SphereEdit: Spherical Semantic Editing in Diffusion Models

Salamata Konate[1,2], Hassan Hamidi[2], Elham Dolatabadi [1,2],Frank Rudzicz [2,3], Laleh Seyyed-Kalantri[1,2,4]
[1]York University, Canada
[2]Vector Institute, Canada
[3]Dalhousie University, Canada
[4]CIFAR Solution Network Member
{skonate, hhamidi, edolatab, lsk}@yorku.ca, fr591304@dal.ca

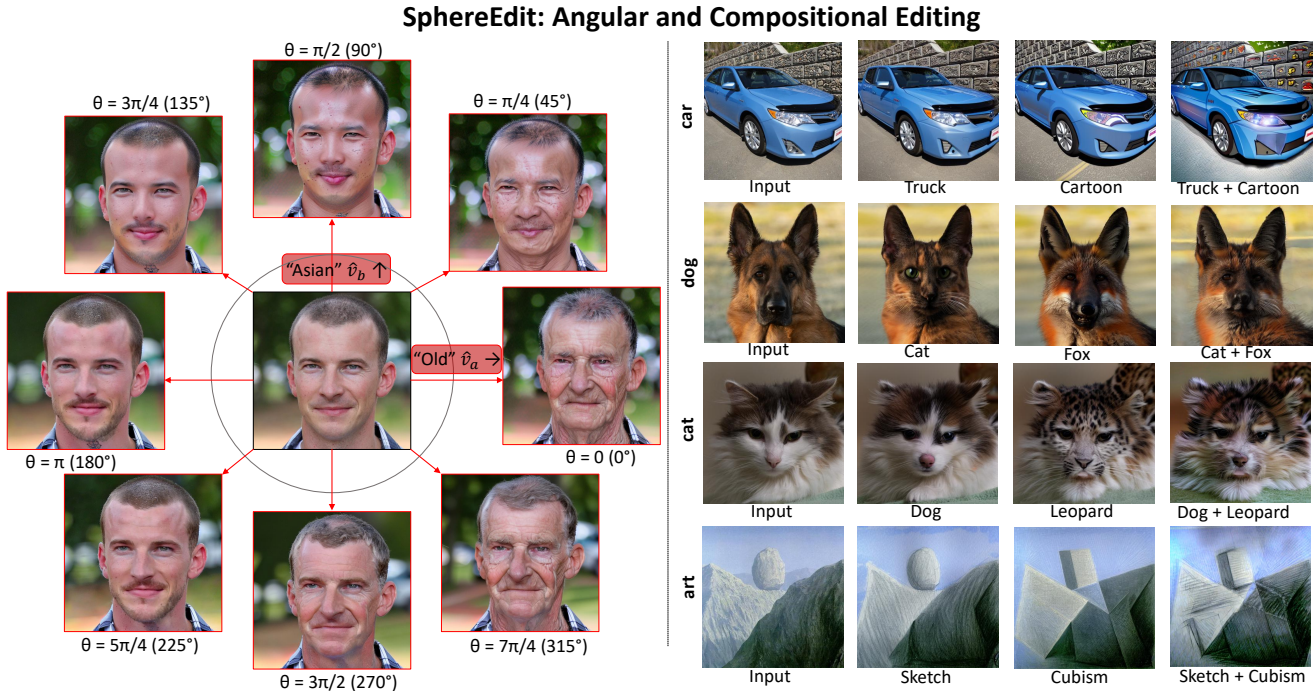

Figure 1. **SphereEdit.** We introduce a training-free, condition-aware diffusion editing method that computes per-attribute directions from noise differences and derives spatial masks from cross-attention. *Left:* Angular control with spherical coefficients $(\lambda_a, \lambda_b) = (\cos\theta, \sin\theta)$ produces a smooth progression from attribute a ($\theta = 0$) to attribute b ($\theta = \frac{\pi}{2}$), with natural blends at intermediate angles. *Right:* Compositional edits across domains: Input → A only → B only → A+B ($\theta = \frac{\pi}{4}$) for *faces, animals, vehicles and arts.* SphereEdit localizes edits with attention-derived masks and composes attributes while preserving identity.

Abstract

Despite significant advances in diffusion models, achieving precise and composable image editing without task-specific training remains a challenge. Existing approaches often rely on iterative optimization or linear latent operations, which are slow, brittle, and prone to attribute entanglement (e.g., editing "lipstick" inadvertently alters skin tone). We introduce SphereEdit, a training-free framework that leverages the spherical geometry of diffusion embeddings and token aware cross-attention to enable interpretable, fine-grained control. We represent semantic attributes as unit vector directions in the denoiser's prediction space and show that antipodal symmetry ('old' is approximately the negation of 'young') naturally supports bidirectional edits, while approximate orthogonality enables clean composition through spherical coefficient. At inference, these directions modulate cross-attention activations,

producing spatially localized edits without optimization or fine-tuning. SphereEdit achieves sharper, more disentangled edits than prior baselines, while remaining plug-and-play and applicable across diverse image editing tasks. The code is available at https://github.com/sala-kon/SphereEdit

1. Introduction

Diffusion models have revolutionized image synthesis, creating unprecedented demand for precise and composable image editing [2, 11, 21, 31]. Their ability to generate photorealistic images from simple text prompts has opened new opportunities in creative design, virtual try-on, content creation, and scientific visualization. In these applications, edits must be identity-preserving, fast, and controllable: adding lipstick should not inadvertently alter skin tone, and inserting eyeglasses should not reshape facial structure or erase subtle identity cues. Despite their impressive generative power, two fundamental challenges persist. First, there is limited understanding of how semantic attributes emerge and interact across denoising steps, making it difficult to predict or disentangle their effects. Second, there remains a lack of interpretable mechanisms for controlling latent space semantics, which forces practitioners to rely on ad hoc prompt engineering or trial-and-error tuning. Together, these gaps hinder progress toward reliable, fine-grained editing: users may obtain visually pleasing results, but without guarantees of consistency, disentanglement, or semantic faithfulness. These gaps limit our ability to achieve interpretable and disentangled editing, despite the remarkable generative quality of modern diffusion pipelines [9, 11, 13, 21, 29].

