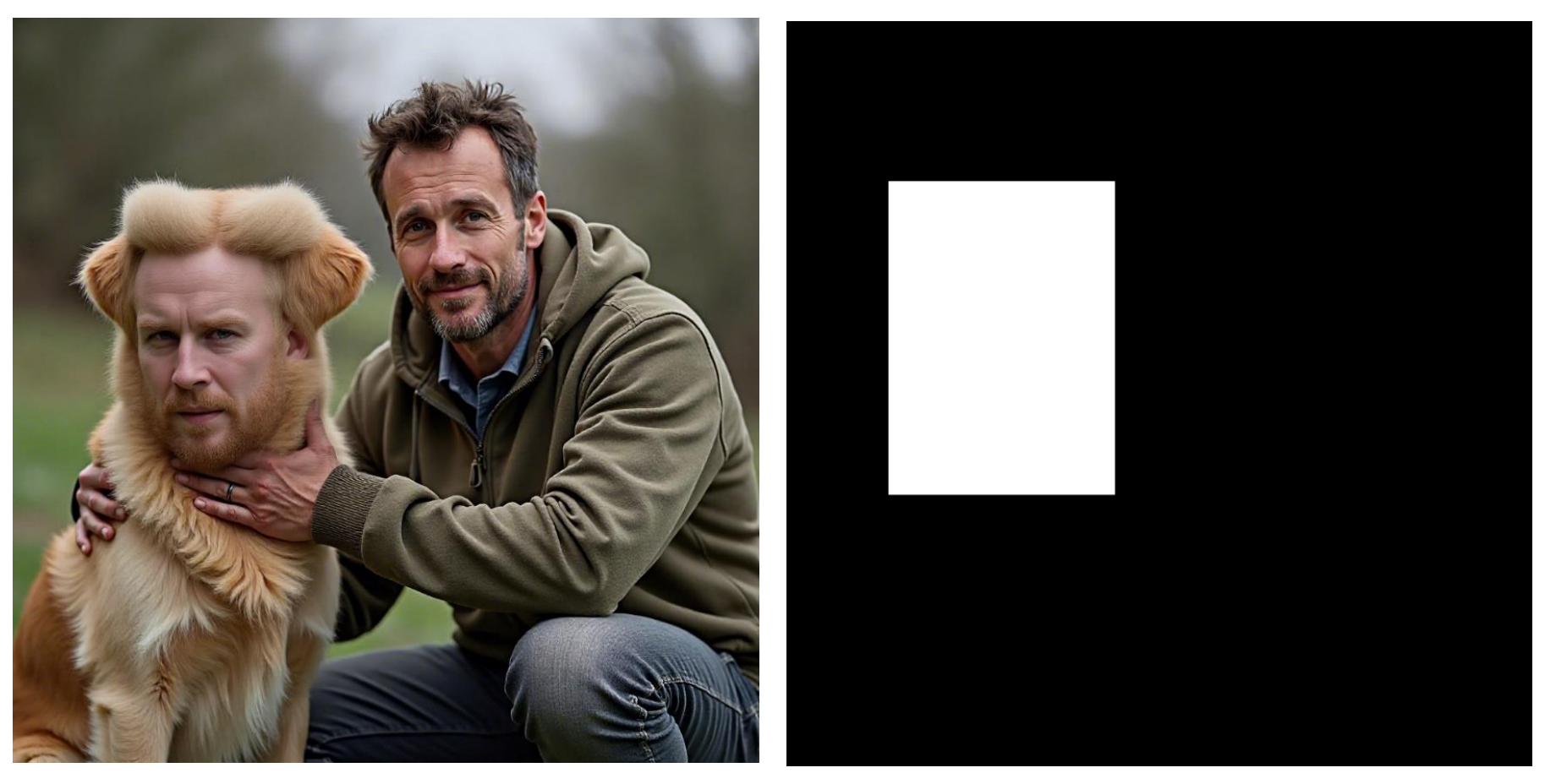


Prompt: ” split-screen composition” can be used to simultaneously generate two vertically stacked images
Top/Bottom: Independent style prompts

