This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

RemEdit: Efficient Diffusion Editing with Riemannian Geometry

Eashan Adhikarla
Lehigh University
Bethlehem, Pennsylvania, USA
eaa418@lehigh.edu

Brian D. Davison
Lehigh University
Bethlehem, Pennsylvania, USA
bdd3@lehigh.edu

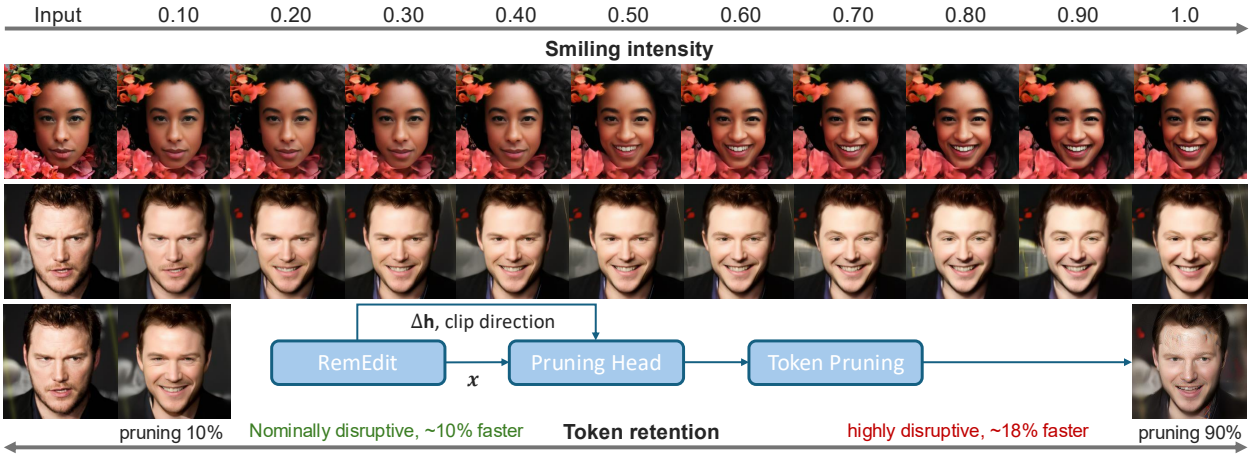

Figure 1. RemEdit maintains semantic fidelity under aggressive token pruning; **90%** pruning is ~**18%** faster yet remains visually acceptable, while **10%** pruning is virtually indistinguishable from the unpruned edit and still ~**10%** faster.

Abstract

Controllable image generation is fundamental to the success of modern generative AI, yet it faces a critical trade-off between semantic fidelity and inference speed. The ***RemEdit*** *diffusion-based framework addresses this trade-off with two synergistic innovations. First, for editing fidelity, we navigate the latent space as a Riemannian manifold. A mamba-based module efficiently learns the manifold's structure, enabling direct and accurate geodesic path computation for smooth semantic edits. This control is further refined by a dual-SLERP blending technique and a goal-aware prompt enrichment pass from a Vision-Language Model. Second, for additional acceleration, we introduce a novel task-specific attention pruning mechanism. A lightweight pruning head learns to retain tokens essential to the edit, enabling effective optimization without the semantic degradation common in content-agnostic approaches. RemEdit surpasses prior state-of-the-art editing frameworks while maintaining real-time performance under 50% pruning. Consequently, RemEdit establishes a new benchmark for practical and powerful image editing. Source code: github.com/eashanadhikarla/RemEdit.*

1. Introduction

The image synthesis capabilities of diffusion models have shifted the focus of generative AI research towards a more challenging frontier: precise and efficient image editing. While diffusion models [13, 41] have demonstrated SOTA synthesis quality, steering their generative process for precise, user-defined edits remains a challenge. Early approaches focused on text-guided manipulation through costly optimization or fine-tuning [19], while others leveraged inversion techniques to enable edits on real images [26]. A pivotal breakthrough was the discovery of the U-Net bottleneck's feature space, termed h-space, which was shown to be remarkably stable and semantically rich [21]. Concurrently, methods like [14, 17] began exploring the internal mechanisms of these models, editing cross-attention maps for more granular control. However, achieving high-fidelity edits that respect the image's original identity often requires intricate tuning. Furthermore, the significant computational cost of the iterative denoising process has motivated a parallel line of research into acceleration, from faster sampling schedules [13, 23] to early efforts in model pruning [7]. This has forced practitioners to choose between powerful but slow methods and faster

but less reliable alternatives. The computational bottleneck is particularly severe in the U-Net's self-attention layers, which, as shown in Fig. 4, can consume over 80% of the total GFLOPs in a single forward pass. While this motivates targeting these layers for acceleration, we argue that existing pruning methods are fundamentally misaligned with the goals of high-fidelity image editing.

To address this trade-off, we propose *RemEdit*, a new framework that jointly optimizes for geometric fidelity and computational efficiency. To our knowledge, *RemEdit* is the first to tackle both challenges for editing within Riemannian diffusion latent spaces. Our work is built on three core pillars of contribution:

1. **Accurate Geodesic Navigation of h-space.** We propose a novel and efficient method for traversing the semantic manifold by learning its local curvature directly. A lightweight Mamba-based predictor efficiently estimates the Christoffel symbols, which in turn define a learnable exponential map. This architecture allows us to solve the geodesic ODE, yielding direct, smooth, and geometrically faithful edits.
2. **High-Fidelity Semantic Control.** We achieve superior control and blending through two key innovations. A dual-SLERP mechanism provides principled, artifact-free interpolation within the manifold-aware latent space. This is guided by a goal-aware prompt enrichment strategy that uses a single pass of a Vision-Language Model (Qwen2-VL) to generate a nuanced, context-aware semantic direction.
3. **Task-Specific Acceleration.** We introduce a novel task-aware attention pruning method that dramatically accelerates inference without sacrificing edit quality. A lightweight "PruningHead," conditioned on the semantic goal of the edit, learns to preserve the most relevant tokens. This enables aggressive acceleration of irrelevant regions while maintaining integrity of the manipulation.

