CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

Dual-Stream Discrete Geometric Gating Network for Alleviating VAE Coupling in Image Editing

Peilin XIONG[1,a)] Honghui YUAN[1,b)] Junwen CHEN[1,c)] Keiji YANAI[1,d)]

Abstract

Diffusion Transformer image editing is bottlenecked by VAE coupling and mask shape overfitting. We propose a Dual-Stream Discrete Geometric Gating Network that separates foreground generation from background context and uses STE-based token-level hard routing between background-anchored and subject-centric positional embeddings. Combined with blurry-mask augmentation, the model escapes rigid mask boundaries while preserving layout. Experiments on DGE-Bench and MagicBrush show stronger identity retention and structural plausibility than ACE++ and FLUX.1-Fill, producing sharper, more natural edits where continuous methods remain blurry or corrupted.

1. Introduction

Instruction-driven image editing must preserve the original scene, keep the edited subject semantically faithful, and allow natural geometric changes such as pose or shape variation. Existing latent diffusion pipelines struggle because image structure and appearance are tightly entangled in the VAE latent space [17]. We refer to this tension as the **VAE Coupling Problem**: stronger adherence to source latents protects layout, while weaker adherence enables geometry changes but risks semantic drift and background corruption.

Current methods mainly fail in two ways. Inpainting-oriented models such as FLUX.1-Fill [2] and ACE++ [11] treat the user mask as a hard geometric boundary, which causes **Mask Shape Overfitting**; edited objects inherit the rough mask contour instead of their natural outline. Training-free editing methods relax this constraint but often depend on sensitive structural control schedules [5], [12], [14]. We propose a **Dual-Stream Discrete Geometric Gating Network** that decouples background preservation from subject generation and uses token-level discrete routing rather than continuous interpolation. A Qwen-Image diagnostic shows that shared positional embeddings (PE) duplicate content, while cross-region attention masks suppress it, suggesting that PE acts as a context selector; our method changes how mask-derived subject tokens enter the geometry rather than removing the mask.

The key design choice is to reformulate local editing as *structural decoupling* rather than masked content replacement. One stream stabilizes the visible scene, while the other stream is allowed to regenerate subject geometry under discrete positional control. This makes the model better suited to edits that require non-rigid expansion, shrinkage, or semantic replacement without damaging nearby context. This perspective also clarifies why common control strategies fail in practice. If the model is forced to stay too close to the source latent, it preserves the scene but cannot execute meaningful geometric change. If it is allowed to deviate too freely, it may satisfy the text prompt while corrupting background structure or drifting away from the target subject identity. Our method addresses this trade-off at the architectural level rather than by hand-tuning an inference schedule.

Another practical requirement is data. Learning non-rigid local editing requires paired samples where the background remains highly consistent while the edited subject changes shape, pose, or category in a semantically valid way. Such examples are rare in raw web-scale corpora, which is why we combine the model design with an automated data curation pipeline and blurry-mask augmentation. Together, these components encourage the network to infer plausible geometry from context instead of copying the literal mask contour.

2. Related Work

2.1 Instruction-Guided Editing and Inpainting

Generative image editing has advanced rapidly with diffusion models [13], [16], [17]. Early text-guided methods [4], [7], [12] and recent instruction-driven systems such as ACE [10] and ACE++ [11] condition generation on multimodal inputs, while modern inpainters such as FLUX.1-Fill [2] excel at photorealistic masked completion. However, both families still struggle with robust structural control. Inpainters often inherit the coarse mask outline too literally, while instruction-editing models can preserve semantics but drift in geometry or background consistency. Our dual-stream design targets this gap by preserving the scene in one stream and regenerating editable subject structure in another.

