

HypDAE: Hyperbolic Diffusion Autoencoders for Hierarchical Few-shot Image Generation

Lingxiao Li^{1,2} Kaixuan Fan¹ Boqing Gong^{2*} Xiangyu Yue^{1*}

¹ MMLab, The Chinese University of Hong Kong ² Boston University
{lxli, bgong}@bu.edu, kxfan127@gmail.com, xyyue@ie.cuhk.edu.hk

Abstract

Few-shot image generation aims to generate diverse and high-quality images for an unseen class given only a few examples in that class. A key challenge in this task is balancing category consistency and image diversity, which often compete with each other. Moreover, existing methods offer limited control over the attributes of newly generated images. In this work, we propose Hyperbolic Diffusion Autoencoders (HypDAE), a novel approach that operates in hyperbolic space to capture hierarchical relationships among images from seen categories. By leveraging pre-trained foundation models, HypDAE generates diverse new images for unseen categories with exceptional quality by varying stochastic subcodes or semantic codes. Most importantly, the hyperbolic representation introduces an additional degree of control over semantic diversity through the adjustment of radii within the hyperbolic disk. Extensive experiments and visualizations demonstrate that HypDAE significantly outperforms prior methods by achieving a better balance between preserving category-relevant features and promoting image diversity with limited data. Furthermore, HypDAE offers a highly controllable and interpretable generation process. Code is available at: <https://github.com/lingxiao-li/HypDAE>.

1. Introduction

Generative models [24, 49, 50, 54] have succeeded in generating high-fidelity and realistic images, partially thanks to a large volume of high-quality data for model training. However, with the widespread presence of long-tail distributions and data imbalances across image categories [27], there are many scenarios in the real world where it is impossible to collect sufficient samples of certain categories for model training. It is difficult for generative models trained on well-sampled categories to generate realistic and diverse images

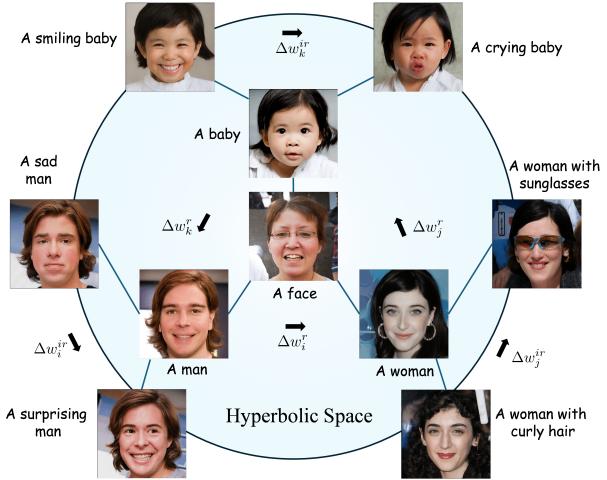


Figure 1. **Illustration of hierarchical text-image representation in hyperbolic space.** Hyperbolic space provides a natural and compact encoding for semantic hierarchies in large datasets. Adjusting high-level, identity-relevant attributes Δw^r alters an image's identity, while modifying low-level, identity-irrelevant attributes Δw^{ir} produces variations within the same identity.

for a novel category given only a few examples. This challenging task is known as few-shot image generation, which aims to synthesize images that preserve the category-level identity of the limited input samples [8, 13, 14, 25–29, 37].

Existing few-shot image generation methods are primarily GAN-based and fall into three categories: *transfer-based* approaches [8, 38], which use meta-learning or domain adaptation for cross-category generalization but often face limited transferability; *fusion-based* approaches [2, 19, 26, 27, 61, 66], which fuse features from multiple exemplars but tend to produce outputs overly similar to the inputs; and *transformation-based* approaches [1, 13, 14, 28, 29], which apply intra-category perturbations without fine-tuning but often lack diversity. Recent diffusion-based methods [5, 33, 51] for *object-level personalization* focus on instance identity, require test-time fine-tuning, and depend on prompt engineering, making them incompatible with our category-level setting without textual inputs or model up-

*Corresponding authors

dates. However, all these methods operate in Euclidean feature space, which limits their ability to capture hierarchical semantic structure—a key requirement for generalizing from few examples in category-level few-shot image generation.

In contrast to prior methods, recent work such as HAE [37] highlights the importance of modeling hierarchical structures in few-shot image generation. Similar to language [12, 41, 57], images also exhibit semantic hierarchies [9, 31], where each image can be viewed as a composition of attributes at different levels. As illustrated in Fig. 1, high-level, identity-relevant attributes (e.g., gender or age) define the semantic core of a category, while low-level, identity-irrelevant attributes (e.g., expression or hairstyle) introduce intra-class variation. Capturing this hierarchical organization is essential for generating diverse yet category-consistent images. Hyperbolic space provides a natural embedding for such structures, as it can represent tree-like relationships with low distortion [41]. This enables infinite layers of semantics to be encoded compactly, allowing identity-preserving edits along radial directions and intra-category diversity through tangential shifts in the latent space, which facilitates a structured and interpretable latent space that supports controllable few-shot image generation.