The synergy of these contributions results in a framework that is both more powerful and significantly more practical than prior work, closing the gap between theoretical robustness and real-world usability. More discussions on key challenges motivating this work are in **Appx. A**.

2. Related Work

2.1. Semantic latent structures in h-space

Kwon et al. first formalized the U-Net bottleneck (h-space) and showed that simple offsets enable reliable edits without retraining [21]. Follow-ups explored compositionality (InjectFusion) [16], PCA directions/semantic axes [10, 50], 3D consistency [10], and h-space inversion for style fidelity [15]. Surveys summarize this trend toward interpretable h-space control [30]. We build on this line, but argue that the stability often attributed to linearity is better explained by an underlying Riemannian structure of diffusion dynamics.

Recent works [28, 46] emphasize that h-space is more than a transient encoding, but is instead a structured, semantically aligned representation. Our work builds on this, proposing that its stability and linearity emerge from the Riemannian geometry intrinsic to diffusion dynamics.

2.2. Geodesic and Manifold-Aware Editing

Manifold assumptions pervade generative modeling, from classic Isomap [43]/LLE [34] to GAN latent geometry [8, 11, 34, 38, 43]. Recent works adapt this view to diffusion via autoencoding latents [24, 32], score Jacobian metrics [1, 29, 35] and shortest paths [30] (with extended discussion in **Appx. A**)In contrast, we learn the h-space connection and integrate an exponential map (mamba-based) to take geodesic steps efficiently, avoiding Jacobian estimation while keeping $O(N)$ complexity.

2.3. Semantic Guidance and Control

Training-free T2I editors localize changes either through attention control or implicit masking. Prompt-to-Prompt (P2P) constrains word-level changes by reusing cross-attention maps [12] and MasaCtrl [3]; LEDITS++ [2] combines fast inversion with implicit attention masking. We include both under identical prompts, inversion depths, and seeds in our comparisons and observe that RemEdit achieves stronger locality and identity preservation at comparable or faster runtime. These methods excel when text alone suffices, but are sensitive to phrasing and lack an explicit geometric constraint. Our approach complements them by constraining edits along learned h-space geodesics with dual-SLERP to balance edit strength and identity. Zero-shot editors built atop modern T2I pipelines further reduce overhead: Null-Text Inversion (optimization-based inversion) [26], Negative-Prompt Inversion (one-shot inversion) [25], Inversion-Free Editing (InfEdit) with modified sampler; no explicit inversion [45], and Lightning-Fast inversion via guided Newton steps [37]. These works achieve strong identity retention and speed but remain prompt-dependent. RemEdit is complementary: by operating in unconditional h-space with a learned connection and dual-SLERP, it enforces identity and locality by design under the same prompts.

2.4. Acceleration of Diffusion Models

Sampling-level accelerators (DDIM/ DPM-Solver/ distillation) reduce steps [13, 23, 36], while token-level methods prune attention with training-free (SiTo [49]/AT-EDM [44]), hybrid (DaTo [48]/CAT [5]), or merging (ToFu [20]) strategies. Training-based approaches (DyDiT [51]/LD-Pruner [4]) restructure modules. Other works like EffDiff [42] focus on real-time manipulation. Editing