[1] The Univ. of Electro-Communications
[a)] xiong-p@mm.inf.uec.ac.jp
[b)] yuan-h@mm.inf.uec.ac.jp
[c)] chen-j@mm.inf.uec.ac.jp
[d)] yanai@cs.uec.ac.jp

2.2 Attention and Layout Control

Several training-free approaches manipulate internal attention or positional information to steer structure [12], [14]. PosBridge [19] and RoPECraft [9], for example, transfer geometry by swapping or optimizing positional embeddings from a reference to a target. These strategies are effective for relatively rigid edits, but they still depend on continuous relaxation or heuristic schedules. When the target object needs substantial non-rigid change, soft interpolation between incompatible coordinate systems often causes ghosting, duplication, or blurry intermediate layouts. Our method instead introduces a learnable discrete gate, so each subject token can commit to one valid geometric regime rather than averaging between two incompatible ones.

3. Method

At a high level, a Main Stream preserves the source scene, a Subject Stream regenerates the editable object, and a lightweight gate selects the positional geometry for each subject token.

3.1 Motivation for Discrete Gating

Recent editing strategies that transplant positional information can reduce coupling, but they still rely on continuous control or manually chosen schedules [19]. This becomes problematic when rotary positional embeddings represent discrete spatial coordinates [18]: interpolating between incompatible geometric states often produces invalid intermediate layouts and blurry generations. Our method therefore uses hard geometric routing instead of soft mixing (Fig. 1).

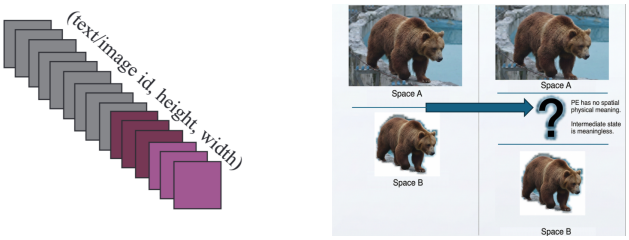


Fig. 1: Discrete vs. soft positional routing. Soft mixing creates invalid intermediate geometry, while hard token-level routing keeps each token on one valid coordinate system.

3.2 Dual-Stream Architecture

The network uses a *Main Stream* to preserve visible scene structure and a *Subject Stream* to generate the editable object (Fig. 2). The Subject Stream is initialized with independent noise, breaking the foreground-background coupling inherited from the original latent. Instead of treating the edited region as a hole to refill, the network treats it as a dedicated generation path conditioned on a stable scene representation.

In implementation terms, the visual sequence contains main-stream tokens, subject-stream tokens cropped from the editable region, and reference edit-image tokens. These tokens interact inside a shared multi-modal transformer, but only the subject branch is allowed to switch its geometric regime. This preserves compatibility with the pretrained backbone while giving the edited region a dedicated path for semantic regeneration.

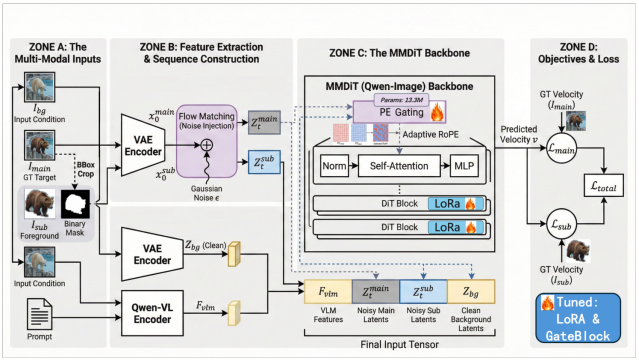


Fig. 2: Overall architecture. (1) The Main Stream preserves the source scene. (2) The Subject Stream starts from independent noise for editable content. (3) The gate routes subject-token PE between background-anchored and subject-centric coordinates.

3.3 STE-Based Routing

Subject tokens choose between a background-anchored geometry PE_{base} and a subject-centric geometry PE_{swap} (Fig. 3). We implement this with a lightweight STE gate [3]: $p_i = \sigma(h_\phi(H)_i)$, $G_i = \text{round}(p_i)$, and $PE_{\text{eff},i} = G_i PE_{\text{base},i} + (1 - G_i) PE_{\text{swap},i}$. The threshold is fixed at 0.5 and the final head is zero-initialized.

