Learn Once, Edit Anywhere: Visual Direction Transfer for Diffusion Models

Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag

Virginia Tech
{ydalva, hidir, pinary}@vt.edu
http://vidit-edit.github.io

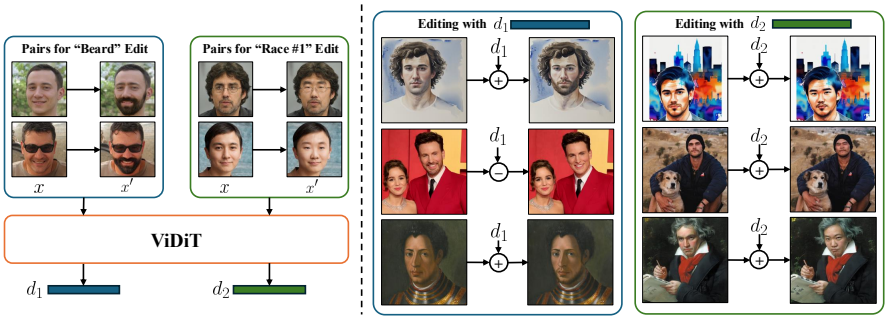

Fig. 1: ViDiT. By learning a continuous direction from a small set of image-edit pairs, ViDiT transfers fine-grained semantic edits to any image in a zero-shot manner. Once learned, each direction enables disentangled, bidirectional attribute control, applied with scale λ_e, across diverse domains including real-world portraits, artistic styles, and illustrations, without any fine-tuning or per-image optimization.

Abstract. The rapid advancement of diffusion models has enabled the generation of high-fidelity images from textual prompts, yet achieving precise, disentangled control over specific attributes remains a significant challenge. A fundamental limitation arises because visual differences between images are often far more descriptive and nuanced than what can be captured through human-crafted text descriptions, which frequently fail to convey fine-grained semantic details. To address this, we introduce **ViDiT (Visual Direction Transfer for Diffusion)**, a framework that expands the editing vocabulary by capturing latent semantics directly from image-edit pairs. ViDiT learns the underlying transformation by optimizing a single, global, and continuous editing direction from a small set of "before-and-after" examples. This optimization process transfers visual changes into the diffusion model's conditioning space, allowing for detailed edits that text alone cannot easily describe. ViDiT operates on a "Learn Once" principle, which completely eliminates the need for model fine-tuning or expensive per-image optimization during inference.

Once learned, these continuous directions enable "Edit Anywhere" capabilities, allowing users to apply highly disentangled manipulations, such as changes in facial features, animal attributes, or artistic styles, to any image in a zero-shot manner with granular control over the edit intensity. Quantitative and qualitative evaluations demonstrate that ViDiT outperforms existing text-based editing methods in maintaining input faithfulness while achieving precise, scalable attribute control.

1 Introduction

Denoising Diffusion Models (DDMs) [13] and Latent Diffusion Models (LDMs) [27] have gained significant popularity in the generative modeling landscape due to their ability to synthesize high-quality, high-resolution images across diverse domains [13, 27]. Their success has led researchers to increasingly leverage them for image editing tasks, ranging from text-driven modifications [12] to layout-based control [37]. However, a fundamental limitation persists in these workflows: human-crafted text descriptions are often too coarse to capture the nuanced visual differences, or "deltas", that define precise, fine-grained edits. While a pair of images can effortlessly demonstrate a subtle semantic shift, translating that same transformation into a textual prompt is inherently constrained by the descriptive bottleneck of natural language. This creates a significant gap between the user's intent and the model's output, as the expressive power of language is frequently surpassed by the complexity of visual information.

This gap is further widened by the fact that diffusion models often remain black boxes, limiting the fine-grained interpretability required for disentangled control. Advancing such control requires understanding their semantic structures, particularly how specific attributes can be manipulated in isolation. In this regard, Generative Adversarial Networks (GANs) have proven to be useful models for achieving interpretable and disentangled edits due to their well-structured latent spaces that offer rich, linearly applicable semantics. By leveraging this organized architecture, specialized discovery frameworks [11, 35] have uncovered significant amounts of semantically meaningful directions [33], allowing for highly nuanced control. In contrast, current research on diffusion models has only uncovered a handful of disentangled latent directions [6, 19]. This disparity arises from the recursive noise estimation and the inherently complex management of variables across multiple timesteps in diffusion architectures, making it difficult to map the fine-grained visual deltas that are easily represented in GANs.

To overcome these limitations, we introduce **ViDiT** (Visual Direction Transfer for Diffusion), a framework that expands the existing editing vocabulary by capturing latent semantics directly from image-edit pairs. Operating on a "Learn Once, Edit Anywhere" principle, ViDiT moves away from a total dependency on human-crafted text by learning to capture latent semantics from a small set of before-and-after visual examples. Rather than a simple feedforward mapping, ViDiT learns the underlying transformation by optimizing a single, global, and **continuous editing direction** that transfers the visual delta into the diffusion model's conditioning space. This allows for detailed attribute control that