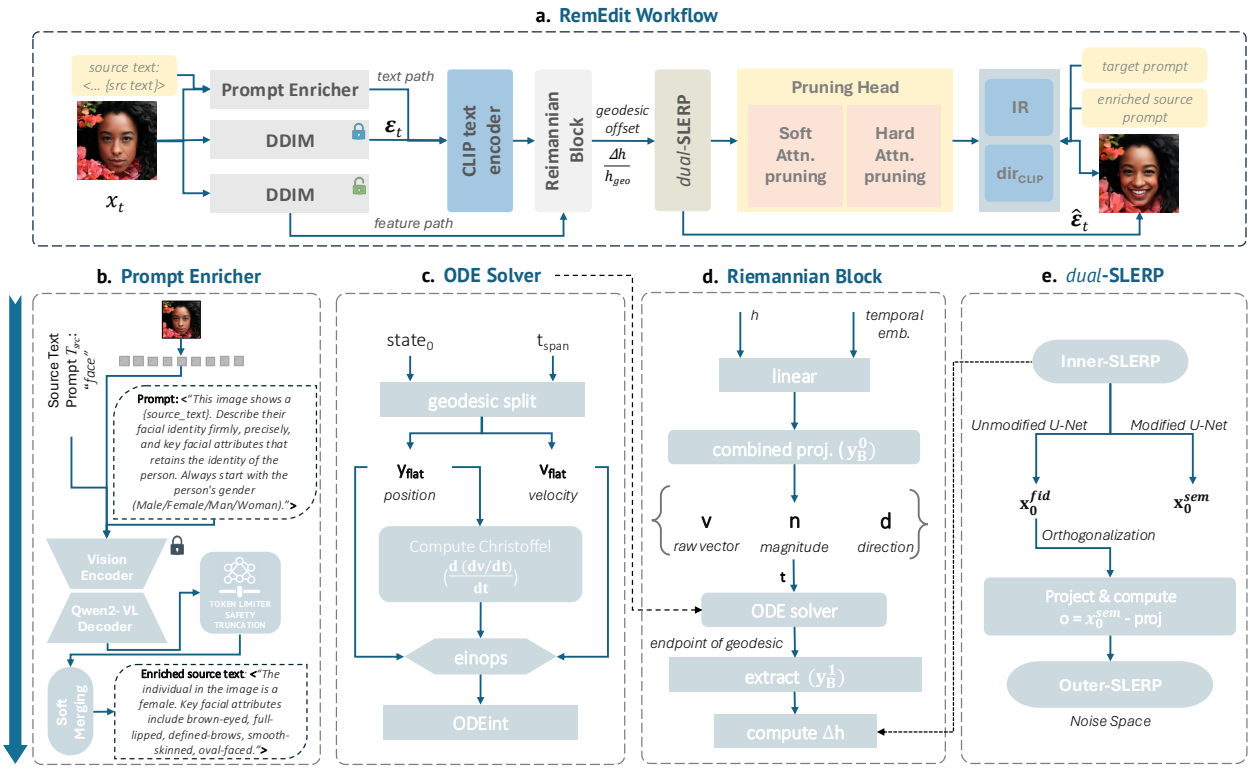

Figure 2. **RemEdit Architecture Diagram.** Overview of our diffusion editing pipeline integrating exponential map for geodesic navigation, dual-SLERP interpolation for fidelity control, Qwen2-VL for prompt enrichment. The individual modules flow from top to bottom.

imposes stricter locality than generation; content-agnostic pruning can harm identity in unedited regions. Despite their ingenuity, the aforementioned methods are designed and evaluated primarily for T2I. The task of semantic editing imposes a much stricter constraint: semantic consistency must be preserved across all unedited regions of the image. A content-agnostic pruner, by its nature, cannot distinguish between tokens that are globally redundant (e.g., a patch of blue sky) and tokens that are locally critical for maintaining the identity of an unedited object. Pruning the latter can lead to unacceptable degradation of the source content. We introduce a task-aware pruner tailored to edit locality.

3. Method

3.1. Riemannian Edit Framework

Our complete architecture is illustrated in Fig. 2.

A central challenge in semantic editing is that linear manipulations in latent space [21] often fail to preserve image realism. H-space methods typically apply a learned linear offset: $h' = h + \Delta h_{\text{linear}}$. Our results show that geodesic updates improve directional alignment and segmentation consistency with only a small increase in computation, later recovered through pruning. Adding a vector to a latent code can push it off the manifold of natural images[1], resulting in artifacts. This raises a critical question:

How can we edit semantically while respecting the data manifold, without high computational cost?

While prior methods acknowledge this geometric structure, they typically rely on post-hoc computations such as estimating metrics from score function Jacobians or solving expensive optimization problems [29, 35].

We estimate the full metric through an ODE solver to capture geodesics curvature and endpoints, while directly learning the manifold's connection.[2] This allows us to compute geodesics—the straightest possible paths on a curved surface—using the mathematical tool of the exponential map.

In Riemannian geometry, the exponential map, denoted $exp_p(v)$, takes a point p on a manifold and a tangent vector v at that point, and maps it to a new point on the manifold by traveling along the geodesic starting at p in the direction v for unit time. The path of this geodesic, $\gamma(t)$, is governed by the geodesic equation, which depends on the manifold's

[1] Linear offsets often leave the manifold, producing edits that no longer preserve the subject's identity.

[2] the rules that govern how a vector is transported along a curve.

Christoffel symbols Γ_{ij}^k:

$$\frac{d^2\gamma^k}{dt^2} + \mathbf{\Gamma}_{ij}^k \frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt} = 0 \tag{1}$$

The key insight is that by learning the Christoffel symbols, we can solve this second-order ODE to efficiently compute the exponential map. As illustrated in our architecture (see Appendix, Fig. 11), our Riemannian block implements this as follows: **(a) Initialization:** We form an initial point y_0 on the manifold by combining the input h-space feature h with the temporal embedding. **(b) Tangent Vector Prediction:** From y_0, we predict an initial tangent velocity vector v_0 using a linear layer followed by a smooth, norm-preserving tanh-based retraction. **(c) ODE Integration:** We learn the Christoffel symbols Γ using a lightweight Mamba network. We then solve the geodesic equation (Eq. 1), formulated as a first-order system, from $t = 0$ to $t = 1$ using a high-precision adaptive ODE solver. **(d) Offset Extraction:** The solution to the ODE at $t = 1$ gives the endpoint of the geodesic, $\gamma(1) = exp_h(v_0)$. We then define our geometrically-sound edit vector as the displacement along this path:

$$\Delta h = \exp_h(v_0) - h \tag{2}$$

This allows us to formulate the h-space update in a principled, on-manifold fashion, where the final edited feature is the geodesic endpoint itself:

$$h'_{\text{RemEdit}} = h + \Delta h_{\text{geo}} = \exp_h(v_0) \tag{3}$$

A breakdown of the exponential-map module and ODE solver configuration is included in the **Appx. C**.

3.2. High-Fidelity Control with Dual-SLERP

The geodesic offset Δh (Sec. 3.1) provides a robust, geometry-aware direction for editing. However, applying it naively as $h' = h + \Delta h$ offers no control over the edit's intensity and risks overpowering the original image's identity. This raises the next critical question:

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

We hypothesized that the latent features in diffusion models approximate a hyperspherical gaussian distribution. Therefore, interpolation should be performed spherically to remain on the data manifold and preserve statistical properties. Linear interpolation (Lerp) fails to preserve the norm of latent vectors and can "fall off" this hypersphere, leading to a collapse in quality. Spherical Linear Interpolation (SLERP) [39, 40], which travels along the great-circle arc between two points, is the natural choice for this space. We employ this insight in a novel dual-SLERP mechanism, illustrated in Fig. 3, for two distinct levels of control.