Despite the advantages of hierarchical representation for few-shot image generations [37], existing methods are dominated by GAN-based approaches and face three primary challenges: **1)** Suboptimal image quality due to the generative constraints of GANs, particularly with insufficient training data; **2)** Reduced diversity from the 1-to-1 mapping between hyperbolic and image spaces, which leads to the loss of high-frequency details when latent codes are insufficiently trained; and **3)** The necessity of labeled data for learning hierarchical latent representations. In parallel, recent advances in diffusion models [5, 33, 47, 51, 62] enable high-quality image generation with rich, diverse details. Moreover, pre-trained foundation models (e.g., Stable Diffusion [50], CLIP [48]) further support adaptation with limited data and strong generalization.

To address the challenges, we propose **Hyperbolic Diffusion AutoEncoders (HypDAE)**, a novel method for leveraging the natural suitability of hyperbolic space for hierarchical latent code manipulation and the generative ability of diffusion models for few-shot image generation. In particular, we separate the representation of given images into two subcodes [47]. Our method begins by training an image encoder to capture high-level semantic subcode, while utilizing a pre-trained Stable Diffusion (SD) [50] model for decoding and modeling stochastic subcode by reversing the generative process. A hyperbolic feature encoder and decoder are then trained to map latent vectors from Euclidean space \mathbb{R}^n to hyperbolic space \mathbb{D}^n , with classification loss ensuring hierarchical image embeddings. Finally, to reduce the data labeling cost for training a hyperbolic encoder, pseudo-labeling is

enabled by zero-shot classification with pre-trained vision models. By capturing the attribute hierarchy among images, **HypDAE** generates new images through two methods: **1)** varying stochastic subcode of the Diffusion Autoencoders, and **2)** randomly shifting semantic subcode in an identity-irrelevant direction to modify category-irrelevant features. Hyperbolic space enables control over the semantic diversity of generated images by adjusting the radii in hyperbolic space. This facilitates hierarchical attribute editing for flexible, high-quality, and diverse few-shot image generation.

Our contributions are summarized as follows:

- We introduce **HypDAE**, an innovative method for few-shot image generation, which learns the hierarchical representation of images in hyperbolic space. To the best of our knowledge, **HypDAE** is the first to enable diffusion models to model semantic hierarchical information in hyperbolic space for few-shot image generation.
- We demonstrate that in our hyperbolic latent space, semantic hierarchical relationships among images are reflected by their distances to the center of the Poincaré disk. This allows for easy interpretation of the latent space.
- By leveraging pre-trained foundation models, **HypDAE** generates high-quality images with rich, diverse, category-irrelevant features from a single image. Extensive experiments show that **HypDAE** significantly outperforms existing few-shot image generation methods without requiring human-annotated class labels, offering state-of-the-art quality, diversity, and flexibility. This demonstrates the robustness of **HypDAE** under settings with automatically generated labels, making it a strong solution for label-free few-shot image generation.

2. Related Work

Few-shot Image Generation. Given a few samples from one novel category, few-shot image generation aims to produce more new images with high quality and various diversity. Recent advances in few-shot image generation encompass various approaches. *Transfer-based methods* [8, 38, 65], utilizing meta-learning or domain adaptation on GANs, often struggle with generating realistic images. *Fusion-based methods*, aligning random vectors with conditional images [26] or framing generation as a conditional task [19, 27, 61], generally produce outputs with limited diversity. *Transformation-based approaches* [28, 29, 37], which transform single conditional images, can lack consistency. Ding *et al.* [13, 14] explored intra-category transformations as identity-irrelevant edits using a single sample. More recently, Diffusion-based *subject-driven generation* [10, 15, 20, 59] adapts models to incorporate new concepts with limited data, exemplified by DreamBooth [51], which assigns unique identifiers to subjects using 3-5 images. This has inspired methods like Custom Diffusion [33] and

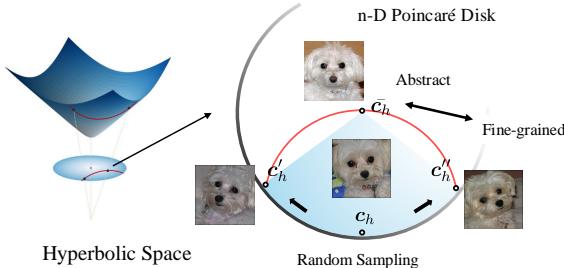


Figure 2. Illustration of the property of hyperbolic space on the Poincaré disk. Given two latent codes of Maltese dog c'_h and c''_h on the edge of Poincaré disk, the geodesic between these two points is the red curve rather than a straight line in Euclidean space. Therefore, their average latent code is calculated as \bar{c}_h , which is closer to the center. Thus, \bar{c}_h can be viewed as the “parent” of c_h , c'_h , and c''_h . One can generate diverse images without changing the category by moving the latent code from one child to another of the same parent in the hyperbolic space.

further improvements [7, 21, 52]. However, these works focus on instance-level identity preservation, require test-time fine-tuning, and rely heavily on prompt engineering. Therefore, their objectives, problem settings, and evaluation protocols differ fundamentally from our task.