Recent work has begun probing the geometry of diffusion representations, modelling denoising trajectories with Riemannian metrics [25] or enforcing isometric constraints [9] to enhance disentanglement. These studies highlight that semantic structure is not arbitrary, but shaped by geometric regularities in the latent space. In parallel, most practical editing methods rely on the cross-attention mechanism, which aligns text conditions with spatial features [5, 32, 33]. While this enables attributes such as "glasses" or "beard" to be added via CLIP embeddings [28], existing approaches face persistent trade-offs. Optimization and inversion techniques (e.g., test-time adaptation) are slow, fragile to prompts, and prone to identity drift [24, 37]. Linear latent operations treat attributes as global shifts, but edits often leak into unrelated regions and fail to compose cleanly across prompts [16]. Prompt- and attention-rewiring strategies improve semantic alignment, but depend on layer- or head-specific heuristics with weak spatial grounding, leading to edit bleeding (e.g., "lipstick" altering cheeks) [8, 10]. Together, these efforts reveal a key limitation: existing methods capture either global geometry or condition alignment, but not both. Effective editing demands a framework that unifies geometric regularity with localized, condition-aware control—without retraining or fragile heuristics.

In this work, we propose SphereEdit, a training-free framework that integrates geometric control with condition-aware spatial localization for precise, composable diffusion editing. SphereEdit exploits the spherical geometry of diffusion embeddings, representing semantic attributes as angular directions on the unit sphere. This formulation naturally supports antipodal symmetry (opposite directions correspond to opposite semantics, e.g., "old" vs. "young") and smooth angular composition without metric estimation. Edits are injected at inference time via cross-attention, requiring no retraining. To ensure spatial precision, we derive condition-specific attention maps that localize where spherical edits apply, restricting changes to relevant regions (e.g., lipstick affects only lips).

Across benchmarks such as FFHQ and CelebA-HQ, SphereEdit yields sharper, more disentangled edits than optimization-, linear-latent, and attention-based baselines, reducing identity drift and improving edit localization.
Our contributions are as follows.

- We propose SphereEdit, an interpretable and stable geometric framework for diffusion editing that represents semantic attributes as unit vector directions on the sphere. This formulation unifies different levels of control: 1D antipodal edits (adding or removing an attribute), 2D angular trade-offs (interpolating between two attributes), and 3D spherical composition (balancing three or more).
- SphereEdit is a training-free, condition-aware mechanism that leverages cross-attention attribution to derive spatially localized masks, ensuring edits remain confined to relevant regions (e.g., lipstick only modifies lips) while avoiding leakage into unrelated areas.
- SphereEdit produces sharper, more disentangled edits and stronger identity preservation than editing baselines as shown in experiments across diverse domains (faces, pets, cars, and arts).

2. Related work

Diffusion Models and Latent Diffusion. Diffusion models have emerged as state-of-the-art generative models, demonstrating impressive performance in high-fidelity image synthesis [11, 30, 31]. These models generate images through iterative denoising, gradually transforming Gaussian noise into realistic images [11, 31]. Subsequent improvements have have enhanced sampling efficiency, including non-Markovian formulations (DDIM) that also enable inversion, as well as alternative noise schedules [23, 31]. Particularly relevant to our work, Latent Diffusion Models (LDMs) [29]shift this process into a compressed latent space, maintaining perceptual quality while dramatically reducing computational cost. While these develop-

ments have focused on generation quality and efficiency, the geometric and semantic organization of latent spaces remains poorly understood. This lack of understanding limits progress toward controllable and interpretable editing.

Image Editing with Diffusion Models. Although diffusion models can perform text-guided edits, naive conditioning often causes semantic entanglement and unintended changes in unrelated regions [17, 34]. Training-free methods address this challenge along several directions. Inversion-based approaches such as DDIM inversion and Null-Text inversion enable faithful reconstruction and better text–image alignment [19, 22, 31], though they remain slow and fragile at inference. To mitigate this limitation, TurboEdit [35] proposes an accelerated inversion procedure, but the speedup often comes at the cost of visual artifacts. Latent-space editing methods such as DiffusionCLIP [15, 17] represent attributes as semantic directions, but lack spatial grounding and frequently leak into irrelevant regions. Cycle-consistency approaches such as Cycle Diffusion [34] preserve global identity, but struggle to support fine-grained or multi-attribute edits. Instruction-driven frameworks such as InstructPix2Pix [4] improve usability by following natural language instructions, but typically sacrifice precision and spatial locality. Despite these advances, precise, compositional editing that preserves identity and remains spatially localized is still unresolved, motivating closer investigation of attention localization and compositional guidance.