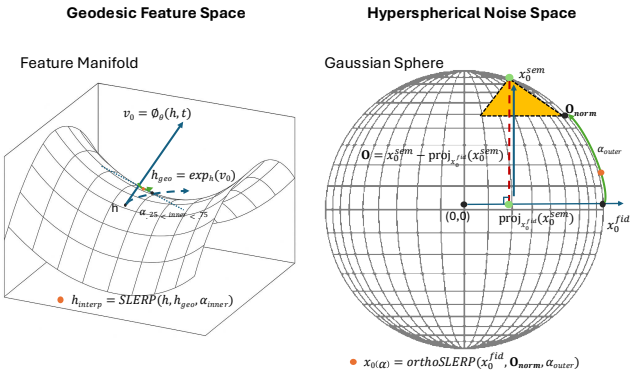

Figure 3. Dual-SLERP two-stage interpolation: **[Left]** Inner SLERP on the Riemannian feature manifold, blending the original feature h and the geodesically shifted feature $\exp_h(v_0)$ via the interpolation parameter α_{inner}. **[Right]** Outer orthogonal SLERP on the hyperspherical noise latent space, projecting the semantic prediction x_0^{sem} onto the fidelity latent x_0^{fid}, extracting the orthogonal component o, and interpolating between x_0^{fid} and o using α_{outer}, thus disentangling attribute edits from identity preservation.

Inner SLERP (Modulating Edit Strength). The first stage of our blending strategy occurs within the feature space, as depicted on the left in Fig. 3. After computing the full geodesic offset Δh using the Riemannian framework (Sec. 3.1), we SLERP between the original feature map h and the fully edited feature map $h_{geo} = h + \Delta h$. This interpolation is defined as: $h' = SLERP(h, h_{geo}, \alpha_{inner})$. The scalar $\alpha_{inner} \in [0, 1]$ acts as a precise knob[3] to control the strength of the semantic edit in feature space, before it is passed to subsequent layers of the U-Net.

Outer SLERP (Preserving Global Fidelity). The second stage occurs at the end of a denoising step to form the final prediction of the clean image, x_0. Our framework naturally produces two predictions: a fidelity-preserving latent vector, x_0^{fid}, from the original U-Net path, and a semantically edited one, x_0^{sem}, from our edited path. To robustly fuse these while preserving the original image's identity, we use an orthogonalized SLERP in noise space (shown in Fig. 3). We first project the semantic latent onto the fidelity latent to find the component of the edit that is orthogonal to the original identity: $o = x_0^{sem} - proj_{x_0^{fid}}(x_0^{sem})$. We then interpolate between the fidelity prediction and orthogonal component o:

$$x_0(\alpha_{outer}) = \Psi(x_0^{fid}, o, \alpha_{outer}) \tag{4}$$

A broader conceptual comparison of interpolation approaches is in **Appx. E**.

3.3. Goal-Aware Prompt Enrichment

A core challenge faced by prior methods is in text-guided editing ambiguity, as concise prompts often lack the speci-

[3] decoupling edit's direction from its magnitude (controlled by SLERP)

ficity to prevent unrelated attribute changes. For an attribute like makeup, a simple source-target pair such as "face" → "face with makeup" provides insufficient context. The model, seeking to satisfy this minimal constraint, is free to explore neighboring manifolds in the data distribution that may introduce unintended changes, such as altering the perceived gender of the subject. This necessitates providing richer, instance-specific context to the edit direction without adding significant computational overhead and w/o re-labeling text for image pair. We call this a "no free lunch" problem, which we solve by introducing a lightweight, single-pass prompt enrichment stage using a pretrained Vision-Language Model (Qwen2-VL). Instead of using a generic source text for all images, we first generate a detailed caption for the specific source image x_0 (details shown in Appendix Fig. 12). This approach effectively grounds the edit (shown in Fig. 6) in the specific visual context of the input image, mitigating unwanted semantic shifts by narrowing the model's exploratory freedom. Further details are in **Appx. D**.

3.4. Task-Specific Attention Pruning

Our framework thus far achieves high-fidelity, controllable editing. However, like all modern diffusion models, it remains computationally expensive. The self-attention blocks are the primary bottleneck, consuming substantial resources: over 4× the GFLOPs of convolution and ResNet blocks at 256×256 resolution, and over 8× at 512×512 resolution. This makes them ideal for optimization.

While token pruning has emerged as a viable acceleration strategy, existing methods are designed for general image generation. This raises a critical question:

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

For pure generation, minor inconsistencies or artifacts may be acceptable. For editing, they represent failure. If editing a "smile" onto a face alters the subject's hair or the background, the edit has failed. This implies that a pruner for editing cannot be content-agnostic; it must be explicitly aware of the editing task. A successful pruning strategy for semantic editing must be conditioned on the edit vector itself. By providing the pruning mechanism with information about the desired semantic change, it can learn to distinguish between tokens that are globally redundant and tokens that are critical for preserving the identity of unedited regions, even if those regions are low-variance.

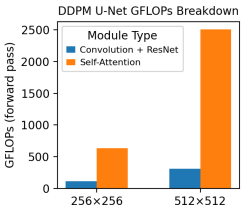

Figure 4. GFLOPs breakdown of the DDPM U-Net at 256^2 and 512^2.

We introduce a novel **task-aware attention pruning** mechanism. Instead of relying on static heuristics, we train a lightweight neural network to dynamically predict token importance, conditioned on both the input features and the semantic goal of the edit.

The training objective is a weighted sum of two competing losses, designed to balance output quality with computational savings: $\mathbf{L} = \mathcal{L}_{\text{fidelity}} + \lambda_{\text{sparsity}} \cdot \mathcal{L}_{\text{sparsity}}$. First, a Fidelity Loss teaches the pruner what to keep by forcing its output to match the original, unpruned model's output. Second, a Sparsity Loss encourages the model to be efficient by pushing it to prune as many tokens as possible. During inference, this trained pruner $\mathcal{P}_\theta$ performs non-differentiable hard pruning.

