RemEdit: Efficient Diffusion Editing with Riemannian Geometry

Lehigh University
Eashan Adhikarla, Brian D. Davison

WACV 2026 TUCSON, AZ 3/6 - 3/10

Introduction

The Trade-off: Fidelity vs Speed

- Controllable image generation faces a critical trade-off between semantic fidelity and inference speed.
- Standard $\boldsymbol{h}$-space methods use simple linear offsets ($\boldsymbol{h}' = \boldsymbol{h} + \Delta\boldsymbol{h}$) which implicitly treat the latent space as Euclidean. This pushes latent codes off the natural image manifold, destroying the subject's identity and creating artifacts.
- Furthermore, U-Net self-attention layers consume over **80%** of total GFLOPs during the forward pass, crippling real-time edge deployment.

The RemEdit Solution

- RemEdit eliminates these artifacts by navigating the $\boldsymbol{h}$-space as a curved Riemannian manifold, calculating exact geodesic paths.
- It introduces a novel task-specific attention pruning mechanism to slash inference times by aggressively pruning tokens without degrading unedited semantic regions.

✓ Riemannian Geodesic Navigation

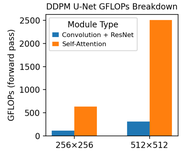
How can we edit semantically while respecting the data manifold, with high computational cost?

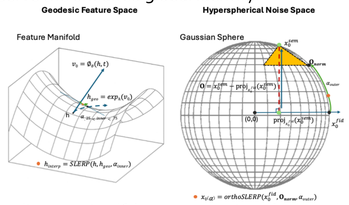
- RemEdit eliminates these artifacts by navigating the $\boldsymbol{h}$-space as a curved Riemannian manifold, calculating exact geodesic paths.
- It introduces a novel task-specific attention pruning mechanism to slash inference times by aggressively pruning tokens without degrading unedited semantic regions.

✓ Dual-SLERP Blending Control

How can we scale the geodesic offset Δh to control the edit strength without introducing artifacts?

- **Inner SLERP (Feature Space):** Interacts on the Riemannian manifold to scale the semantic strength of the edit.
- **Outer SLERP (Noise Space):** Extracts the orthogonal component of the edit $o = x_0^{sem} - proj_{x_0^{fid}}(x_0^{sem})$ and fuses it spherically to preserve the unedited global identity:





✓ Task-aware Attention Pruning

- A dynamic Pruning Head, conditioned on the *edit vector itself*, learns to retain only the tokens critical to the specific semantic goal.

How can we prune computationally expensive tokens without violating the core constraint of image editing: preserving the semantic content of unedited regions?

RemEdit Framework

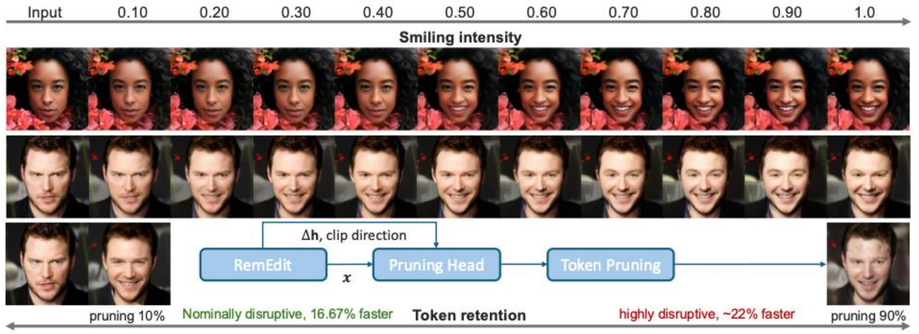


Fig. 1 RemEdit maintains semantic fidelity under aggressive token pruning; **90%** pruning is **~18%** faster yet remains visually acceptable, while **10%** pruning is virtually indistinguishable from the unpruned edit and still **~10%** faster.