Attention Localization and Compositional Guidance. Latent diffusion models leverage cross-attention layers to align textual prompts with visual features, enabling semantic control and spatial localization of edits. Cross-attention mechanisms model interactions between image features and text tokens, and DAAM aggregates these signals across layers and heads to produce spatial attribution maps [32]. Beyond attention control, guidance composition techniques combine multiple conditions either by additive score fusion or by multiplicative product-of-experts formulations. Composable Diffusion enables multi-concept generation using energy based models [20], while SEGA injects semantic guidance terms during sampling to amplify or suppress specific concepts [3, 12]. Other approaches refine attention directly: Attend-and-Excite strengthens object presence [7], while GLIGEN [18] and MasaCtrl [5] preserve spatial layouts through structure-guided control While effective, these methods often average attention across layers or heads, diluting condition specificity and producing blurred attribute boundaries. Moreover, when edits target the same region, they often conflict, causing entangled results and reducing compositional precision. These challenges motivate a complementary research that grounds editing in the semantic geometry of diffusion representations, as we discuss next through CLIP semantics and latent geometry.

CLIP Semantics and Latent Geometry. The joint text–image embedding space learned by CLIP enables semantic manipulation through directional arithmetic, where attributes are modeled as vectors that can be added, subtracted, or inverted [26, 27]. StyleCLIP maps CLIP-derived directions into StyleGAN's latent space for controlled face editing [26], while DiffusionCLIP extends this idea to diffusion models [15]. More recently, geometric perspectives have emerged: Riemannian diffusion interprets latent spaces through manifold geometry [13], and near-isometric mappings between semantic and visual representations have been explored to preserve disentanglement [9].Despite these advances, most approaches rely on linear composition of attribute directions, which often causes interference when multiple semantics overlap. Geometric methods offer promising structure, but they remain largely theoretical and do not directly yield practical tools for localized, compositional editing. Exploring spherical structures, with their inherent antipodal symmetry and angular composition, remains an open direction for diffusion models.

3. Method

Our approach is guided by four key criteria: (i) localized, only regions relevant to the *target* attribute are modified; (ii) composable, allowing multiple attributes should combine predictably without unexpected interaction; (iii) stable, with bounded magnitude to prevent over-saturation; and (iv) training-free, requiring no model finetuning.

To satisfy these criteria, we (1) extract per-attribute directions $\mathbf{v}$ from differences in denoiser predictions, (2) localize them with condition-aware attention maps, (3) resolve conflicts via overlap-aware orthogonalization, and (4) combine them through spherical coefficient for stable and interpretable trade-offs.

3.1. Latent Diffusion Preliminaries

Following latent diffusion framework [31], an image $x_0 \in \mathbb{R}^{H\times W\times 3}$ is encoded by a pretrained VAE encoder $\mathcal{E}$ into a latent $z_0 = \mathcal{E}(x_0)$. The forward process corrupts z_0 with Gaussian noise,

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative schedule.

A UNet denoiser ϵ_θ predicts noise conditioned on noisy latent z_t, timestep t, and text embedding c:

$$\hat{\epsilon} = \epsilon_\theta(z_t, t, c) \tag{2}$$

We obtain latent trajectories $\{z_t\}$ via DDIM inversion [31] and steer the denoising by modifying $\hat{\epsilon}$ at inference. The VAE decoder $\mathcal{D}$ is used only for visualization.