The gating network predicts token-wise probabilities from the unified visual representation and discretizes them into binary routing decisions in the forward pass. Boundary tokens can retain background coordinates for smooth fusion, while interior subject tokens can use subject-centric coordinates for geometric freedom.

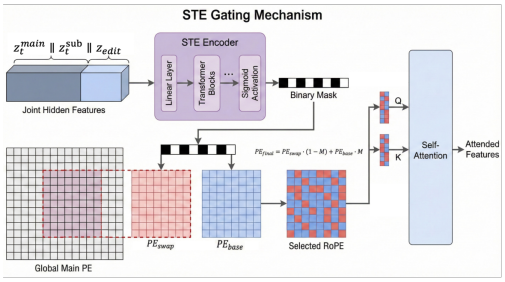


Fig. 3: STE gating predicts a binary decision G_i for each subject token, selecting background-anchored or subject-centric PE.

3.4 Data Pipeline and Training Signal

We construct training pairs with stable backgrounds and meaningful local geometric changes (Fig. 4). The pipeline uses a VLM and segmentation model to discover change regions, verify them, and filter out samples that fail strict background consistency checks [1], [6], [8]. We further apply **blurry-mask augmentation** so the network cannot succeed by copying mask contours literally; it must infer plausible shape from context and semantics.

The model is trained end-to-end with the standard denoising objective of the backbone while the STE gate is optimized through the straight-through gradient estimator. As a result, the routing module is supervised indirectly by generative quality: it learns to select the positional regime that mini-

mizes structural artifacts, background leakage, and subject corruption. This avoids introducing a separate geometric annotation signal while still producing interpretable binary decisions during inference.

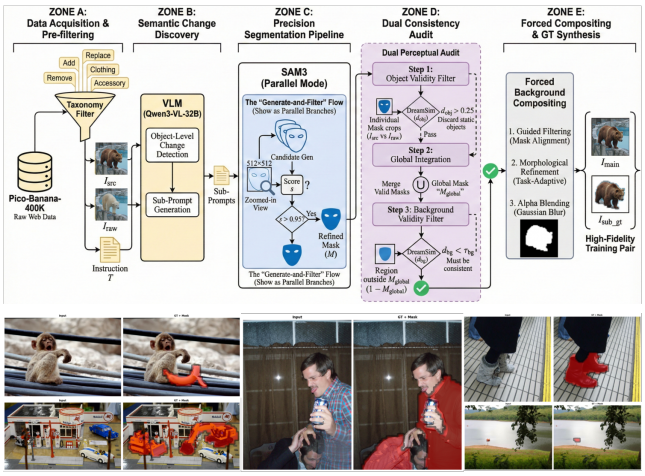


Fig. 4: Automatic data construction and representative pairs: discover change regions, verify background consistency, and train with mask perturbation to reduce contour copying.

4. Experiments

4.1 Setup

We evaluate on **DGE-Bench**, a curated benchmark derived from a large-scale image editing corpus, and on **MagicBrush** [20] for zero-shot transfer. We compare against two strong mask-based baselines, FLUX.1-Fill [2] and ACE++ [11]. Our main metrics are DINO for identity retention, DreamSim for structural plausibility, and text-image alignment scores for instruction following.

The model is built on Qwen-Image fine-tuning with LoRA-style adaptation and lightweight STE gating modules. We use independent subject-stream noise, classifier-free guidance rescaling, and latent blending for stable background preservation. Because MIRU space is limited, detailed hyperparameters and task-wise preprocessing are deferred to supplementary material.

DGE-Bench is designed to emphasize the cases where geometry matters most: subject addition, removal, and replacement under strict background preservation. MagicBrush evaluates whether the same model transfers to a different editing distribution without benchmark-specific tuning.