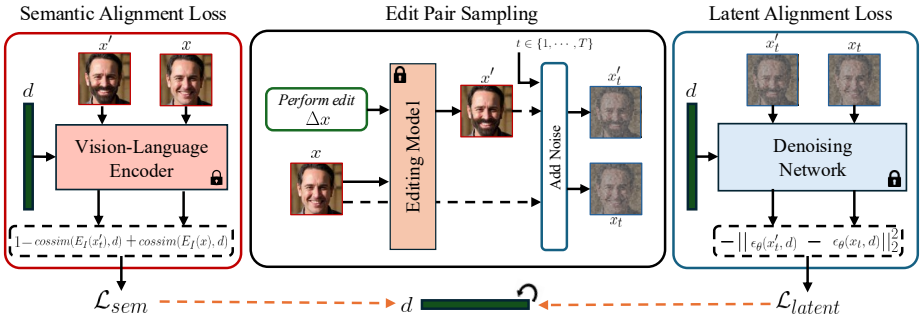

Fig. 2: ViDiT Framework. Given a pair of images (x, x'), where $x' = x + \Delta x$ is obtained by applying an edit Δx using any editing method, we optimize a continuous latent direction $\mathbf{d}$ using two complementary objectives. The **Semantic Alignment Loss** $\mathcal{L}_{\text{sem}}$ encourages $\mathbf{d}$ to be geometrically aligned with x' and misaligned with x in the feature space of a frozen Vision-Language Encoder (E). The **Latent Alignment Loss** $\mathcal{L}_{\text{latent}}$ maximizes the difference in the denoising predictions of the frozen Denoising Network (ϵ_θ) between the noised edited image x'_t and the noised original x_t, grounding $\mathbf{d}$ in the diffusion model's native latent space. Both losses jointly update $\mathbf{d}$, while all other components remain frozen. Once learned, $\mathbf{d}$ can be applied zero-shot to any image at inference time.

text alone cannot easily describe, all without requiring base-model finetuning or per-sample test-time optimization.

Our main contributions are outlined as follows:

- We present a universal framework that learns **continuous editing directions** from visual examples, effectively expanding the editing vocabulary of pre-trained diffusion models without requiring base model finetuning.
- We demonstrate the versatility of our direction transfer by successfully drawing from diverse sources, including unconditional generative spaces (e.g., StyleSpace [33]), domain adaptation models (e.g., StyleGAN-NADA [9]), and diffusion-based editors (e.g., Prompt2Prompt [12]).
- We showcase the **"Edit Anywhere"** capability of our approach by transferring a wide range of fine-grained semantics across diverse categories, including human faces, artistic styles, and extending these to real-world, in-the-wild images.
- Our evaluations demonstrate that ViDiT achieves superior disentanglement and identity preservation compared to state-of-the-art text-based diffusion editors, while providing granular control over edit intensity.
- We share our source code to facilitate further research in visual direction transfer for diffusion models.

2 Related Work

Latent Space Exploration of Diffusion Models. Text-to-image diffusion models utilize pre-trained text encoders to represent semantically rich information within their latent spaces, enabling high-quality image generation from textual conditions [27]. Recent research has sought to uncover the underlying semantic structure of this feature space to enable more precise control. Some approaches focus on specific architectural components, such as the representations encoded in the bottleneck block of the denoising network [19], while others investigate the transformation of representations across different imaging domains [32]. Inspired by latent space exploration in GANs, researchers have attempted to identify directions targeted by specific input semantics [24]; however, these methods often struggle to generalize to large-scale models like Stable Diffusion, as the diffusion latent space lacks the compact, linear structure that makes GAN spaces amenable to direction discovery. Recent efforts have addressed this by decomposing latent variables into energy functions of semantic significance [21] or using contrastive learning to discover unsupervised disentangled directions [6]. Despite these advances, such methods remain constrained by the inherent complexity of the diffusion latent space and are limited to directions already present within the model's learned representation. **ViDiT** bridges this gap by transferring well-understood semantics from diverse external sources directly into accessible, continuous editing directions.

Editing Visual Semantics with Diffusion Models. Editing with diffusion models typically relies on text prompts or structural guidance to modify attributes [12, 37]. While effective, these methods are often limited by the descriptive bottleneck of natural language. SEGA [2] utilizes semantic guidance to steer the diffusion process without additional training, while InstructPix2Pix [3] learns to follow human-edited instructions through a conditional diffusion model. Structure-based methods such as PnP-Diffusion [31] and MasaCtrl [4] inject intermediate features to preserve layout and identity during editing, and Asyrp [19] and DiffAE [25] explore semantic latent spaces within the denoising architecture itself. Recent works have also explored diverse conditioning strategies: IP-Adapter [34] injects image-based prompts via decoupled cross-attention, while Weights2Weights [7] and Attribute-Control [1] identify linear directions in the weight or CLIP embedding spaces, respectively. However, all of these approaches either depend on coarse textual descriptions, require large-scale training data, or are restricted to directions expressible through text-level features alone. **ViDiT** bridges these paradigms by learning the desired edit once from visual examples and applying it as a global, **continuous editing direction**, an approach that captures attribute-level semantics that are self-evident in visual pairs but remain difficult for textual or structural instructions to define precisely.

Combining Structured Generative Models and Diffusion. Recent literature has explored hybridizing the structured control of GANs with the generative power of diffusion models [30]. More relevant to our work, $\mathcal{W}+$ Adapter [20] and

Fig. 3: Directions transferred with ViDiT. ViDiT transfers fine-grained editing directions from diverse supervision sources to in-the-wild images in a zero-shot manner. **(a)** Directions transferred from StyleGAN2 [17] trained on FFHQ [16], demonstrating facial attribute edits such as *Bald*, *Gender*, *White Hair*, and *Baldness*. **(b)** Directions transferred from StyleGAN2 trained on AFHQ [5], showcasing animal attribute edits including *Lynx*, *Nose*, *Lion*, and *Skin Color*. **(c)** Style directions transferred from StyleGAN-NADA [9], applying artistic transformations such as *Ukiyoe* and *Pixar* to diverse real-world images. **(d)** Semantic directions transferred from Prompt2Prompt [12], enabling attribute edits such as *Gender* and *Beard* on real portraits. Across all cases, edits are disentangled and faithfully preserve the identity, background, and artistic style of the input.