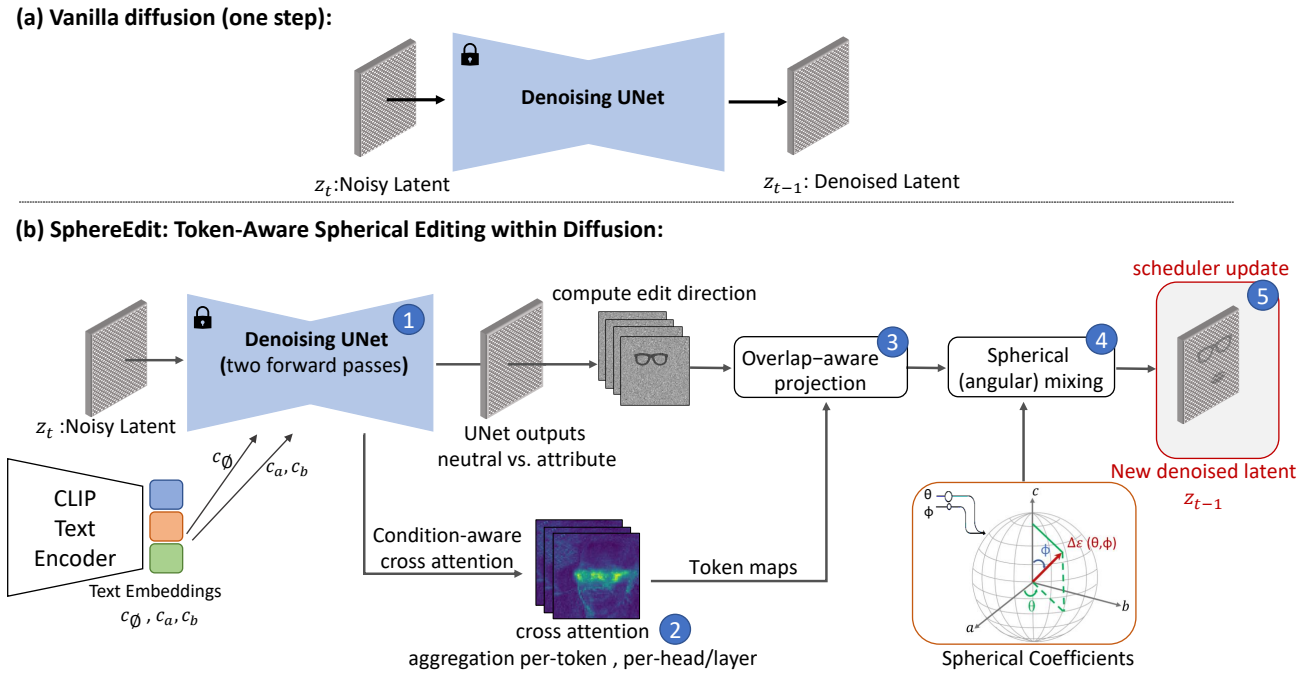

Figure 2. **SphereEdit framework. (a)** Vanilla diffusion: the UNet predicts noise and the scheduler updates the latent from z_t to z_{t-1}. **(b)** SphereEdit within the same step: (1) run two UNet passes per attribute (neutral vs. attribute) to form a per-attribute edit direction via noise difference; (2) aggregate cross-attention into condition maps—used to detect spatial overlap between attributes; (3) reduce interference in overlap regions via a local masked projection; (4) combine attribute directions with angular (spherical) coefficients; (5) add the mixed edit to the base prediction and let the scheduler produce z_{t-1}. No retraining.

3.2. SphereEdit for Semantic Editing

(1) Attributes Directions on the Unit Sphere. Let $\mathcal{A} = \{a_1, \ldots, a_K\}$ denote a set of K semantic attributes (e.g., glasses, beard, young), each associated with a conditioning text embedding c_k The unconditional case is represented by a null token $\varnothing$ with embedding $c_\varnothing$. At denoising step t, we define a binary mask $m_k(t) \in \{0, 1\}$ that specifies whether attribute a_k is active. This scheduling mechanism enables fine-grained temporal control over edits along the denoising trajectory. For example, "glasses" can be applied only at later steps, allowing the facial structure to be reconstructed first.

To isolate the semantic contribution of an attribute a_k, we compare the denoiser's predictions under two conditions: conditioned on a_k and on the unconditional null token. Subtracting these predictions cancels shared structure from z_t and reveals the semantic contribution unique to a_k.

Formally at step t, we approximate the local linearization of the denoiser in the text-conditioning space. The attribute direction $\mathbf{v}_k(t)$ is defined as:

$$\mathbf{v}_k(t) = \epsilon_\theta(z_t, t, c_k) - \epsilon_\theta(z_t, t, c_\varnothing), \tag{3}$$

Intuitively, $\mathbf{v}_k(t)$ approximates the denoiser's direction to the attribute embedding, i.e., "what the model thinks a_k looks like" at timestep t.

To avoid scale imbalance between different attributes, we normalize $\mathbf{v}_k(t)$ to unit vector: $\hat{\mathbf{v}}_k(t) = \frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|_2}$.

(2) Spherical Geometry of Diffusion Latents. Recent studies indicate that diffusion latents exhibit an approximately spherical geometry [9, 13], a structure that naturally supports smooth semantic interpolations and consistent attribute transitions. Analogous findings in variational autoencoder, show that latent variables also align on spherical manifolds, clustering semantically related concepts (e.g., male attributes) near one another while placing contrasting concepts on opposing regions [36]. A geometrically sound latent space for a diffusion model provides a precise control over attributes directly in the latent space [9].

Building on this insight, we treat each attribute a_k as occupying a region on the unit sphere, with its direction $\hat{\mathbf{v}}_k$, pointing toward its semantic locus. When editing an image, e.g, adding glasses, we move along $\hat{\mathbf{v}}_{glasses}$ on the sphere.

For three active attributes (e.g glasses, lipstick, young), we assign each attribute a spherical coefficient λ_k that determines its contribution on the unit sphere. Using the standard spherical–Cartesian transformation [1], these coefficients are defined as:

$$\lambda_1 = \cos\theta, \quad \lambda_2 = \sin\theta\cos\phi, \quad \lambda_3 = \sin\theta\sin\phi. \quad (4)$$

Here, θ, ϕ are the spherical angles that determine how contributions are distributed among attributes on the unit sphere. Each normalized attribute direction $\hat{\mathbf{v}}_k$ is then transformed by its spherical coefficient λ_k, yielding the transformed direction $\hat{\mathbf{v}}_k^{Tr}$.

$$\hat{v}_k^{Tr} = \lambda_k \hat{v}_k, \qquad k \in 1, 2, 3. \quad (5)$$

Equation 5 naturally extends from lower to higher dimensions. In the one-dimensional case where only a single attribute a_k is active, it reduces to moving along a line (e.g x-axis with $\theta = 0^\circ$), with the scale factor controlling edit intensity. In the two-dimensional case, with two attributes, it simplifies to a circle in the $x - y$ plane, parameterized by θ with $0^\circ \leq \theta < 360^\circ$. In the three -dimensional case, with three attributes, it recovers the full spherical formulation, with both θ and ϕ controlling the allocation across three dimensions. Additionally, each direction can be scaled by a factor between 0 and 1 to adjust the visibility of its edit; for single-attribute edits, a larger scaling factor can be used to intensify the effect.

Therefore, our proposed method controls the edit through the attribute direction $\hat{\mathbf{v}}_k$.

(3) Localization with condition-aware attention Applying $\mathbf{v}_k(t)$ globally risks unintended edits in irrelevant regions. To avoid this, we compute condition-aware cross-attention maps M_k that highlight spatial regions associated with the attribute condition. Unlike prompt-wide maps, M_k focuses only on the relevant condition span, preventing leakage. For example, "glasses" attends to the eye region, while "beard" attends to the lower face.