Given input features $X \in \mathbb{R}^{B\times C\times H\times W}$ and a task-defining semantic vector $\mathbf{d}_{\text{edit}} \in \mathbb{R}^{D_{\text{clip}}}$, we reshape X into a token sequence $T \in \mathbb{R}^{B\times N\times C}$, where $N = H \cdot W$. A learned pruning function $\mathcal{P}_\theta$ estimates token importance:

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

Top-k indices are selected with $k = \lfloor N \cdot (1-\rho) \rfloor$, and attention is computed only over retained tokens:

$$A = \text{Softmax}\left(\frac{Q_{\text{kept}} K_{\text{kept}}^{\top}}{\sqrt{C}}\right) V_{\text{kept}}$$

The attention output is restored to the full token space and projected, yielding the final output X_{out}:

$$X_{\text{out}} = X + W_o(\text{Scatter}(A, \mathcal{I}_{\text{keep}}))$$

See **Appx. B** for more detail on pruning.

4. Experiments

We conduct a comprehensive set of experiments to validate *RemEdit*: the accuracy, the effectiveness of our semantic control mechanisms, and the efficiency of our task-aware pruner. We evaluate our method on several challenging benchmarks, including CelebA-HQ [18], LSUN-Church [47], and AFHQ-Dog [6] each over 256×256.

4.1. Implementation Details

Our framework is trained in an efficient few-shot manner. Unlike baseline methods such as Asyrp [21] which require multiple iterations over 1000+ images, RemEdit achieves superior results by training on only 500 images for two-shot setup (n=2). We observe that a single training iteration is sufficient for the Riemannian block to capture the high-level semantics of an edit, though training for approximately three iterations significantly refines the output by removing minor artifacts in detailed regions like eyebrows and hair texture. All experiments are conducted using the DDIM architecture on 256×256 resolution images unless otherwise specified. We report wall-clock end-to-end inference measurements for the pipeline.

Table 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).

Method	Smiling		Sad		Tanned		Inference Time
	S_{dir} ↑	Seg. Cons. (%) ↑	S_{dir} ↑	Seg. Cons. (%) ↑	S_{dir} ↑	Seg. Cons. (%) ↑	(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**

Table 2. Quantitative evaluation on the **LSUN-Church** dataset.

Method	Department Store		Ancient		Red Brick	
	S_{dir} ↑	Seg. Cons. (%) ↑	S_{dir} ↑	Seg. Cons. (%) ↑	S_{dir} ↑	Seg. Cons. (%) ↑
Diffusion CLIP [19]	0.1300	54.50	0.1976	**64.82**	0.2085	65.83
BoundaryDiffusion [53]	0.1866	56.72	0.2034	60.13	0.2112	66.10
Asyrp [21]	0.1932	57.62	0.2087	62.65	0.2170	67.42
Prompt-to-prompt (p2p) [12]	0.1820	53.40	0.1890	58.75	0.1985	61.30
LEdits++ [2]	0.1895	58.85	0.2045	59.20	0.2130	64.95
RemEdit	**0.1959**	**59.03**	**0.2193**	63.91	**0.2200**	**68.20**

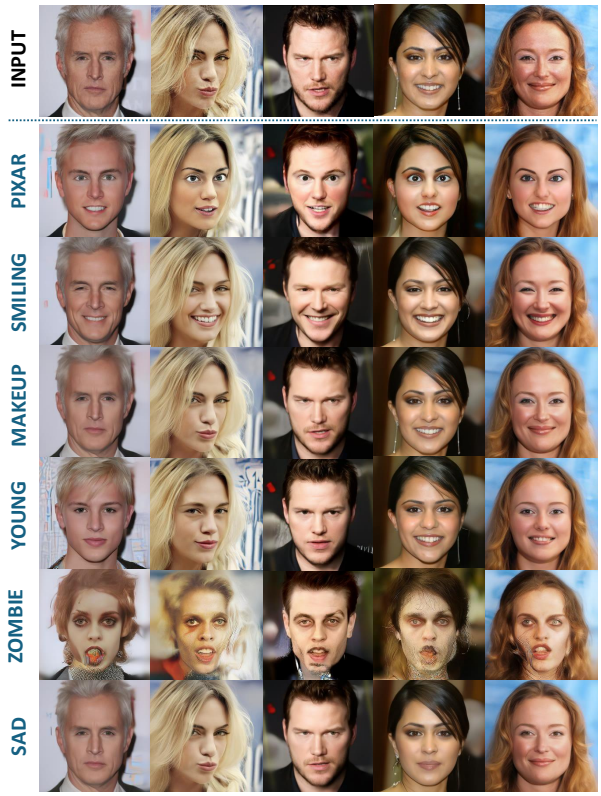

Figure 5. Editing results of RemEdit on CelebA-HQ dataset where attributes {"Sad", "Smiling", "Makeup", "Young"} are Human in-distribution and {"Zombie", "Pixar"} are Human out-of-distribution.

Table 3. Reconstruction fidelity of *RemEdit* under different inversion depths t_0, forward DDIM steps S_{for}, and generative refinement steps S_{gen}. Each cell reports MAE / LPIPS / SSIM. Lower MAE/LPIPS and higher SSIM indicate better fidelity.

	S_{gen} = 6	S_{gen} = 40	S_{gen} = 500
t_0 = 300			
S_{for} = 6	0.047 / 0.185 / 0.732	0.061 / 0.221 / 0.704	0.063 / 0.224 / 0.694
S_{for} = 40	0.027 / 0.110 / 0.863	0.023 / 0.091 / 0.891	0.023 / 0.086 / 0.895
S_{for} = 500	0.024 / 0.095 / 0.885	0.020 / 0.073 / 0.914	0.019 / 0.065 / 0.923
t_0 = 450			
S_{for} = 6	0.055 / 0.208 / 0.673	0.073 / 0.255 / 0.655	0.077 / 0.260 / 0.643
S_{for} = 40	0.031 / 0.128 / 0.827	0.025 / 0.100 / 0.880	0.024 / 0.093 / 0.885
S_{for} = 500	0.028 / 0.108 / 0.862	0.024 / 0.076 / 0.910	0.020 / 0.068 / 0.919
t_0 = 600			
S_{for} = 6	0.084 / 0.283 / 0.501	0.101 / 0.325 / 0.564	0.106 / 0.330 / 0.552
S_{for} = 40	0.047 / 0.175 / 0.706	0.029 / 0.120 / 0.852	0.028 / 0.108 / 0.862
S_{for} = 500	0.041 / 0.147 / 0.778	0.024 / 0.087 / 0.893	0.022 / 0.076 / 0.907