Concept Sliders [10] leverage the disentangled capabilities of structured latent spaces to enhance diffusion-based editing. Concept Sliders [10] utilize low-rank adapters (LoRAs) to learn specific concepts, often employing samples from a pre-trained StyleGAN as a source for semantic directions, but require training a separate LoRA per concept, which is again bounded by a text prompt. The $\mathcal{W}+$ Adapter [20] introduces an additional architectural module to map StyleGAN's $W+$ latent space directly to the diffusion model, enabling the transfer of GAN-based edits to individual images. Similarly, DiffAE [25] learns a semantic encoder alongside the diffusion decoder to enable structured latent manipulation, though this requires training a dedicated architecture from scratch. In contrast, **ViDiT** eliminates the need for both model fine-tuning and the introduction of additional adapter modules. By learning a single, global, and **continuous editing direction** directly from visual deltas, our framework enables a "Learn Once, Edit Anywhere" capability applicable to any image in a zero-shot manner. This bypasses the overhead of training per-concept LoRAs or specialized latent mapping layers, providing a more lightweight and scalable solution for incorporating complex visual semantics into pre-trained diffusion models.

3 Method

3.1 Preliminaries

Diffusion Models. Diffusion models [13, 27] learn to generate data through an iterative reverse denoising process. Given a clean image x_0, the forward process adds Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ over T timesteps to produce a noisy latent x_t. A denoising network ϵ_θ is trained to predict this noise via the objective:

$$\mathcal{L}_{DM} = \mathbb{E}_{x_0,\epsilon,t}[||\epsilon - \epsilon_\theta(x_t, t)||_2^2] \tag{1}$$

During inference, sampling begins at $x_T \sim \mathcal{N}(0, 1)$ and iteratively approaches x_0. To steer this process, Classifier-Free Guidance (CFG) [14] computes a weighted combination of conditional and unconditional noise predictions:

$$\tilde{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \phi) + \lambda_g(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi)) \tag{2}$$

where c is a text-derived conditioning embedding, ϕ is the null-text embedding, and λ_g is the guidance scale. A key observation is that the CFG residual $(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \phi))$ acts as a *directional signal* in the noise-prediction space that steers generation towards c. ViDiT exploits this structure by learning a continuous direction $\mathbf{d}$ that replaces the text conditioning signal entirely, allowing semantic edits to be expressed as directions in this space rather than as natural language descriptions.

3.2 ViDiT: Transferring Visual Deltas

Motivation. Existing text-driven editing methods rely on natural language to specify semantic changes [2, 3, 12]. While convenient, this imposes a *linguistic bottleneck*: many fine-grained visual concepts, precise facial structure, subtle texture shifts, nuanced style changes, are self-evident in images but extremely difficult to describe in text. A prompt like "a person with a beard" captures only a coarse prior and inevitably entangles the edit with unrelated attributes such as age or skin tone. By contrast, a set of before-and-after image pairs $\{(x, x')\}$, where $x' = x + \Delta x$ is obtained by applying an edit Δx through any editing method, encodes the precise visual delta without linguistic ambiguity. Our key insight is that this delta can be *transferred* into a compact, continuous direction $\mathbf{d}$ that lives in the diffusion model's conditioning space and faithfully replicates the transformation on any new image.

Overview. ViDiT identifies a continuous latent direction $\mathbf{d}$, which acts as a transferable visual prompt, by optimising it against a small set of N image-edit pairs $\{\mathcal{X}_{\text{input}}, \mathcal{X}_{\text{edited}}\}$. The direction $\mathbf{d}$ is initialised randomly within the model's text-embedding space and refined using a dual-objective loss that enforces alignment at two complementary levels of abstraction: globally, via semantic feature matching, and locally, via dense latent-space correspondence. Crucially, the denoising network ϵ_θ and the vision-language encoder E_I are both kept frozen throughout; only $\mathbf{d}$ is updated. This design ensures that the model's generative prior is not disturbed and that $\mathbf{d}$ genuinely captures the target transformation rather than overfitting to the specific images seen during training.

Semantic Alignment Loss. To provide a global semantic anchor, we leverage a frozen Vision-Language Encoder E_I [26]. For $\mathbf{d}$ to represent the intended semantic shift, it must be geometrically close to the edited images in the feature space of the encoder, while remaining distant from the originals. We formulate $\mathcal{L}_{\text{sem}}$ as a contrastive objective:

$$\mathcal{L}_{\text{sem}} = 1 - \text{cossim}(E_I(x'), \mathbf{d}) + \text{cossim}(E_I(x), \mathbf{d}) \tag{3}$$