(4) Overlap-aware Orthogonalization. Attributes may attend to overlapping spatial supports (e.g., mouth and beard). Naively summing directions can cause entanglement. A global orthogonalization would distort semantics, so we enforce orthogonality only inside overlapping regions, preserving meaning elsewhere. To avoid this, we restrict orthogonalization to the exact areas of overlap, preserving each attribute's meaning elsewhere.

Let $\mathcal{R}_k(t)$ be the spatial support of a_k at timestep t, obtained by thresholding its heatmap $M_k(x, y, t)$ at quantile q. Similarly, $\mathcal{R}_j(t)$ is defined for the attribute a_j. The overlap mask between a_j and a_k is therefore:

$$S_{jk}(x, y, t) = \mathbf{1}\big[(x, y) \in \mathcal{R}_j(t) \cap \mathcal{R}_k(t)\big]. \quad (6)$$

We then decorrelate $\mathbf{v}_k$ from $\mathbf{v}_j$ only where overlaps occur by subtracting its masked Gram–Schmidt projection:

$$\mathbf{v}_k \leftarrow \mathbf{v}_k - \langle \mathbf{v}_k, \mathbf{v}_j \rangle_{S_{jk}} (\mathbf{v}_j \odot S_{jk}), \quad (7)$$

where $\langle \cdot, \cdot \rangle_{S_{jk}}$ is a masked inner product, and $\odot$ indicates element-wise multiplication with the overlap mask.

With such overlap-aware disentanglement in place, the resulting directions can now be combined, which we achieve next through a spherical parameterization that yields stable and interpretable trade-offs.

(5) Final editted direction When applying multiple attributes (e.g editing age and race), we simply sum the transformed unit vector direction $\hat{\mathbf{v}}_k^{Tr}$ obtained via Eq. 5 for the active attributes. The composed edit direction is then defined as follows:

$$\Delta\epsilon_t = \sum_{k:m_k(t)=1} \hat{\mathbf{v}}_k^{Tr}. \quad (8)$$

The edited denoiser output becomes:

$$\hat{\epsilon}(t) = \epsilon_{\text{base}}(t) + \Delta\epsilon_t, \quad (9)$$

which replaces the original denoiser output in the DDIM update, progressively steering the latent trajectory toward the edited image while retaining fidelity to the original content.

3.3. Algorithmic Summary

Algorithm 1 outlines our editing procedure per denoising step. It summarizes how attribute directions are extracted, localized, disentangled, and combined via spherical composition before being injected into the DDIM update.

Algorithm 1 SphereEdit (one denoising step t)

Compute base prediction $\epsilon_{\text{base}} \leftarrow \epsilon_\theta(z_t, t, c_\varnothing)$
for each active attribute a_k **do**
  Direction $\mathbf{v}_k \leftarrow \epsilon_\theta(z_t, t, c_k) - \epsilon_{\text{base}}$
  Normalize $\hat{\mathbf{v}}_k \leftarrow \mathbf{v}_k / \|\mathbf{v}_k\|_2$
  Attention map $M_k \leftarrow \text{CrossAttention}(z_t, c_k)$
end for
for each pair (j, k) of active attributes **do**
  Overlap $S_{jk} \leftarrow \text{Overlap}(M_j, M_k)$
  Orthogonalize $\mathbf{v}_k \leftarrow \mathbf{v}_k - \langle \mathbf{v}_k, \mathbf{v}_j \rangle_{S_{jk}} (\mathbf{v}_j \odot S_{jk})$
end for
Coefficients $\lambda_k \leftarrow \text{SphericalCoefficients}(\theta, \phi)$
Edited prediction $\hat{\epsilon}(t) \leftarrow \epsilon_{\text{base}} + \sum_k \lambda_k \hat{\mathbf{v}}_k$
Update latent z_{t-1} via DDIM using $\hat{\epsilon}(t)$

4. Experiments

To assess the effectiveness of SphereEdit for disentangled and composable semantic editing, we conducted both qualitative and quantitative experiments. We compare our framework against state-of-the-art baselines, including inversion-based, attention-manipulation, and latent-direction methods. Our results highlight SphereEdit's superior spatial precision, composability, and identity preservation.

