7. Appendix

A. Problem Statement

The goal of diffusion-based image editing is to develop a framework that is simultaneously faithful to user intent, geometrically accurate, and computationally efficient. While prior work has shown that adding a semantic offset Δh to the bottleneck features in h-space is an effective manipulation strategy [21, 27], this approach presents a tripartite challenge that has not been holistically addressed.

First, the standard operation, $h' = h + \Delta h$, implicitly treats this space as Euclidean, which can lead to artifacts and off-manifold results. This leaves a foundational question unresolved:

What is the optimal formulation for the edit vector Δh that respects the intrinsic, non-Euclidean structure of the data manifold?

Second, even with a geometrically sound direction, applying the edit requires fine-grained control to preserve the subject's identity and avoid artifacts. This is made more difficult by the inherent ambiguity of simple text prompts, which can lead to unintended semantic shifts. This raises the question:

How can an edit be applied with tunable strength and guided by precise, context-aware instructions to ensure high fidelity?

Finally, any method that increases model sophistication to improve fidelity risks exacerbating the already significant computational cost of diffusion models, hindering practical application. This leads to the third challenge:

How can a high-fidelity editing process be made computationally efficient without compromising the quality and semantic consistency of the edit?

This paper addresses these interconnected problems. We hypothesize that a truly robust solution requires a unified framework: one that models h-space as a Riemannian manifold to derive a principled edit vector, introduces advanced blending and guidance mechanisms for high-fidelity control, and integrates a task-aware acceleration strategy to ensure efficiency. Our work aims to formalize and implement such a holistic framework.

B. Details on Task-aware Pruning

1. **Importance Scoring:** The input X is reshaped into a token sequence $T \in \mathbb{R}^{B\times N\times C}$, where $N = H \times W$. The `PruningHead` function $\mathcal{P}_\theta$ computes importance scores S:

$$S = \mathcal{P}_\theta(T, \mathbf{d}_{\text{edit}}) \in [0,1]^{B\times N} \tag{5}$$

2. **Index Selection:** Given a pruning ratio ρ, we keep $k = \lfloor N \cdot (1-\rho) \rfloor$ tokens. The indices $\mathcal{I}_{\text{keep}}$ of these tokens are selected:

$$\mathcal{I}_{\text{keep}} = \text{topk}_{\text{indices}}(S, k) \tag{6}$$

3. **Pruned Attention:** The query (Q), key (K), and value (V) projections are gathered using the selected indices to form pruned sets $Q_{\text{kept}}, K_{\text{kept}}, V_{\text{kept}} \in \mathbb{R}^{B\times k\times C}$. Attention is computed only on this reduced set:

$$A_{\text{pruned}} = \text{Softmax}\left(\frac{Q_{\text{kept}} K_{\text{kept}}^T}{\sqrt{C}}\right) V_{\text{kept}} \tag{7}$$

4. **Scattering and Output:** The attended features $A_{\text{pruned}} \in \mathbb{R}^{B\times k\times C}$ are scattered back into a zero tensor of the original size, $A_{\text{result}} \in \mathbb{R}^{B\times N\times C}$, at their original positions $\mathcal{I}_{\text{keep}}$. The final output X_{out} is computed via the residual connection with the output projection W_o:

$$X_{\text{out}} = X + W_o(A_{\text{result}}) \tag{8}$$

C. Exponential Map Architecture

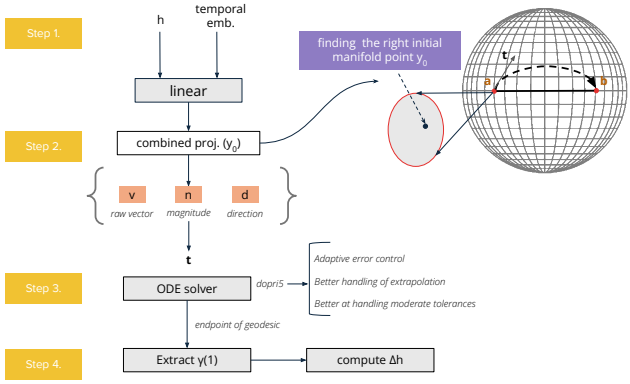

Figure 11. Our ExpoMap module learns the local geometry of the h-space manifold by estimating Christoffel symbols and solving the geodesic ODE using a numerical integrator (`Dopri5`). It takes as input a semantic direction and timestep, maps it through a learnable projection to obtain the initial tangent vector v_0, and integrates along the manifold to compute the final offset $\Delta h = \gamma(1) - h$. This enables geometry-aware edits without post-hoc Jacobian estimation.

D. Workflow details on Qwen2-VL

While CLIP embeddings have proven effective for guiding semantic edits, they often encode social or aesthetic biases and operate within a relatively narrow distribution of concepts. To mitigate this limitation, we enrich the CLIP embedding with additional context from a frozen Qwen2-VL model, effectively forming a broader `CLIP+Qwen2` joint

Table 5. Comparison of interpolation methods for diffusion model editing. While LERP is simple, it often produces artifacts by deviating from the data manifold. SLERP improves upon this by preserving the hyperspherical structure, and NoiseDiffusion further refines this for natural images by correcting the noise distribution. Our proposed dual-SLERP provides granular control by operating on both the learned feature manifold and the noise space, enabling disentangled control over edit strength and identity preservation.

Feature	LERP (Linear)	SLERP (Spherical) [39, 40]	NoiseDiffusion [52]	Dual-SLERP (Ours)
Manifold Consistency	Off-manifold	On hypersphere	Corrective (projects to valid noise distribution)	**Dual-manifold aware** (Riemannian & Hyperspherical)
Semantic Blending	Limited / Entangled	Fine-grained	Fine-grained (Corrected)	**Hierarchical & Disentangled**
Artifact Risk	High	Moderate (on natural images)	Lower (for natural images)	**Lowest** (due to orthogonality)
Control Mechanism	Coarse (single parameter)	Smooth & Structured	Multi-parameter (correction + interpolation)	**Disentangled control** (edit strength vs. fidelity)

embedding space. This expanded representation captures fine-grained, instance-specific visual semantics—especially useful for under-specified prompts like "face" or "young woman." While our Riemannian editing and dual-SLERP modules refine the geometry of noise and feature spaces, this multimodal enrichment ensures alignment at the textual level. Together, they allow RemEdit to perform edits that are both geometrically coherent and semantically precise, even under vague or biased attribute conditions.

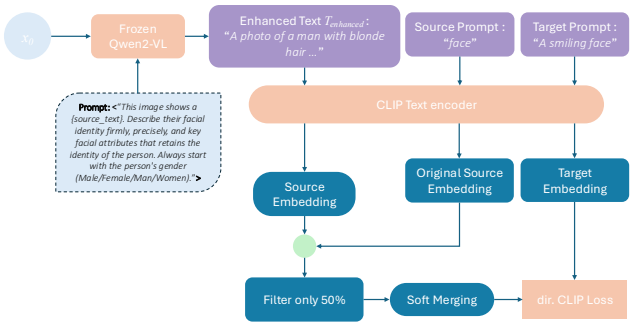

Figure 12. The source image x_0 and a descriptive prompt are fed into a frozen Qwen2-VL model to generate a detailed, instance-specific caption, T_{enhanced}. This enhanced text provides a richer source embedding, constraining the edit to be more faithful to the original image's core attributes.

E. Image Interpolation in Diffusion Models

To better situate our contribution, we provide a conceptual comparison of different interpolation strategies in Tab. 5. NoiseDiffusion [52] mitigates this by projecting noise back to a valid prior. Our proposed dual-SLERP mechanism goes further, being dual-manifold aware: inner SLERP operates in the learned Riemannian h-space, while outer SLERP modulates fidelity in noise space. This disentangled scheme allows separate control over semantic strength and identity preservation, enabling precise and robust image editing.