Algorithm 1 Learning direction $\mathbf{d}$ with ViDiT

Require: Pre-trained diffusion model $\epsilon_\theta(x_t, c)$, pre-trained CLIP Image Encoder $E_I(x)$, a set of N image-edit pairs $\{(x_1, x'_1), \cdots, (x_N, x'_N)\}$ obtained from any editing source, randomly initialized direction $\mathbf{d}$, learning rate η.
while training **do**
  for $i = 1, \cdots, N$ **do**
    Sample pair (x_i, x'_i)
    Sample $\epsilon \sim \mathcal{N}(0, 1)$, $t \sim \mathcal{U}(1, T)$
    $x_{i,t} = x_i + \alpha^t \epsilon, \quad x'_{i,t} = x'_i + \alpha^t \epsilon$
    $\mathcal{L}_{\text{latent}} = -||\epsilon_\theta(x'_{i,t}, \mathbf{d}) - \epsilon_\theta(x_{i,t}, \mathbf{d})||_2^2$
    $\mathcal{L}_{\text{sem}} = 1 - \text{cossim}(E_I(x'_i), \mathbf{d}) + \text{cossim}(E_I(x_i), \mathbf{d})$
    $\mathcal{L} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{latent}}$
    $\mathbf{d} \leftarrow \mathbf{d} - \eta \nabla_{\mathbf{d}} \mathcal{L}$
  end for
end while

This loss places $\mathbf{d}$ in the correct conceptual neighbourhood of the target concept. However, because CLIP operates on globally pooled, high-level embeddings, it is insensitive to fine-grained spatial structure. A direction satisfying $\mathcal{L}_{\text{sem}}$ alone may capture the right concept yet still alter unrelated facial features, since CLIP cannot distinguish between a change in beard texture and a correlated change in jaw shape. This motivates our second loss term.

Latent Alignment Loss. To capture high-frequency structural details and enforce pixel-level correspondence between pairs, we introduce $\mathcal{L}_{\text{latent}}$. Unlike GANs, which condition on compact low-dimensional latent codes, diffusion models produce spatially-aligned, high-dimensional noise-prediction maps, $\epsilon_\theta(x_t, \mathbf{d})$, that encode dense per-pixel information. We exploit this property by maximising the difference between noise predictions for the noised edited image x'_t and the noised original x_t at random timesteps $t \sim \mathcal{U}(1, T)$:

$$\mathcal{L}_{\text{latent}} = -\mathbb{E}_{x_0, \epsilon, t}[||\epsilon_\theta(x'_t, \mathbf{d}) - \epsilon_\theta(x_t, \mathbf{d})||_2^2] \tag{4}$$

Intuitively, this encourages $\mathbf{d}$ to produce predictions that maximally distinguish the edited from the unedited image in the denoising network's internal feature space, grounding the direction in the model's native representation. In practice, we observe that $\mathcal{L}_{\text{latent}}$ is critical for preserving fine-grained identity cues such as hair follicle structure, specular reflections on glasses, and skin texture, details that $\mathcal{L}_{\text{sem}}$ alone cannot recover. The two losses are complementary: $\mathcal{L}_{\text{sem}}$ steers $\mathbf{d}$ towards the correct semantic concept, while $\mathcal{L}_{\text{latent}}$ sharpens its spatial precision. The total training objective is $\mathcal{L} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{latent}}$. We provide the training algorithm for ViDiT in Alg. 1.

3.3 The "Edit Anywhere" Principle

Once $\mathbf{d}$ has been optimized, a one-time process per editing concept that requires no fine-tuning of the base model, it effectively expands the diffusion model's

vocabulary with a new, reusable semantic token. At inference time, $\mathbf{d}$ can be injected into the CFG formulation to steer the denoising trajectory of *any* image towards the target concept, regardless of domain, style, or content.

Editing via Direction Injection. We modify the inference-time CFG to incorporate the learned direction:

$$\bar{\epsilon}_\theta(x_t, c, \mathbf{d}) = \tilde{\epsilon}_\theta(x_t, c) + \lambda_e(\epsilon_\theta(x_t, \mathbf{d}) - \epsilon_\theta(x_t, \phi)) \tag{5}$$

where the additional term replaces the standard text-conditioning residual with the visual-delta residual, and $\lambda_e \in \mathbb{R}$ provides continuous, granular control over edit intensity. Setting $\lambda_e = 0$ recovers the unedited output, while increasing λ_e progressively strengthens the effect. Crucially, this requires no per-image optimization or test-time adaptation: the same $\mathbf{d}$ applies zero-shot to any input.

Multi-directional Editing. A key advantage of the additive formulation in Eq. 5 is that multiple independently learned directions can be composed at inference time without retraining. Given a set of directions $D = \{d_1, \ldots, d_k\}$ with corresponding editing scales $\{\lambda_{e_1}, \ldots, \lambda_{e_k}\}$, the noise prediction generalizes to:

$$\hat{\epsilon}_\theta(x_t, c, D) = \tilde{\epsilon}_\theta(x_t, c) + \sum_{i=1}^{|D|} \lambda_{e_i}(\epsilon_\theta(x_t, d_i) - \epsilon_\theta(x_t, \phi)) \tag{6}$$

This allows complex combinations of attributes, for example, simultaneously applying a smile direction and an age direction, while preserving strong disentanglement between each component, since each d_i was optimized independently to isolate a single semantic concept.