We also distinguish between global and local evaluation. Global metrics measure whether the final image remains coherent as a whole, while local metrics computed on the edited region reveal whether the model actually performs the intended transformation instead of merely producing visually smooth but semantically weak inpainting. This distinction is important because a baseline can look acceptable globally while still failing inside the edited subject region.

4.2 Main Results

Table 1 shows that the proposed network improves both identity preservation and structural naturalness. On DGE-Bench, it achieves the best DINO score and the lowest DreamSim distance, indicating that the edited subject remains faithful while adopting more plausible geometry.

Table 1: Quantitative comparison on **DGE-Bench**. We report both global image quality and local precision computed inside the edit bounding box.

Global Quality (DGE-Bench)				
Method	**DINO↑**	**DSim↓**	**Sig-I↑**	**Sig-T↑**
ACE++	0.982	0.081	0.954	0.705
FLUX.1-Fill	0.978	0.076	0.947	0.632
Ours	**0.988**	**0.054**	**0.965**	**0.776**

Local Precision (BBox)				
Method	**Sig-T↑**	**DINO↑**	**DSim↓**	**Sig-I↑**
ACE++	0.390	0.852	0.303	0.856
FLUX.1-Fill	0.262	0.807	0.321	0.835
Ours	**0.503**	**0.902**	**0.175**	**0.907**

Table 2: Zero-shot results on the MagicBrush test set. Final Turn and All Turn are reported separately, and Avg. Rank averages the 10 displayed metrics.

Final Turn					
Method	**L1↓**	**L2↓**	**CLIP-I↑**	**DINO↑**	**CLIP-T↑**
ACE++	0.081	0.032	0.891	0.817	0.307
FLUX.1-Fill	**0.075**	**0.025**	0.898	0.839	0.297
Q-Control	0.078	0.030	0.898	**0.840**	0.322
Ours	0.076	0.032	**0.907**	0.839	**0.324**

All Turn + Avg. Rank						
Method	**L1↓**	**L2↓**	**CLIP-I↑**	**DINO↑**	**CLIP-T↑**	**Rank↓**
ACE++	0.047	0.017	0.936	0.894	0.306	3.7
FLUX.1-Fill	0.044	**0.013**	**0.947**	**0.919**	0.300	2.1
Q-Control	0.044	0.016	0.943	0.912	**0.315**	2.2
Ours	**0.043**	0.017	0.946	0.910	**0.315**	**2.0**

The improvement is not limited to one metric family. Higher DINO indicates stronger identity retention, lower DreamSim reflects more perceptually plausible structure, and the text-alignment gains show that the model is not simply preserving the source image conservatively. Instead, it executes the requested edit while maintaining layout consistency. This is the behavior expected from successful structural decoupling.

Fig. 5 shows the qualitative effect of this design. The network is not forced to terminate generation at the original mask border; instead, it can expand, shrink, or reshape the subject according to the instruction while maintaining the surrounding structure. Table 1 (Local Precision) makes the same point quantitatively inside the edited region: the full model improves local text alignment by a large margin while also preserving better local identity and geometry than the two mask-bound baselines.

Table 2 shows strong zero-shot transfer on MagicBrush: the full model attains the best Avg. Rank while remaining competitive on reconstruction and DINO.

Fig. 5 shows the same pattern in direct comparison against ACE++ and FLUX.1-Fill. Baselines often remain tied to the supplied mask and produce truncated or blurry objects, while the proposed network generates more complete subjects with cleaner boundaries.

4.3 ICE-Bench Transfer

To test whether the same decoupling principle generalizes beyond DGE-Bench, we evaluate on ICE-Bench local editing



(a) Object addition and removal (b) Replacement and removal

Fig. 5: Comparison against strong mask-based baselines on challenging local edits.

