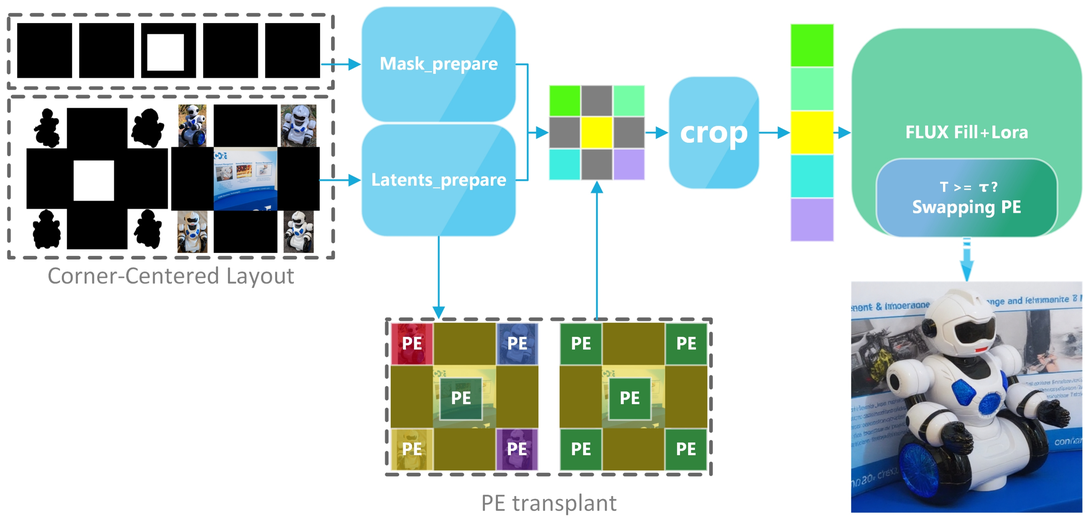
PosBridge: Multi-View Positional Embedding Transplant for Identity-Aware Image Editing

Peilin Xiong **Junwen Chen** **Honghui Yuan** **Keiji Yanai**
Department of Informatics, The University of Electro-Communications, Tokyo, Japan

Introduction

- Training-free **PosBridge** enables identity-aware, spatially controllable insertion.
- SwPE aligns structure; Corner-Centered tokens cut bias and memory.
- Optional class-agnostic LoRA adds detail; strong identity and coherence.

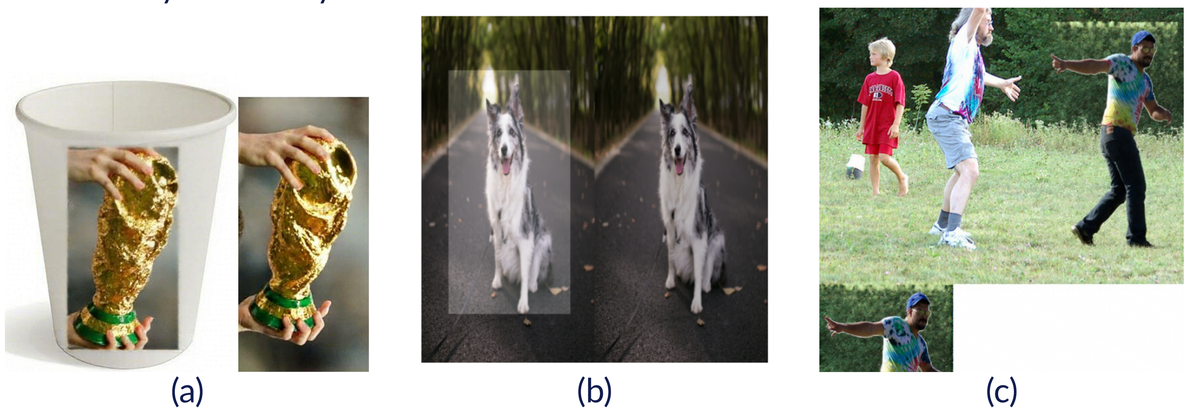
Method



- **Overview:** Training-free, identity-aware, spatially controllable insertion built on pre-trained MMDiT [1].
- **SwPE (mask-guided):** Early steps align structure; later steps blend details.
- **Corner-Centered + tokens:** Reduce spatial bias; restructure tokens for efficiency.
- **Optional LoRA:** Parameter-efficient detail enhancement [2].

Positional Embedding Effects

- Shared PEs duplicate structure across regions.
- Visual priors can override prompts when halves are identical.
- Boundary continuity leaks.



(a) (b) (c)

Experiments

- Baselines: AnyDoor [3], Insert-Anything (trained on FLUX.1-Fill, same as ours) [4]; Dataset: DreamBooth [5]; Metrics: CLIP-I, DINOv2, C+D.
- Training: FLUX.1-Fill backbone at 512×512 on a single A100; optional LoRA (rank 32) trained 6500 steps using infrequent T5 text tokens (e.g., ``Kwa'', ``McKe'') as class prompts; inference uses $\tau = 2$.

Methods marked with * use single-reference input. C+D is the average of CLIP-I and DINOv2.

Method	DINOv2 ↑	CLIP-I ↑	C+D ↑
AnyDoor*	0.7247	0.8662	0.7954
Insert-Anything*	**0.7638**	0.8777	**0.8208**
Copy-Paste	0.7760	0.8882	0.8321
Ours (SwPE, training-free)	0.7002	0.8701	0.7852
Ours (LoRA + SwPE)*	0.7149	0.8778	0.7963
Ours (LoRA)	0.7264	0.8815	0.8039
Ours (LoRA + SwPE)	0.7351	**0.8827**	0.8089

Conclusion and Future Work

Identity-consistent, controllable insertion via positional-embedding transplant + Corner-Centered Layout; LoRA refines appearance. Next: attention-guided adaptive τ and post-swap masking.