Table 4. Ablation study on the core components of RemEdit The results show that each component provides a distinct benefit. **Time (s) is measured end-to-end, including VLM captioning, inversion, geodesic ODE solving, and decoding.**

Method Configuration	Dir. CLIP ↑	Seg. Cons. (%) ↑	Time (s) ↓
Euclidean offset (h-space baseline)	0.190	87.9	**1.82**
+ Geodesic Navigation (Ours)	0.196	89.5	2.76
Linear Interpolation (no SLERP)	0.188	86.7	2.74
Inner-only SLERP	0.191	88.2	2.78
Outer-only SLERP	0.192	88.4	2.77
Dual-SLERP Blending (Ours, unpruned)	**0.198**	**92.4**	2.89
Full RemEdit + Pruning (ρ = 0.2)	0.192	90.1	2.38
Full RemEdit + Pruning (ρ = 0.5)	0.184	89.5	2.31

4.2. Analysis of Geometric Semantic Control

4.2.1. Quantitative Analysis

We first evaluate RemEdit's core editing, highlighting the synergy between geodesic navigation and dual-SLERP blending. Following recent SOTA methods [19, 21, 53],

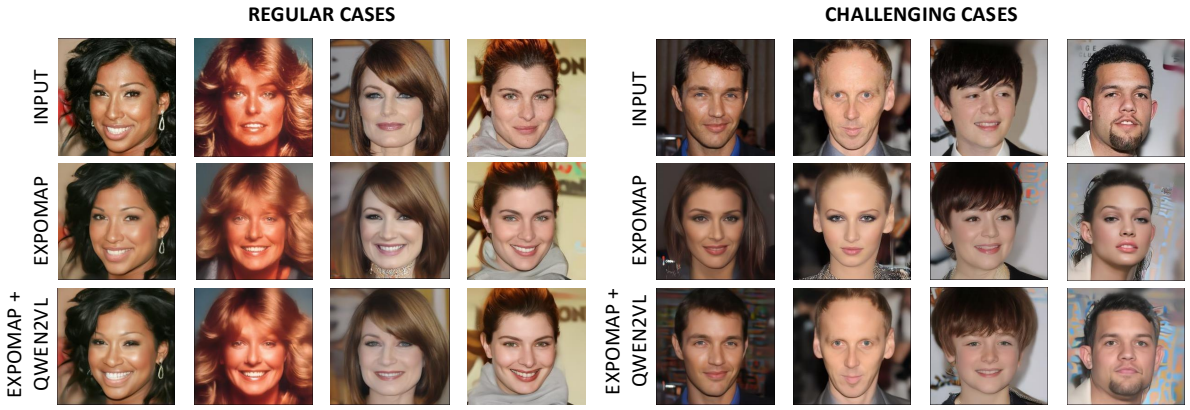

Figure 6. Showing through a "makeup" task how the usage of Qwen2 VL for fine grained text injection corrects some of the failure cases.

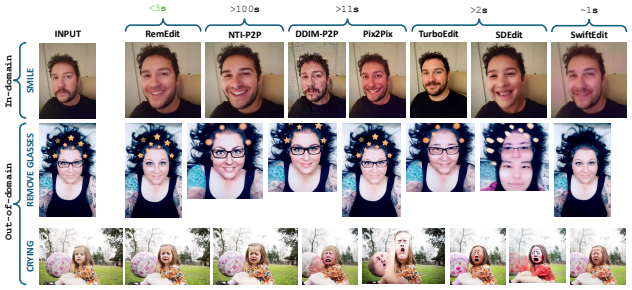

Figure 7. Zero shot qualitative comparison on RealEdit benchmark dataset. Identical prompts, inversion depth, and seeds. RemEdit achieves consistent attribute transfer with minimal collateral change while preserving identity.

Tab. 1 and 2 compare RemEdit on CelebA-HQ and LSUN-Church using directional CLIP similarity (S_{dir}) [33] for semantic alignment and segmentation consistency [22] for identity preservation. We chose 250 samples per task attribute. RemEdit consistently outperforms prior work, including Asyrp, across attributes like Smiling," Sad," and "Tanned." This highlights the effectiveness of our exponential map: by solving a geodesic ODE, the Riemannian block ensures semantically precise edits, while dual-SLERP preserves identity.

4.2.2. Qualitative Analysis

Qualitative results across attributes are shown in Fig. 5. RemEdit successfully manipulates attributes from simple expressions ("Sad," "Smiling") to more complex transformations ("Zombie," "Pixar") while maintaining a high degree of photorealism and identity. Fig. 10 provides a direct comparison against the Asyrp baseline. Even when running with only 40 denoising steps, RemEdit produces results that are qualitatively superior to Asyrp running at 1000 steps, successfully editing difficult cases where the baseline struggles. This highlights the efficiency and robustness of our geometric approach. As seen in Fig. 6, Qwen2-VL improves clarity on underspecified prompts, whereas regular cases and the underlying geometry remain stable.

4.2.3. Ablation Study

Reconstruction Quality. Tab. 3 investigates the reconstruction fidelity of our model under various inversion and generation settings. We analyze the impact of the inversion depth (t_0), the number of forward DDIM steps (S_{for}), and the number of generative steps (S_{gen}). The results, measured by MAE, LPIPS, and SSIM, show that a deeper inversion (higher (t_0) and a sufficient number of generative steps are crucial for high-fidelity reconstruction, confirming that our method adheres to the expected behavior of diffusion-based inversion pipelines.