Table 3: ICE-Bench results under modified and original prompt/input protocols. Because the original prompts may include placeholders such as `<Mask>` and `<SOURCE>`, the two blocks should be read as protocol-specific references rather than a strict horizontal comparison.

Orange data set (modified inputs/prompts)				
Method	**AES↑**	**IMG↑**	**PF↑**	**SRC↑**
ACE++	0.480	**0.575**	0.575	0.925
Q-Control	0.472	0.569	**0.588**	0.937
Ours	**0.494**	0.565	0.582	**0.942**

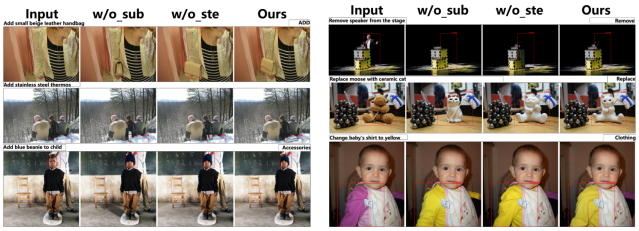
Original ICE-Bench inputs/prompts				
Method	**AES↑**	**IMG↑**	**PF↑**	**SRC↑**
ACE	0.480	0.487	**0.643**	0.916
ACE++	**0.481**	**0.577**	0.616	0.930
UltraEdit	0.455	0.458	0.442	**0.951**

tasks (Tasks 17 and 19–22) [15] using the latest protocol-specific evaluation.

Under the modified input/prompt protocol, Table 3 shows that our method achieves the best AES and SRC while remaining competitive on IMG and PF.

4.4 Ablation

Ablation confirms that discrete routing is the key ingredient rather than increased flexibility alone.



(a) Background leakage and weak localization (b) Sharper generation with full routing

Fig. 6: Ablation visualization. Without discrete routing, the edited region tends to remain blurry or entangled with background texture.

Fig. 6 and Table 4 show that the full model provides the best balance between structure preservation and instruction execution; without discrete routing, local text alignment can rise slightly while geometric plausibility deteriorates.

4.5 Efficiency

The gate adds only 13.31M parameters, far below a ControlNet-style copied branch (∼1.13B), and costs 0.539 versus 9.279 TFLOPs/step (∼17× less). LoRA adapters are fused before inference, so the main overhead comes from the Subject Stream itself, which adds roughly 30–40% latency and peak memory for a ∼25% edit region. This is substantially lighter than attaching a full auxiliary control branch while still giving the editable region its own geometric degrees of freedom.

Table 4: Ablation study on **DGE-Bench**. We evaluate both global image quality and local editing precision.

Global Quality				
Config	**DINO↑**	**DSim↓**	**Sig-I↑**	**Sig-T↑**
LoRA training	0.9858	0.0652	0.9572	0.7731
No STE	0.9786	0.0892	0.9463	**0.7840**
Full version	**0.9877**	**0.0538**	**0.9646**	0.7763

Local Precision (BBox)				
Config	**Sig-T↑**	**DINO↑**	**DSim↓**	**Sig-I↑**
LoRA training	0.4642	0.8695	0.2388	0.8811
No STE	0.4630	0.8750	0.2371	0.8849
Full version	**0.5030**	**0.9024**	**0.1750**	**0.9071**



Fig. 7: Subject-stream visualization across multiple examples. The proposed network generates concentrated, high-fidelity subject content instead of blurry copies.

4.6 Limitations

Localization is derived from the mask bounding box rather than a precise contour. This makes coarse masks usable in practice, but irregular masks can enlarge the editable support, and cluttered scenes or translucent objects remain challenging because clean foreground-background separation is inherently ambiguous. When stricter adherence to the original support is needed, mask-based blending can pull the fusion boundary back toward the user region.