Real Image Editing. ViDiT directions generalize robustly to real-world photographs. To edit a real image, we first invert it into the model's latent trajectory using DDPM Inversion [15], which produces a sequence of noisy latents $\{x_T, \ldots, x_0\}$ that approximately reconstruct the input when denoised. We then perform the reverse process with the direction-injected noise prediction:

$$\bar{\epsilon}_\theta(x_t, \mathbf{d}) = \epsilon_\theta(x_t, \phi) + \lambda_e(\epsilon_\theta(x_t, \mathbf{d}) - \epsilon_\theta(x_t, \phi)) \tag{7}$$

The success of this zero-shot transfer to real images confirms that $\mathbf{d}$ encodes a fundamental semantic vector within the diffusion model's latent space, rather than an overfitted representation tied to the distribution of training pairs. As demonstrated in Fig. 3, the learned directions transfer faithfully across synthetic portraits, animal photographs, and stylized artwork alike.

4 Experiments

To evaluate the effectiveness of ViDiT in transferring semantically rich latent directions, we conduct experiments across human faces, cats, and dogs. We demonstrate that by transferring visual deltas into a continuous direction $\mathbf{d}$, our framework achieves a "Learn Once, Edit Anywhere" capability that generalises across synthetic, real-world, and stylised images.

Experimental Setup. We utilise Stable Diffusion [27] for all experiments. For supervision, we primarily use off-the-shelf StyleGAN2 [17] models trained on FFHQ [16], AFHQ-Cats, and AFHQ-Dogs [5]. GANs are chosen as the primary supervision source because they currently provide the highest-quality disentangled image-edit pairs; however, ViDiT is source-agnostic and we additionally demonstrate directions transferred from StyleGAN-NADA [9] and Prompt2Prompt [12] in Fig. 3. We optimize $\mathbf{d}$ using N=1000 samples with the AdamW optimiser [22] for 1000 iterations. Once learned, $\mathbf{d}$ acts as a new "token" in the model's vocabulary, enabling zero-shot editing in $\sim$5 seconds on a single NVIDIA L40 GPU.

4.1 Ablation Studies

We evaluate four core design choices: the inference timestep T, the number of training pairs N, the choice of inversion method, and the contribution of each loss term. Identity preservation is assessed via LPIPS [38], DINO [23], and DreamSim [8], and semantic alignment via CLIP-T [26] and SigLIP-T [36]. Qualitative results are shown in Figs. 4 and 6 and quantitative results in Tab. 1.

Ablation on Inference Timestep. The injection timestep T controls the trade-off between edit strength and identity preservation. When $\mathbf{d}$ is applied at the full denoising timestep, the edit is introduced early in the generative process, allowing the model to gently incorporate the semantic change while maintaining structural coherence. Reducing to $0.4T$ applies the direction later, where the denoising network operates on lower-noise latents and produces more aggressive edits, useful when a strong effect is desired, but at the cost of noticeable identity drift, as shown in Fig. 4(a). In our default configuration we use the full timestep, which offers the best balance between edit fidelity and identity preservation.

Ablation on Sample Size. A key practical question is how many image-edit pairs are needed to learn a reliable direction. As shown in Tab. 1, ViDiT already captures meaningful semantic content with as few as N=10 pairs, demonstrating that the method is not data-hungry. However, DINO and LPIPS scores improve consistently as N increases, with N=1000 yielding the best identity preservation overall. This suggests that a larger pool of pairs provides broader coverage of the target concept's variation, leading to a more robustly disentangled direction that generalises better at inference time. Notably, CLIP-T peaks at N=100, indicating a mild trade-off between semantic alignment and structural faithfulness at very large N, which we address through our dual-loss formulation.

Ablation on Inversion Method. For real image editing, we ablate the choice of inversion strategy used to map the input into the model's latent trajectory. As shown in Fig. 6(b), DDIM Inversion [29] introduces noticeable drift in fine facial details, degrading structural consistency between the original and edited outputs. DDPM Inversion [15], by contrast, produces a more faithful latent trajectory that preserves high-frequency identity cues such as hair texture, glasses frames, and skin tone, yielding significantly cleaner edits. We therefore adopt DDPM Inversion as our default for real image editing throughout all experiments.