1. **FLUX.1-dev Split-Screen Generation**
2. **Our Embedding Swap**

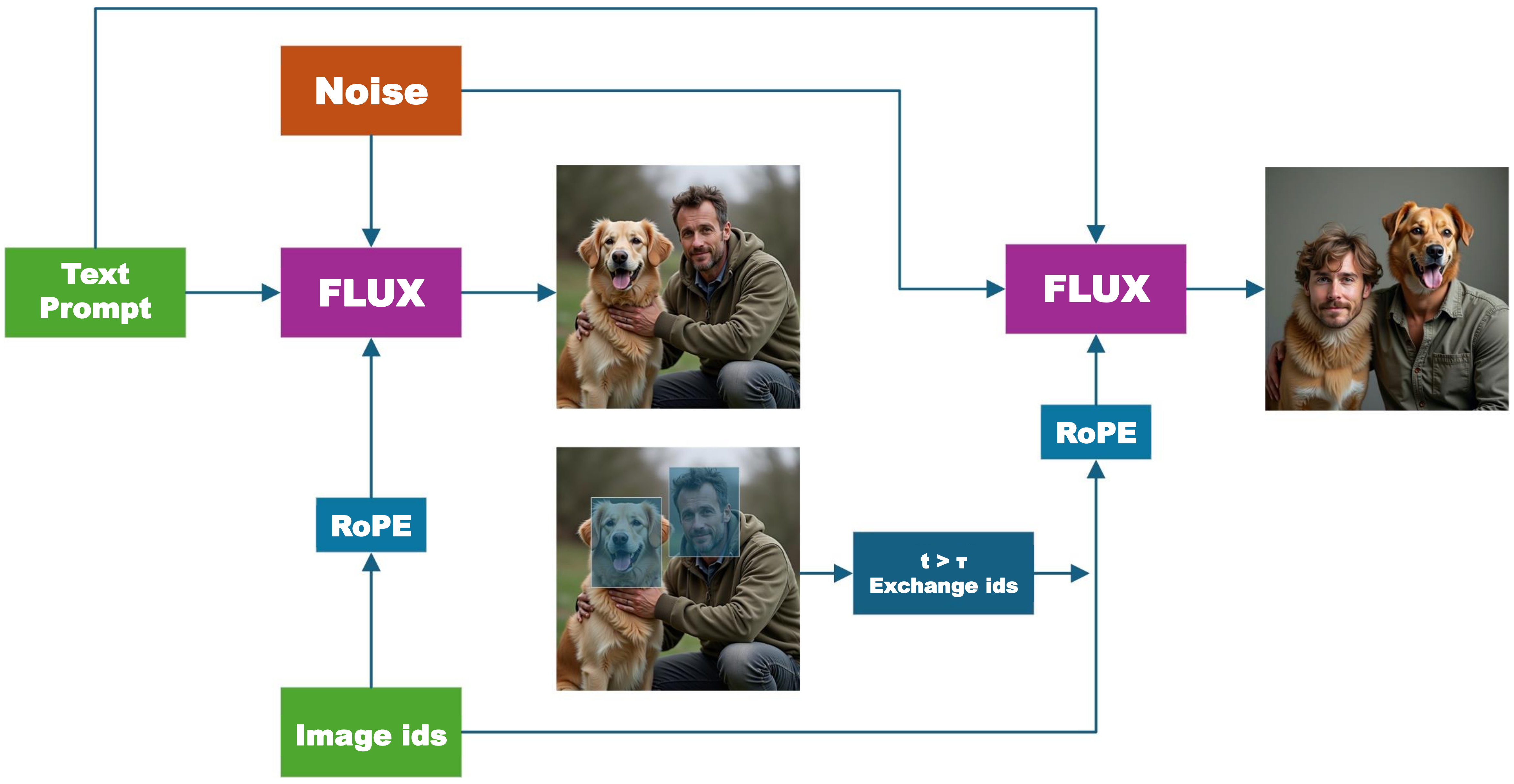
Localized Editing via FLUX-Fill Integration





• Binary mask pinpoints the edit region
• Positional-embedding swaps apply the change locally
• Background pixels are preserved automatically

Method



- We manipulate img-ids (latent coordinates) that FLUX uses to understand spatial relationships via RoPE (Rotary Positional Embedding).
- RoPE [2] injects relative positional information into Transformers by rotating queries and keys, enabling attention to remain sensitive to relative positions.

Key Mechanism:

EARLY STAGE (t > τ)	**LATE STAGE (t ≤ τ)**
-Swap IDs across regions	-Revert to native IDs
-Forces donor structure via RoPE	-RoPE harmonizes details

Key Innovation:

Timestep-controlled RoPE manipulation
⇥ Coarse structure transfer ⟶ Seamless refinement

[1] Black Forest Labs, S. Batifol, A. Blattmann, *et al.*, “FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space,” *arXiv preprint arXiv:2506.15742*, 2025.
[2] Su J, Ahmed M, Lu Y, et al. Roformer: Enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063.

2025.07