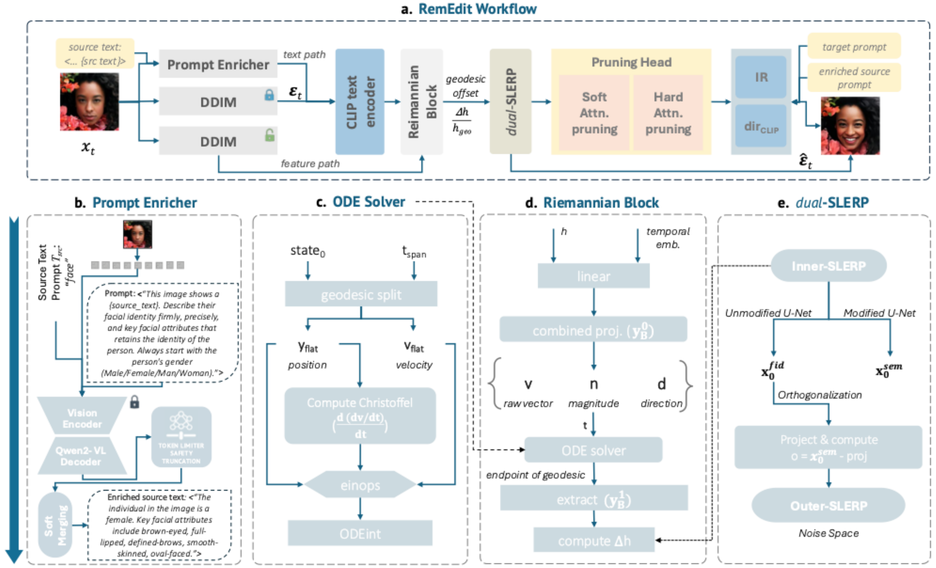


Fig. 2 Overview of our diffusion editing pipeline integrating exponential map for geodesic navigation, dual-SLERP interpolation for fidelity control, Qwen2-VL for prompt enrichment. The individual modules flow from top to bottom.

Experiments

Qualitative Comparisons

Best Efficiency

Method	Smiling S_{dir} ↑	Smiling Seg. Cons. (%) ↑	Sad S_{dir} ↑	Sad Seg. Cons. (%) ↑	Tanned S_{dir} ↑	Tanned Seg. Cons. (%) ↑	Inference Time (sec) ↓
StyleCLIP [31]	0.130	86.80	0.149	85.50	0.152	84.30	8.5
StyleGAN-NADA [9]	0.160	89.40	0.161	87.70	0.166	88.50	12.3
Diffusion-CLIP [19]	0.170	**93.70**	0.163	**89.93**	0.174	**92.85**	45.2
BoundaryDiffusion [53]	0.170	90.40	0.166	89.02	0.177	85.71	38.7
Asyrp [21]	0.190	87.90	0.159	88.90	0.177	89.31	28.9
Prompt-to-prompt (p2p) [12]	0.165	85.20	0.152	84.10	0.158	86.30	145.0
LEdits++ [2]	0.182	89.70	0.169	87.80	0.175	88.90	20.1
RemEdit	**0.1982**	92.41	**0.1792**	89.72	**0.1948**	92.18	**2.8**

Identity preservation

Tab. 1 Quantitative comparison of diffusion-based image editing methods on CelebA-HQ dataset. Results show (S_{dir}), Seg. Cons., and inference time across three facial attribute editing tasks (Smiling, Sad, Tanned).

Fig. 3 RemEdit is **~21x** and **~36x** times faster than asyrp and NT-P2P

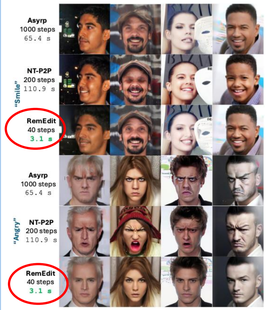
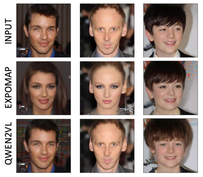
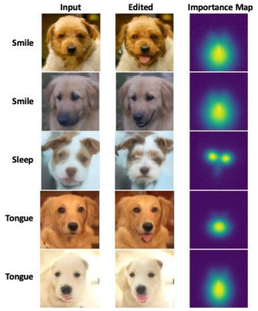


Fig. 4 Showing through a "makeup" task how the usage of Qwen2-VL for fine grained text injection corrects some of the failure cases.



- Vague prompts (e.g., "face" → "face with makeup") gives the model too much freedom, often resulting in unintended identity or gender shifts.

Fig. 5 Token importance heatmap per image for visualizing what the pruning head attends to.



High Speed Inference

- Pruning 20% of tokens drops inference time from 2.89s to 2.38s with negligible metric decay.
- Under aggressive 90% token pruning, the edit remains semantically correct and visually robust.

Zero-shot Performance

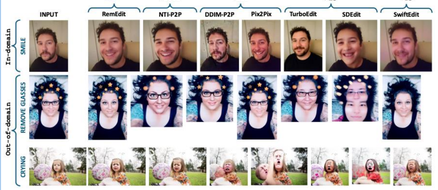


Fig. 6 Zero-shot comparison on RealEdit. Under identical settings, RemEdit ensures consistent attribute transfer while strictly preserving identity.

In Fig. 3, the model executes flawless edits in only 40 steps (3.1s), successfully manipulating difficult images where heavy baselines like Asyrp (1000 steps, 65.4s) and NT-P2P (200 steps, 110.9s) fail to preserve structure.

Conclusion

By integrating Riemannian geodesic navigation with task-aware attention pruning, RemEdit resolves the diffusion bottleneck; delivering high-fidelity, identity-preserving semantic control at accelerated inference speeds.

Name: Eashan Adhikarla
Address: 113 Research Drive
Lehigh University, Bethlehem, PA

Email: eaa418@lehigh.edu
Website: eashanadhikarla.github.io