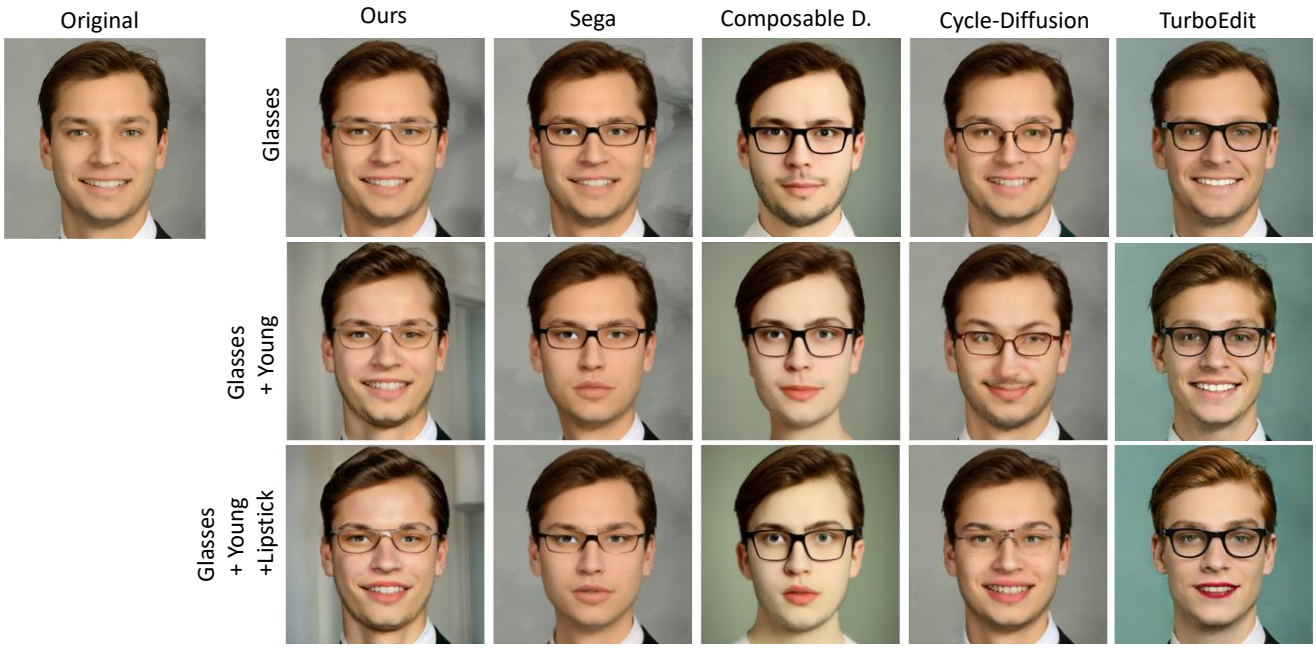

Figure 3. **Qualitative results.** We compare SphereEdit with SEGA [3], Composable Diffusion[20], Cycle Diffusion[34], and TurboEdit[35] for single and composable attributes *Glasses*, *Glasses+Young*, and *Glasses+Young+Lipstick*. Our method produced more localized edits while preserving identity.

4.1. Experimental Setup

All experiments are conducted with Stable Diffusion v1.5. We use 50 denoising steps and set the quantile threshold to $q = 0.8$ unless otherwise specified. For spherical composition, we fix the angular parameters to $\theta = \phi = \pi/2$, which corresponds to activating all defined attributes, unless stated otherwise. Attribute-specific editing schedules are controlled by varying the start and end timesteps depending on the task. To ensure reproducibility, we fix the random seed to 0. On an NVIDIA L40 GPU, a full edit with three attributes takes approximately 12 seconds.

4.2. Qualitative Results

Fig. 1 and 3 show the qualitative results of our editing experiments. Our method enables smooth angular traversal, spatially precise edits, and compositional control across diverse domains. Figure 1 illustrates these capabilities. On the left, varying the spherical angle between two semantic directions: race and age. Our result produces a continuous interpolation that gradually shifts identity along both axes. This shows how spherical parameterization turns angular coordinates into interpretable control: moving to the opposite angle inverts the attribute (e.g., "old" vs. "young") even when only the "old" embedding is provided. In addition, intermediate angles yield meaningful blends without identity drift. On the right, we show results beyond faces, including cars, dogs, cats, and arts. For each input, SphereEdit applies two single-attribute edits (e.g., Truck, Cartoon; Cat, Fox; Dog, Leopard) and then composes them. Edits remain localized and disentangled. The truck structure is preserved when cartoonized, the dog retains shape when gaining fox-like fur, and the cat smoothly acquire leopard textures. This demonstates that SphereEdit applies consistently across domains beyond human faces, including animals and vehicles.

In Figure 3, we compare SphereEdit against state-of-the-art baselines: SEGA [3], Composable Diffusion [20], Cycle-Diffusion [34], and TurboEdit [35]. SphereEdit consistently produces sharper, better localized, and more identity-preserving edits. For example, Composable Diffusion and TurboEdit often alter facial identity and introduce global tone shifts. SEGA and Cycle-Diffusion tend to under-edit subtle attributes, failing to capture fine-grained changes. In multi-attribute compositions such as "Glasses+Young" and "Glasses+Young+Lipstick," SphereEdit confines edits to the intended regions while preserving identity, while TurboEdit introduces artifacts and Composable Diffusion entangles attributes.

4.3. Quantitative Results

To validate the effectiveness of SphereEdit, we conduct quantitative comparisons against state-of-the-art methods on two compositional editing tasks: Glasses+Mustache and Eyebrow+Lipstick. All methods are evaluated on the same set of 100 images from the FFHQ dataset [14], with results reported in Table 1.

Table 1. **Quantitative results.** We compare SphereEdit with SEGA [3], Composable Diffusion[20], Cycle Diffusion[34], and TurboEdit[35] on two compositional edits *Glasses+Mustache* and *Eyebrow+Lipstick*. We report CLIP-T [27] for text alignment, CLIP-I [27] and DINO [6] for content and identity preservation, and LPIPS [38] for visual quality (higher is better except for LPIPS). SphereEdit achieves the strongest overall, leading most metrics and remaining competitive where it is not the top performer.