Controlled Image Generation. We also conduct an ablation study on individual contributions, with results presented in Tab. 4. We start with a strong baseline implementing an Asyrp-style h-space edit. We then incrementally add our proposed modules: first the geodesic navigation, then the dual-SLERP blending to form the full, unpruned RemEdit model, and finally our task-specific pruning at two different ratios (ρ = 0.2 and ρ = 0.5). We evaluate each configuration on directional CLIP similarity S_{dir}, segmentation consistency SC, and inference time per image.

4.3. Analysis of Task-Specific Acceleration

As shown in Fig. 1, RemEdit maintains fidelity even under aggressive pruning. At a 10% pruning ratio, the edited output is visually indistinguishable from the unpruned version, while still providing a 10% speed-up. Even at an aggressive 90% pruning ratio, which yields an ~18% speed-up, the edit remains semantically correct and visually acceptable. This degradation is a direct result of our task-aware approach. Token retention patterns are illustrated in Fig. 8. The importance maps clearly show that the model learns to focus on the semantically relevant regions (e.g., the mouth for "Smiling," the entire face for "Sleep") while correctly identifying the background as prunable.

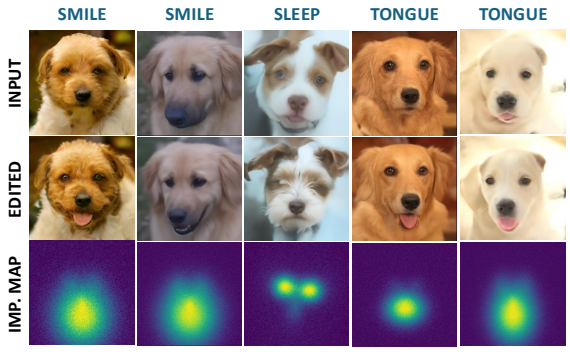

Figure 8. Token importance map heatmap per image for visualizing what the pruning head attends to.

Figure 9. We also demonstrate the semantic control capability of RemEdit through multi-attribute editing.

5. Results and Discussion

Our experimental results validate the effectiveness of the RemEdit framework. The quantitative comparisons in Tables 1 and 2 demonstrate a clear improvement over existing methods in semantic alignment and identity preservation.

The ablation study in Tab. 4 provides a deeper insight into the trade-offs involved. Our contributions on the geodesic approach and dual-SLERP blending modules significantly improve both the S_{dir} score (from 0.190 to 0.198) and segmentation consistency (from 87.9% to 92.4%), confirming their role in enhancing edit quality and fidelity. This gain, however, comes at the cost of increased inference time, rising from 1.82s for the baseline to 2.89s for the full unpruned RemEdit, due to the ODE-solving computation.

This is where our task-specific pruning demonstrates its value. By pruning just 20% of the tokens, we reduce inference time to $2.38s$ while maintaining performance metrics close to the full model. Even with an aggressive 50% pruning ratio, the model remains competitive with the baseline in quality (S_{dir} = 0.184 vs 0.190) while being substantially faster than the unpruned version. It confirms that our pruning method successfully makes our geometric editing framework more practical, closing the performance gap with simpler methods while retaining a clear advantage in edit quality.

Furthermore, the token-importance maps in Fig. 8 confirm our task-aware pruning strategy by focusing on relevant regions (mouth, tongue, face) and assign lower background weights; quality is stable with 20–50% pruning. Fig. 9 shows multi-attribute examples (Tanned+Smiling, Makeup+Smiling) with no interference or identity change observed here. On difficult cases as in Fig. 10, RemEdit reaches edit in ˜40 steps (˜3.1s), Asyrp ˜1000 (˜65.4s), NT-P2P ˜200 (˜110.9s); panels show differences in identity and locality at the cost of inference speed.

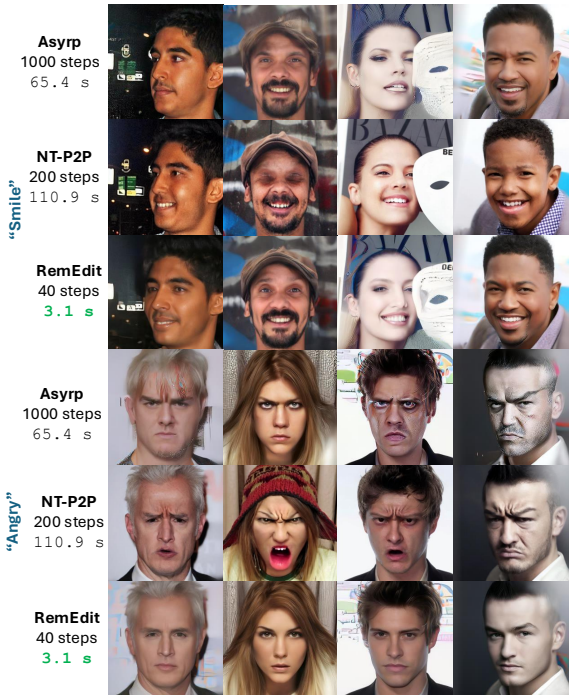

Figure 10. Solving difficult cases in only 40 steps.

6. Conclusion

In this work, we introduced RemEdit, a novel framework for high-fidelity and efficient controlled image generation. We addressed the fundamental trade-off between accuracy and speed by making contributions on three fronts. First, we proposed a new way to navigate the semantic h-space by modeling it as a Riemannian manifold and solving for geodesic paths, which we showed improves semantic quality. Second, we introduced a dual-SLERP and Qwen2-VL based text-injection blending mechanism for fine-grained control via textual and image features over edit strength and identity preservation, further boosting fidelity. Finally, we developed a task-aware attention-pruning method that makes our geometrically sophisticated approach computationally practical, significantly accelerating inference while preserving the quality of the edit.

While RemEdit demonstrates high semantic fidelity and practical efficiency, it relies on pretrained models such as Qwen2-VL and CLIP. Future work could explore learned, domain-specific guidance for out-of-distribution tasks.