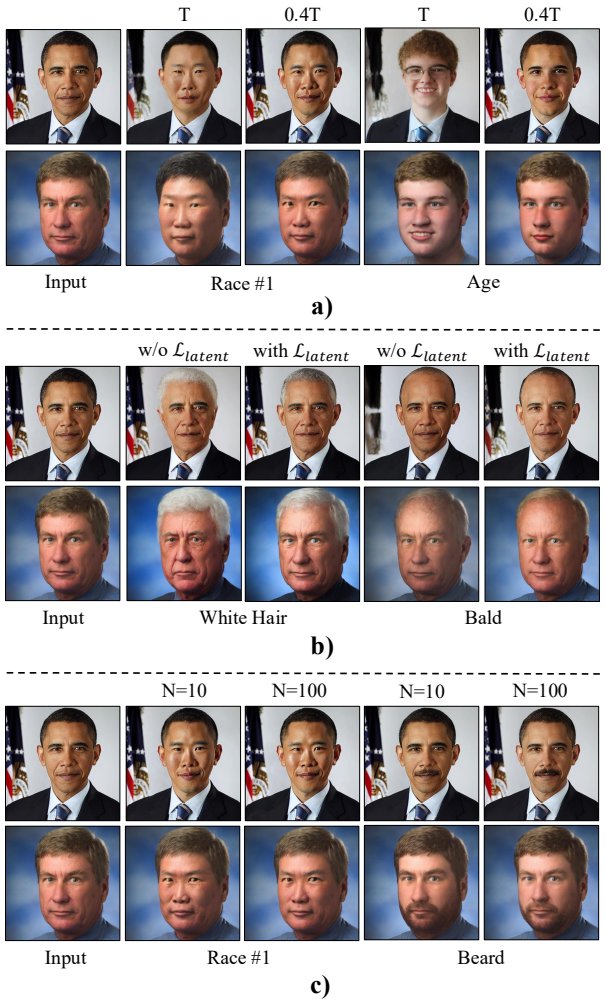

Fig. 4: Qualitative Ablation Results. (a) Effect of inference timestep T: applying the edit at a reduced timestep ($0.4T$) yields stronger edits at the cost of greater identity deviation, while the full timestep T produces more conservative, identity-preserving results. **(b)** Impact of $\mathcal{L}_{\text{latent}}$: without it, edits exhibit entanglement and unintended structural changes (e.g. skin tone shift in White Hair, over-smoothing in Bald); adding $\mathcal{L}_{\text{latent}}$ produces cleaner, more disentangled edits while preserving identity. **(c)** Effect of training sample size N: increasing from N=10 to N=100 improves edit sharpness and disentanglement, with diminishing returns suggesting that meaningful directions can be learned from very few pairs.

Ablation on Loss Terms. We isolate the contribution of $\mathcal{L}_{\text{latent}}$ by training with only the semantic alignment loss $\mathcal{L}_{\text{sem}}$. While $\mathcal{L}_{\text{sem}}$ alone is sufficient to identify the target concept, the resulting direction moves the output in the

Table 1: Quantitative Ablation Results. Analysis of sample size N and the impact of the Latent Alignment Loss ($\mathcal{L}_{\text{latent}}$). Each variant is evaluated with LPIPS [38], DINO [23], DreamSim [8], CLIP-T [26], and SigLIP-T [36].

Method	LPIPS↓	CLIP-T↑	DINO↑	SigLIP-T↑	DreamSim↑
N=10	<u>0.098</u>	0.403	<u>0.869</u>	0.143	**0.872**
N=100	0.136	**0.425**	0.842	**0.148**	0.771
w/o $\mathcal{L}_{\text{latent}}$	0.121	<u>0.409</u>	0.832	0.143	0.860
Ours (ViDiT)	**0.093**	0.406	**0.891**	<u>0.147</u>	<u>0.861</u>

correct semantic direction, the absence of $\mathcal{L}_{\text{latent}}$ leads to entangled edits that undesirably alter facial structure and skin tone, as visible in Fig. 4(b). This is because $\mathcal{L}_{\text{sem}}$ operates on globally pooled CLIP features, which lack the spatial resolution to enforce pixel-level identity consistency. Adding $\mathcal{L}_{\text{latent}}$ grounds $\mathbf{d}$ in the denoising network's dense, spatially-aligned prediction space, yielding a marked improvement in LPIPS and DINO without sacrificing semantic alignment, confirming that the two losses are complementary rather than competing.

4.2 Qualitative Results

Fig. 3 demonstrates ViDiT's editing capabilities across a wide range of semantics and domains. Our method transfers fine-grained directions spanning facial attributes such as age, gender, beard, and eyeglasses, as well as animal domains including cats and dogs. A key property of our edits is their high degree of disentanglement: only the targeted attribute is modified while all other aspects of the image, including identity, background, and artistic style, are faithfully preserved. Furthermore, the learned direction $\mathbf{d}$ generalizes robustly beyond its supervision source: edits optimized on structured GAN pairs apply equally well to stylized illustrations, classical paintings, and real-world photographs, confirming that $\mathbf{d}$ encodes a fundamental semantic vector within the diffusion model rather than an overfitted representation tied to the source domain.

Fig. 6(c) further demonstrates the compositional power of ViDiT. Since each direction is learned independently and injected additively at inference time, multiple directions can be combined simultaneously without retraining, producing coherent multi-attribute edits such as simultaneously applying *Asian*, *Beard*, and *Smile* while preserving strong disentanglement between each component.

4.3 Quantitative Comparisons

Comparison with Visual Conditioning Methods. To quantify the advantage of visual-delta supervision over text-based and visual conditioning alternatives, we evaluate all methods on 100 generated samples. As shown in Tab. 2, text-based and weight-space methods suffer from significant semantic

Fig. 5: Qualitative Comparison with Diffusion-based Image Editing Methods. We compare our approach with Concept Sliders [10], SEGA [2], InstructPix2Pix [3], and Prompt2Prompt [12]. ViDiT outperforms all baselines in achieving disentangled edits across both single semantics and combined attributes (e.g. "Race" and "Beard"), while faithfully preserving the identity and structure of the input.

entanglement, unintentionally altering subject identity to match coarse linguistic or model priors. By contrast, ViDiT outperforms all competing methods on identity-preservation metrics (CLIP-I, SigLIP-I, DINO) while remaining competitive in semantic alignment.

Comparison with Text-driven Editing Methods. We benchmark ViDiT against text-driven editing methods, SEGA [2], Prompt2Prompt [12], InstructPix2Pix [3], and Concept Sliders [10], on 200 input-edit pairs across "Asian" (Race #1) and "Smile" semantics. Identity preservation is evaluated with LPIPS [38] and DINO [23], and semantic alignment with CLIP-T [26] and SigLIP-T [36]. Importantly, our edits are applied without access to any target text prompt, relying solely on the direction transferred from visual pairs. As shown in Tab. 3, ViDiT significantly outperforms all text-driven baselines on identity preservation while remaining competitive in semantic alignment. An extended qualitative comparison against a broader set of baselines, including PnP-Diffusion [31], MasaCtrl [4], Asyrp [19], and DiffAE [25], is provided in Fig. 6(a), further confirming the superiority of ViDiT in preserving identity and disentangling targeted attributes.