	Glasses + Mustache				Eyebrow + Lipstick			
	Clip-T ↑	Clip-I ↑	DINO ↑	LPIPS ↓	Clip-T ↑	Clip-I ↑	DINO ↑	LPIPS ↓
Sega [3]	0.2226	0.7395	0.6271	**0.1796**	0.2230	0.6960	0.6011	0.1870
Composable [20]	0.2337	0.6353	0.3226	0.3107	0.2290	0.6227	0.3860	0.3589
CycleDiff [34]	0.2260	0.6909	0.4774	0.2142	0.2238	0.6562	0.5229	0.2332
TurboEdit [35]	0.2203	0.6771	0.6250	0.2248	**0.2333**	0.5871	0.4702	0.2252
Ours	**0.2369**	**0.7573**	**0.7026**	0.1919	0.2321	**0.7275**	**0.6836**	**0.1436**

Evaluation metrics. We report four metrics. CLIP-T [27] measures text alignment, capturing how well the generated image reflects the requested attributes. CLIP-I [27] evaluates identity preservation, quantifying whether the subject remains consistent after editing. DINO [6] assesses structural similarity, ensuring that key image features such as geometry and layout are maintained. Finally, LPIPS [38] measures low-level visual fidelity, captures differences in texture, color, and appearance where lower values indicate fewer visual artifacts.

On Glasses+Mustache, SphereEdit achieves the strongest text alignment (CLIP-T), identity preservation (CLIP-I), and structural similarity (DINO), while performing competitively on visual quality (LPIPS), just behind SEGA. On Eyebrow+Lipstick, SphereEdit achieves the best CLIP-I, DINO, and LPIPS scores, while TurboEdit slightly surpasses it in CLIP-T by 0.0012. Overall, SphereEdit demonstrates the most consistent improvements across all metrics, confirming its ability to deliver identity-preserving, structurally coherent, and visually faithful edits in challenging multi-attribute settings.

4.4. Ablation Studies

On the ablation study, we evaluate two key components of our method. First, we study antipodal control, showing that attributes form bidirectional directions (e.g., "old" vs. "young"). Second, we analyze the effect of condition-aware masking, combining both the mask threshold q and attention-based gating, to assess how they influence spatial locality and attribute disentanglement.

Ablation on antipodal control. SphereEdit represents attributes as directions on the unit sphere, where positive and negative scales correspond to opposite semantics. In Fig. 4, we visualize edits for three attributes: "Age", "Female", and "Cyprus" (a race/ethnicity condition). For each case, the negative scale removes the attribute, while the positive scale accentuates it. For example, decreasing "Age" removes wrinkles and softens facial features, while increasing it adds age-related texture. Similarly, scaling along the "Female" direction adjusts facial softness and expression, and scaling "Cyprus" alters ethnicity-related appearance cues. Across all cases, the original identity is preserved and edits remain localized, demonstrating that our spherical formulation naturally supports antipodal symmetry: moving in opposite directions along the same attribute vector yields interpretable, reversible changes.

Figure 4. **Ablation on Antipodal control.** Edits for age, gender, and race using negative and positive scales (-7.5,+7.5). Negative scaling inverts the attribute (e.g., old → young), while positive scaling amplifies it, showing SphereEdit's bidirectional control.

Ablation on condition-aware masking. We analyze the role of condition-aware masking in controlling the spatial extent of edits (Fig. 5) when applying two attributes edit (glasses and lipstick). In Fig. 5(a), we vary the quantile threshold q used to binarize attention maps. At low

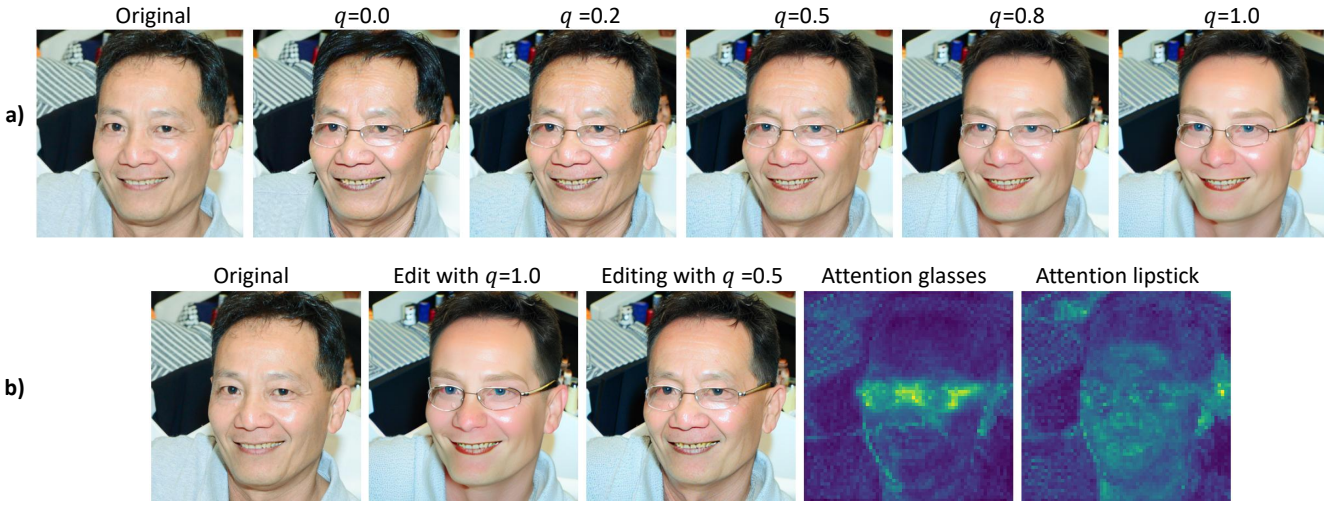

Figure 5. **Ablation on condition-aware masking.** (a) Edits with attributes *glasses* and *lipstick* at different thresholds ($q = 0, 0.2, 0.5, 0.8, 1$), where $q = 0$ means empty mask and $q = 1$ means no mask. (b) Comparison of edits with and without masking ($q = 0.5$), along with cross-attention maps for *glasses* and *lipstick*.