References

[1] Simone Azeglio and Arianna Di Bernardo. What's inside your diffusion model? a score-based riemannian metric to explore the data manifold. *arXiv preprint arXiv:2505.11128*, 2025. 2

[2] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8861–8870, 2024. 2, 6

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2

[4] Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2024. 2

[5] Xinle Cheng, Zhuoming Chen, and Zhihao Jia. Cat pruning: Cluster-aware token pruning for text-to-image diffusion models. *arXiv preprint arXiv:2502.00433*, 2025. 2

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 5

[7] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Advances in Neural Information Processing Systems*, 2023. 1

[8] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. 2

[9] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 6

[10] René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE, 2024. 2

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020. 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 6

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[14] Zhiyuan Hu, Jiancheng Lyu, Dashan Gao, and Nuno Vasconcelos. Pop: Prompt of prompts for continual learning. *arXiv preprint arXiv:2306.08200*, 2023. 1

[15] Ayodeji Ijishakin, Ming Liang Ang, Levente Baljer, Daniel Chee Hian Tan, Hugo Laurence Fry, Ahmed Abdulaal, Aengus Lynch, and James H Cole. H-space sparse autoencoders. In *Neurips Safe Generative AI Workshop 2024*, 2024. 2

[16] Jaeseok Jeong, Mingi Kwon, and Youngjung Uh. Training-free content injection using h-space in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5151–5161, 2024. 2

[17] KJ Joseph, Prateksha Udhayanan, Tripti Shukla, Aishwarya Agarwal, Srikrishna Karanam, Koustava Goswami, and Balaji Vasan Srinivasan. Iterative multi-granular image editing using diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8107–8116, 2024. 1

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 5

[19] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 1, 6

[20] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1383–1392, 2024. 2

[21] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 1, 2, 3, 5, 6, 11

[22] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5549–5558, 2020. 7

[23] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022. 1, 2

[24] Zeyu Lu, Chengyue Wu, Xinyuan Chen, Yaohui Wang, Lei Bai, Yu Qiao, and Xihui Liu. Hierarchical diffusion autoencoders and disentangled image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5374–5383, 2024. 2

[25] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2063–2072. IEEE, 2025. 2

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6038–6047, 2023. 1, 2

[27] Toan Nguyen, Kien Do, Duc Kieu, and Thin Nguyen. h-edit: Effective and flexible diffusion-based editing via doob's h-transform. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28490–28501, 2025. 11

[28] Jinhyeong Park, Muhammad Shaheryar, Seangmin Lee, and Soon Ki Jung. Navigating h-space for multi-attribute editing in diffusion models. In *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 1129–1133. IEEE, 2025. 2

[29] Sung Woo Park, Hyomin Kim, Kyungjae Lee, and Junseok Kwon. Riemannian neural sde: Learning stochastic representations on manifolds. *Advances in Neural Information Processing Systems*, 35:1434–1444, 2022. 2, 3

[30] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36: 24129–24142, 2023. 2

[31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. 6

[32] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7

[34] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000. 2

[35] Shinnosuke Saito and Takashi Matsubara. Image interpolation with score-based riemannian metrics of diffusion models. *arXiv preprint arXiv:2504.20288*, 2025. 2, 3

[36] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

[37] Dvir Samuel, Barak Meiri, Haggai Maron, Yoad Tewel, Nir Darshan, Shai Avidan, Gal Chechik, and Rami Ben-Ari. Lightning-fast image inversion and editing for text-to-image diffusion models. *arXiv preprint arXiv:2312.12540*, 2023. 2

[38] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. 2

[39] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985. 4, 12

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4, 12

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[42] Nikita Starodubcev, Dmitry Baranchuk, Valentin Khrulkov, and Artem Babenko. Towards real-time text-driven image manipulation with unconditional diffusion models. *arXiv preprint arXiv:2304.04344*, 2023. 2

[43] Joshua Tenenbaum. Mapping a manifold of perceptual observations. *Advances in neural information processing systems*, 10, 1997. 2

[44] Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K Jha, and Yuchen Liu. Attention-driven training-free efficiency enhancement of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16080–16089, 2024. 2

[45] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *arXiv preprint arXiv:2312.04965*, 2023. 2

[46] Zizheng Yang, Hu Yu, Bing Li, Jinghao Zhang, Jie Huang, and Feng Zhao. Unleashing the potential of the semantic latent space in diffusion models for image dehazing. In *European Conference on Computer Vision*, pages 371–389. Springer, 2024. 2

[47] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. 5

[48] Evelyn Zhang, Bang Xiao, Jiayi Tang, Qianli Ma, Chang Zou, Xuefei Ning, Xuming Hu, and Linfeng Zhang. Token pruning for caching better: 9 times acceleration on stable diffusion for free. *arXiv preprint arXiv:2501.00375*, 2024. 2

[49] Evelyn Zhang, Jiayi Tang, Xuefei Ning, and Linfeng Zhang. Training-free and hardware-friendly acceleration for diffusion models via similarity-based token pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9878–9886, 2025. 2

[50] Zijian Zhang, Luping Liu, Zhijie Lin, Yichen Zhu, and Zhou Zhao. Unsupervised discovery of interpretable directions in h-space of pre-trained diffusion models. *arXiv preprint arXiv:2310.09912*, 2023. 2

[51] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. *arXiv preprint arXiv:2410.03456*, 2024. 2

[52] PengFei Zheng, Yonggang Zhang, Zhen Fang, Tongliang Liu, Defu Lian, and Bo Han. Noisediffusion: Correcting noise for image interpolation with diffusion models beyond spherical linear interpolation. In *The Twelfth International Conference on Learning Representations*, 2024. 12

[53] Ye Zhu, Yu Wu, Zhiwei Deng, Olga Russakovsky, and Yan Yan. Boundary guided learning-free semantic control with diffusion models. *Advances in Neural Information Processing Systems*, 36:78319–78346, 2023. 6