User Study and Rescoring Analysis. We conducted a perceptual study with 40 participants on Prolific.com on real images, comparing ViDiT against SEGA [2], InstructPix2Pix [3], Prompt2Prompt [12], and Concept Sliders [10]. Participants rated 60 input-edit pairs on a 1–5 scale for semantic success and disentanglement quality. As reported in Tab. 4, ViDiT achieves the highest mean preference score, confirming that our zero-shot edits are perceived as more disentangled and semantically faithful than existing text-driven alternatives.

Table 2: Comparison with Visual Conditioning and Text-based Methods. Evaluated on 100 generated samples using CLIP-T [26], SigLIP-T [36], CLIP-I, SigLIP-I, and DINO [23].

Method	CLIP-T↑	SigLIP-T↑	CLIP-I↑	SigLIP-I↑	DINO↑
Text Prompt	**0.405**	0.135	0.775	0.875	0.822
Weights2Weights [7]	0.401	<u>0.146</u>	0.686	0.837	0.642
IP-Adapter [34]	0.394	**0.148**	0.816	0.890	0.767
Attribute-Control [1]	0.356	0.129	0.829	0.911	0.881
W+ Adapter [20]	0.355	0.130	0.723	0.824	0.759
Ours (ViDiT)	<u>0.395</u>	0.143	**0.831**	**0.913**	**0.886**

Table 3: Comparison with Text-driven Editing Methods. ViDiT is compared against text-driven baselines using LPIPS [38], DINO [23], CLIP-T [26], SigLIP-T [36], and DreamSim [8].

Method	LPIPS↓	CLIP-T↑	DINO↑	SigLIP-T↑	DreamSim↑
SEGA [2]	0.179	0.388	0.714	0.134	0.757
Prompt2Prompt [12]	0.074	**0.408**	0.867	<u>0.143</u>	0.869
InstructPix2Pix [3]	0.059	0.403	0.851	**0.145**	0.877
Concept Sliders [10]	0.121	0.325	0.842	0.097	0.844
Ours (ViDiT)	**0.030**	<u>0.407</u>	**0.929**	0.139	**0.905**

To further assess disentanglement, we perform a rescoring analysis measuring how CLIP classification probabilities for specific attributes shift following an edit, following [6, 28]. We apply four editing directions, Asian (Race #1), Smile, Gender, and Beard, to 100 Stable Diffusion-generated images and report the resulting shifts in Tab. 5. Targeted edits strongly increase the probability of the corresponding attribute (diagonal entries), while off-diagonal interactions remain minimal for unrelated attributes, confirming disentanglement. Some correlated shifts are semantically expected: the Gender edit towards femininity notably reduces the Beard score, and the Beard edit slightly diminishes Smile, consistent with known biases in the underlying generative model.

5 Limitations

While ViDiT faithfully represents the semantic transformation encoded in the input-edit pairs, the disentanglement of the learned direction is inherently bounded by the disentanglement of the supervision source. When pairs exhibit residual entanglement, the learned direction may reflect those characteristics. Additionally, due to biases present in CLIP and Stable Diffusion, certain attributes such as age can become entangled with correlated visual cues (e.g., white hair or eyeglasses). At present, ViDiT optimizes a single global direction per concept,

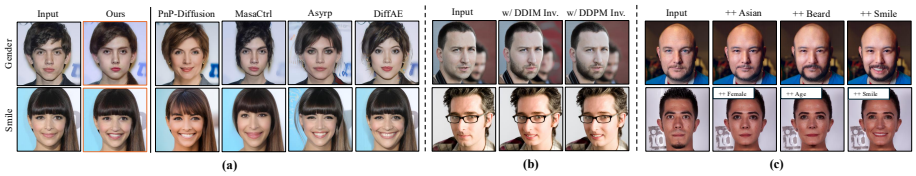

Fig. 6: Additional Qualitative Results. (a) Extended comparison with diffusion-based editing methods on *Gender* and *Smile* semantics. ViDiT produces more identity-preserving and disentangled edits compared to PnP-Diffusion [31], MasaCtrl [4], Asyrp [19], and DiffAE [25], which tend to alter unrelated facial attributes or global image structure. **(b)** Ablation on the inversion method: DDPM Inversion [15] yields superior structural consistency and identity preservation compared to DDIM Inversion [29], which introduces noticeable drift in fine facial details. **(c)** Compositional editing via multiple simultaneous directions. Applying *++Asian*, *++Beard*, and *++Smile* (top row), and *++Female*, *++Age*, and *++Smile* (bottom row) simultaneously produces coherent, disentangled multi-attribute edits without any retraining.

Table 4: User Study. Mean preference scores (1–5) from 40 participants on disentanglement and semantic faithfulness across 60 input-edit pairs.

Method	Preference ↑
SEGA [2]	1.96
InstructPix2Pix [3]	2.06
Prompt2Prompt [12]	2.74
Concept Sliders [10]	3.25
Ours (ViDiT)	**3.36**

Table 5: Rescoring Analysis. Shift in CLIP classification probability (%) per attribute. Rows: applied edit; columns: measured attribute.