thresholds, masks cover broad regions, causing edits to leak into unrelated areas—for example, lipstick spills into skin or global tone shifts occur. Moderate thresholds ($q \approx 0.5$) yield well-localized edits, with glasses confined to the frames and lipstick restricted to the lips, preserving identity. Very high thresholds ($q > 0.8$) prune masks too aggressively, fragmenting the target region and reducing identity preservation. In Fig. 5(b), we visualize edits with and without masking. Without masking, lipstick spreads beyond the lips and alters skin tone, whereas with masking ($q = 0.5$), edits are confined to the condition's attended regions. The corresponding attention maps clarify this effect: *glasses* produce sharp, localized activations around the eyes, while *lipstick* is weaker and more diffuse, explaining its sensitivity to thresholding. Overall, condition-aware masking proves critical for disentangled, multi-attribute edits, and we find $q = 0.8$ offers the best compromise between edit strength and spatial precision.

5. Limitations

SphereEdit delivers localized and composable control, but some aspects remain open for further study. First, our evaluation focuses on up to three attributes; while the spherical formulation naturally generalizes to higher dimensions, it is not yet clear how interpretable and stable such multi-way trade-offs will be in practice. Exploring this setting could shed light on the scalability of spherical editing to more complex attribute compositions. Second, although antipodal symmetry emerges consistently in the diffusion embedding space (e.g., old vs. young), its semantics can be ambiguous in certain application domains. For example, in medical imaging it is unclear whether the "opposite" of a disease corresponds to health or to a different condition, and in demographic attributes such as race, opposites are not well defined. This raises important questions about how spherical geometry interacts with sensitive attributes.

6. Conclusion

We introduce SphereEdit, a training-free and interpretable framework for precise, composable diffusion editing. By representing semantic attributes as unit-norm directions in the denoiser's prediction space, SphereEdit leverages antipodal symmetry for bidirectional edits, overlap-aware orthogonalization for disentanglement, and spherical parameterization for stable multi-attribute composition. Our experiments demonstrate sharper, more localized, and more identity-preserving edits than inversion, attention-rewiring, and latent-direction baselines, across domains including faces, animals, and vehicles.

SphereEdit achieves state-of-the-art practical results and provides a geometric view that links semantic editing to spherical coordinate systems, yielding angular controls that are both intuitive and interpretable. While open challenges remain, such as scaling to higher-order attribute combinations and clarifying the meaning of antipodal semantics in domains like medical imaging, our work establishes SphereEdit as an interpretable, training-free approach for controllable and composable editing in diffusion models.

Acknowledgments

This work was supported in part by the Connected Minds Program through the Canada First Research Excellence Fund Grant #CFREF-2022-00010. We gratefully acknowledge the Vector Institute for providing high-performance computing resources. Additional support was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant awarded to Dr. Laleh Seyyed-Kalantari, as well as by the Google Research Scholar Award and the Canadian AI Safety Institute Research Program at CIFAR.

References

[1] Constantine Balanis. *Advanced Engineering Electormagnetics*. John Wiley Sons, first edition, 2008. ¡Any extra info, e.g. URL or series¿. 4

[2] Valentina Bazyleva, Nicolo Bonettini, and Gaurav Bharaj. X-edit: Detecting and localizing edits in images altered by text-guided diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2729–2739, 2025. 2

[3] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36: 25365–25389, 2023. 3, 6, 7

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[5] 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, 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 7

[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[9] Jaehoon Hahm, Junho Lee, Sunghyun Kim, and Joonseok Lee. Isometric representation learning for disentangled latent space of diffusion models. *arXiv preprint arXiv:2407.11451*, 2024. 2, 3, 4

[10] 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

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

[12] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37: 66743–66772, 2024. 3

[13] Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022. 2, 3, 4

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[15] 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. 3

[16] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2

[17] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 3

[18] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023. 3

[19] Haonan Lin, Yan Chen, Jiahao Wang, Wenbin An, Mengmeng Wang, Feng Tian, Yong Liu, Guang Dai, Jingdong Wang, and Qianying Wang. Schedule your edit: A simple yet effective diffusion noise schedule for image editing. *Advances in Neural Information Processing Systems*, 37:115712–115756, 2024. 3

[20] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pages 423–439. Springer, 2022. 3, 6, 7

[21] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 289–299, 2023. 2

[22] 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*, pages 6038–6047, 2023. 3

[23] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International*

conference on machine learning, pages 8162–8171. PMLR, 2021. 2

[24] Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15912–15921, 2023. 2

[25] 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

[26] 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. 3

[27] 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. 3, 7

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 2

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021. 2, 3

[32] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*, 2022. 2, 3

[33] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. Promptcharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024. 2

[34] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023. 3, 6, 7

[35] Zongze Wu, Nicholas Kolkin, Jonathan Brandt, Richard Zhang, and Eli Shechtman. Turboedit: Instant text-based image editing. In *European Conference on Computer Vision*, pages 365–381. Springer, 2024. 3, 6, 7

[36] Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4503–4513, Brussels, Belgium, 2018. Association for Computational Linguistics. 4

[37] Brian Yang, Huangyuan Su, Nikolaos Gkanatsios, Tsung-Wei Ke, Ayush Jain, Jeff Schneider, and Katerina Fragkiadaki. Diffusion-es: Gradient-free planning with diffusion for autonomous and instruction-guided driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15342–15353, 2024. 2

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7