References

[1] Esser et al., 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis.

[2] Hu et al., 2022. LoRA: Low-Rank Adaptation of Large Language Models.

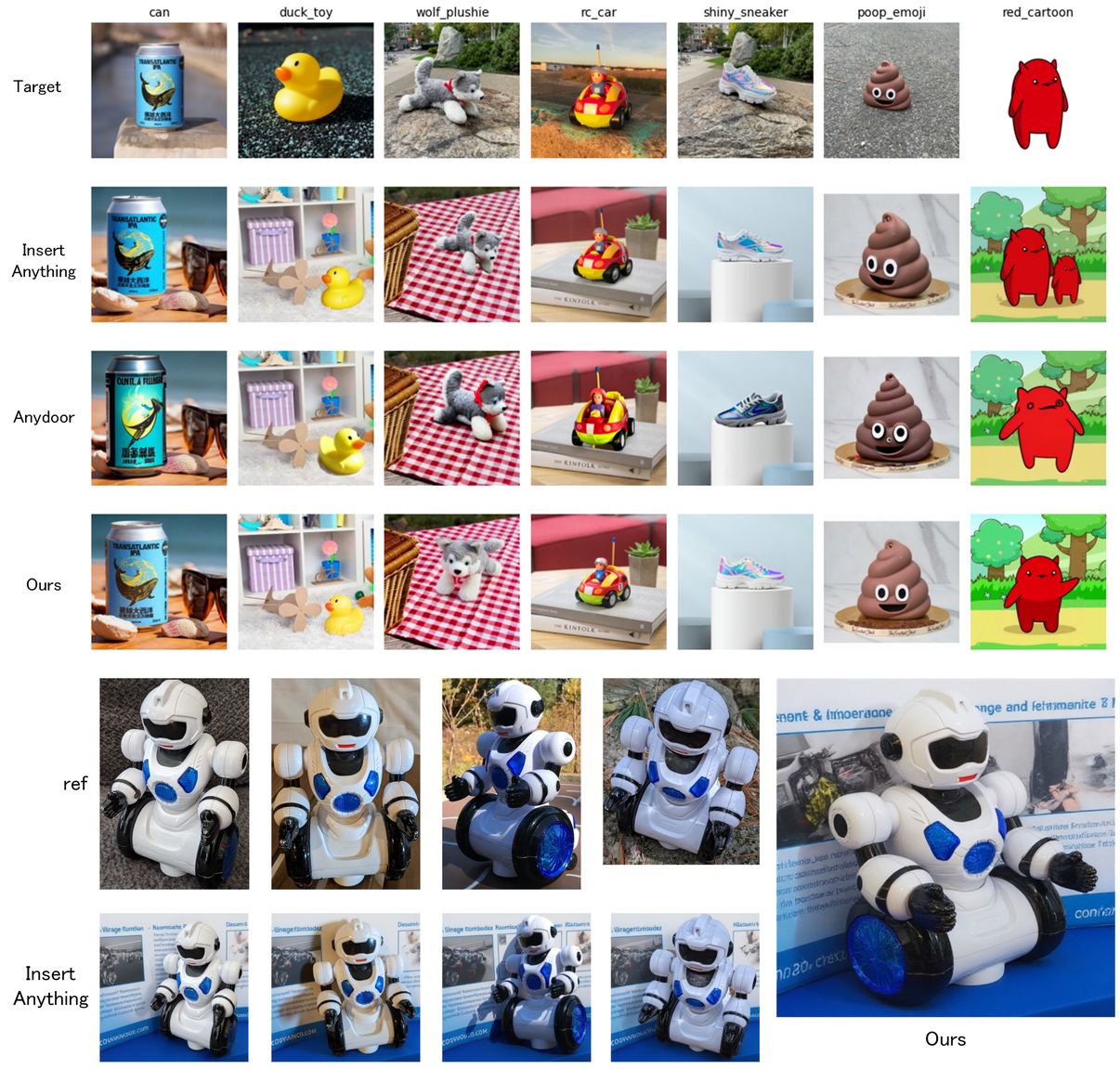
[3] Chen et al., 2024. Anydoor: Zero-shot object-level image customization.

[4] Song et al., 2025. Insert Anything: Image Insertion via In-Context Editing in DIT.

[5] Ruiz et al., 2023. DreamBooth: Fine-Tuning Text-to-Image Diffusion Models for Subject-Driven Generation.

Experiments (Qualitative)

- Qualitative comparisons assess consistency, realism, and controllability.
- Single-reference often fails to harmonize: shape distortion, inconsistent shading, identity collapse; Insert-Anything shows copy-paste artifacts.
- Ours uses four references + SwPE to form robust, pose-consistent representations; better geometry, viewpoint consistency, and seamless scene integration.



Ablation Studies

- LoRA increases detail and identity consistency (higher CLIP-I). SwPE improves spatial structure and layout (higher DINOv2).
- LoRA+SwPE yields the best overall performance.
- Results (512×512): SwPE 0.7002/0.8701; LoRA 0.7264/0.8815; LoRA+SwPE 0.7351/0.8827.
- Corner-Centered multi-reference reduces copy-paste vs. single reference.
- τ effect: 2–4 optimal; 8 boundary artifacts; 16/49 ghosting; we use 2.