	Asian	Smile	Gender	Beard
Asian	**53.6**	15.8	−12.1	−3.5
Smile	−20.4	**41.2**	11.8	−7.2
Gender	2.0	−4.3	**94.7**	−19.8
Beard	−2.6	−8.1	−0.05	**28.3**

which may limit expressiveness for highly localized or spatially varying edits where different regions of the image require different degrees of modification. As the broader ecosystem of editing methods continues to improve, ViDiT stands to directly benefit from higher-quality supervision sources, and exploring this further remains an interesting avenue for future work. Similarly, extending the framework to richer semantic concepts beyond facial and artistic attributes such as object-level or scene-level edits represents a natural and compelling direction for future investigation.

6 Ethics Statement

ViDiT does not introduce any new bias of its own. It learns directions by transferring semantics from pretrained models, including the supervision sources and the CLIP [26] and Stable Diffusion [27] models it builds on, and therefore inherits whatever biases those models already encode rather than creating additional

ones. Any demographic correlations observed in our edits, such as the residual cross-attribute shifts in our rescoring analysis (Tab. 5), originate in these pretrained components.

We anonymize racial attributes throughout: directions are referred to only by non-semantic numeric labels (e.g., "Race #1", "Race #2"), which identify directions inherited from the source latent space and carry no ordering, ranking, or preference for any group. We deliberately avoid explicitly naming any race, and our use of these directions is intended purely to demonstrate transfer capability across diverse semantics.

Finally, like all image synthesis and editing technologies, ViDiT carries a risk of misuse for the generation of deceptive or manipulated media [18]. We acknowledge this concern and encourage responsible deployment, including the development of appropriate detection and attribution tools alongside such capabilities.

7 Conclusion

We introduced ViDiT, a framework that expands the editing vocabulary of pretrained diffusion models by transferring visual deltas directly from image-edit pairs into continuous, transferable latent directions. Operating on a "Learn Once, Edit Anywhere" principle, ViDiT learns a single global direction from a small set of before-and-after examples without fine-tuning the base model or requiring per-image optimization at inference and applies it zero-shot to any image, including real photographs and stylized artwork.

By framing the problem as visual-delta transfer rather than text-prompt engineering, ViDiT overcomes the descriptive bottleneck of natural language and achieves significantly stronger identity preservation than text-driven and instruction-following baselines. The dual-objective training, combining semantic alignment via a frozen vision-language encoder with latent alignment via the denoising network's native prediction space, is key to achieving disentangled edits that preserve fine-grained identity cues. Quantitative benchmarks and a user study with 40 participants confirm that our zero-shot edits are perceived as more disentangled and semantically faithful than existing alternatives.

While we leverage structured generative models as a high-quality supervision source, the framework is entirely source-agnostic: directions learned from domain-adaptation models, diffusion-based editors, or any method producing before-and-after image pairs are equally well supported, pointing toward a general paradigm for incorporating complex visual semantics into diffusion models from any generative source.

8 Acknowledgements

This research is supported by the National Science Foundation under Grant No. 2543524.

References

1. Baumann, S.A., Krause, F., Neumayr, M., Stracke, N., Sevi, M., Hu, V.T., Ommer, B.: Continuous, Subject-Specific Attribute Control in T2I Models by Identifying Semantic Directions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
2. Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., Kersting, K.: SEGA: Instructing text-to-image models using semantic guidance. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=KIPAIy329j
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
5. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
6. Dalva, Y., Yanardag, P.: Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. arXiv preprint arXiv:2312.05390 (2023)
7. Dravid, A., Gandelsman, Y., Wang, K.C., Abdal, R., Wetzstein, G., Efros, A.A., Aberman, K.: Interpreting the weight space of customized diffusion models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems
8. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
9. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators (2021)
10. Gandikota, R., Materzyńska, J., Zhou, T., Torralba, A., Bau, D.: Concept sliders: Lora adaptors for precise control in diffusion models. arXiv preprint arXiv:2311.12092 (2023)
11. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. arXiv preprint arXiv:2004.02546 (2020)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. arXiv preprint arXiv:2304.06140 (2023)
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
17. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)

18. Korshunov, P., Marcel, S.: Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
19. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. arXiv preprint arXiv:2210.10960 (2022)
20. Li, X., Hou, X., Loy, C.C.: When stylegan meets stable diffusion: a $\mathcal{W}_+$ adapter for personalized image generation. arXiv preprint arXiv: 2311.17461 (2023)
21. Liu, N., Du, Y., Li, S., Tenenbaum, J.B., Torralba, A.: Unsupervised compositional concepts discovery with text-to-image generative models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2085–2095 (October 2023)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
24. Park, Y.H., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the latent space of diffusion models through the lens of riemannian geometry. arXiv preprint arXiv:2307.12868 (2023)
25. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10619–10629 (2022)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
30. Song, K., Han, L., Liu, B., Metaxas, D., Elgammal, A.: Stylegan-fusion: Diffusion guided domain adaptation of image generators. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5453–5463 (2024)
31. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
32. Wu, C.H., De la Torre, F.: 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. pp. 7378–7387 (2023)
33. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. arXiv preprint arXiv:2011.12799 (2020)
34. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
35. Yüksel, O.K., Simsar, E., Er, E.G., Yanardag, P.: Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In: Proceedings of

the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14263–14272 (October 2021)
36. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
37. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)