5. Conclusion

The proposed dual-stream discrete geometric gating network reformulates image editing as a structural decoupling problem. By combining dual-stream generation with STE-based routing, the method preserves scene layout while allowing edited subjects to adopt natural shapes beyond rigid mask boundaries. The gains on DGE-Bench, MagicBrush, and protocol-specific ICE-Bench evaluation suggest that discrete geometric routing transfers beyond the training benchmark.

More broadly, the results indicate that explicit geometric routing is a practical alternative to soft positional interpolation when local edits require large non-rigid shape changes. This makes the approach promising for future instruction-guided editors that must balance strong scene preservation with genuinely flexible subject generation.

An immediate next step is to combine discrete routing with tighter region localization and more adaptive subject-stream tokenization. We expect these extensions to further reduce unnecessary computation while improving robustness on thin structures, transparent materials, and highly cluttered scenes.

References

[1] Bai, S., Cai, Y., Chen, R., Chen, K., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Chen, X., Huang, Q., Li, K. and Lin, Z.: Qwen3-VL Technical Report, *arXiv preprint arXiv:2511.21631* (2025).

[2] Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., M"uller, J., Podell, D., Rombach, R., Saini, H., Sauer, A. and Smith, L.: FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, *arXiv preprint arXiv:2506.15742* (2025).

[3] Bengio, Y., L'eonard, N. and Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation, *arXiv preprint arXiv:1308.3432* (2013).

[4] Brooks, T., Holynski, A. and Efros, A. A.: InstructPix2Pix: Learning to Follow Image Editing Instructions, *CVPR* (2023).

[5] Cao, M., Wang, Y., Sebe, N. and de Geus, D.: MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing, *ICCV* (2023).

[6] Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., R"adle, R., Afouras, T., Mavroudi, E., Xu, K. et al.: SAM 3: Segment Anything with Concepts, *ICLR* (2026).

[7] Couairon, G., Lefort, T., Kadkhodamohammadi, A., Clercq, O., Sigal, L. and Lalonde, J.-F.: DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance, *ICLR* (2023).

[8] Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T. and Isola, P.: DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data, *NeurIPS* (2023).

[9] Gokmen, A. B., Ekin, Y., Bilecen, B. B. and Dundar, A.: RoPECraft: Training-Free Motion Transfer with Trajectory-Guided RoPE Optimization on Diffusion Transformers, *NeurIPS* (2025).

[10] Han, Z., Jiang, Z., Pan, Y., Zhang, J., Mao, C., Xie, C., Liu, Y. and Zhou, J.: ACE: All-round Creator and Editor Following Instructions via Diffusion Transformer, *ICLR* (2025).

[11] Han, Z., Jiang, Z., Pan, Y., Zhang, J., Mao, C., Xie, C., Liu, Y. and Zhou, J.: ACE++: Instruction-Based Image Creation and Editing via Context-Aware Content Filling, *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* (2025).

[12] 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).

[13] Ho, J., Jain, A. and Abbeel, P.: Denoising Diffusion Probabilistic Models, *NeurIPS* (2020).

[14] 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).

[15] Pan, Y., He, X., Mao, C., Han, Z., Jiang, Z., Zhang, J. and Liu, Y.: ICE-Bench: A Unified and Comprehensive Benchmark for Image Creating and Editing, *ICCV*, pp. 16586–16596 (2025).

[16] Peebles, W. and Xie, S.: Scalable Diffusion Models with Transformers, *CVPR* (2023).

[17] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models, *CVPR* (2022).

[18] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W. and Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding (RoPE), *Neurocomputing*, Vol. 568, p. 127063 (2024).

[19] Xiong, P., Chen, J., Yuan, H. and Yanai, K.: PosBridge: Multi-View Positional Embedding Transplant for Identity-Aware Image Editing, *Proc. of the British Machine Vision Conference (BMVC)* (2025).

[20] Zhang, K., Mo, L., Chen, W., Sun, H. and Su, Y.: MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing, *NeurIPS* (2023).