Hyperbolic Representation Learning. Recent advancements have introduced hyperbolic space into deep learning [31, 41, 42, 55, 57], initially in NLP for hierarchical language modeling [41, 42, 57]. Riemannian optimization techniques have enabled model training within hyperbolic spaces [3, 4]. Building on this, Ganea *et al.* [16] developed hyperbolic adaptations of core neural network tools such as multinomial logistic regression, feed-forward, and recurrent networks. Applications of hyperbolic geometry have since expanded to image [31, 37], video [55], graph data [6, 44], and 3D shape generation [36]. Desai *et al.* [11] introduced a hyperbolic contrastive model for hierarchical text-image pairing, while HAE [37] demonstrated hyperbolic space’s advantages in few-shot image generation, although GAN-based results remained limited in quality. To our knowledge, this work is the first to integrate hyperbolic representation with diffusion models for high-quality few-shot image generation.

3. Method

To generate diverse new images from a few reference images while preserving their identity, it is essential for our model to capture both identity-relevant features (to maintain identity) and identity-irrelevant features (to enhance diversity). To achieve this, we design a diffusion autoencoder that extracts elementary identity-relevant features via a semantic encoder and identity-irrelevant features via a stochastic encoder. Furthermore, a hyperbolic encoder-decoder is introduced to further disentangle high-order semantic features. Unlike HAE, which applies hyperbolic representations in GANs, our method integrates hyperbolic semantics into diffusion

models. We introduce a stable training pipeline, and encode hyperbolic embeddings as diffusion conditions, enabling scalable and interpretable few-shot generation.

We give details of our method in the following sections. Specifically, Sec. 3.1 details the modeling of hierarchical representations, Sec. 3.2 elaborates on the framework and loss functions of HypDAE, and Sec. 3.3 and Sec. 3.4 discuss the pseudo-labeling process and hyperbolic latent editing algorithms, respectively.

3.1. Preliminary

Hierarchical Learning in Hyperbolic Space. A key challenge in multi-level semantic editing is deriving a hierarchical representation from real images, as illustrated in Fig. 2. To address this, we utilize *hyperbolic* space as the latent space, given its suitability for hierarchical structures. Unlike Euclidean spaces (zero curvature) or spherical spaces (positive curvature), hyperbolic spaces exhibit negative curvature, making them ideal for modeling hierarchical data [41, 42]. As a continuous analog of trees [41], hyperbolic space enables hierarchical representation through its exponential radius growth, facilitating structured modeling across text, images, and videos [18].

The n -dimensional hyperbolic space, \mathbb{H}^n , is a homogeneous, simply connected Riemannian manifold with constant negative sectional curvature¹. We use the Poincaré disk model following [31], $(\mathbb{D}^n, g^{\mathbb{D}})$, a preferred choice in gradient-based learning [16, 31, 41, 42, 55, 57], defined by $\mathbb{D}^n = \{x \in \mathbb{R}^n : \|x\| < 1\}$ with the Riemannian metric:

$$g_x^{\mathbb{D}} = \lambda_x^2 \cdot g^E, \quad (1)$$

where $\lambda_x = \frac{2}{1 - \|x\|^2}$, and g^E is the Euclidean metric tensor $g^E = \mathbf{I}^n$. The induced distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{D}^n$ can be defined by:

$$d_{\mathbb{D}}(\mathbf{x}, \mathbf{y}) = \text{arccosh} \left(1 + 2 \frac{\|\mathbf{x} - \mathbf{y}\|^2}{(1 - \|\mathbf{x}\|^2)(1 - \|\mathbf{y}\|^2)} \right). \quad (2)$$

A geodesic is defined as a locally minimized-length curve between two points, as illustrated in Fig. 2. The midpoint of this geodesic, closer to the origin, represents the mean of two latent codes in hyperbolic space. This property ensures that the mean of two leaf embeddings is not another leaf but their hierarchical parent, as illustrated in Fig. 2 on a 2-D Poincaré disk [55]. Embeddings near the disk’s edge correspond to fine-grained images, whereas those closer to the center represent more abstract features, such as an average face.

3.2. Hyperbolic Diffusion AutoEncoders (HypDAE)

The overall pipeline of HypDAE, illustrated in Fig. 3, consists of two stages: **1**) In Stage I, a semantic encoder maps

¹We set the curvature $c = -1$ in this paper.

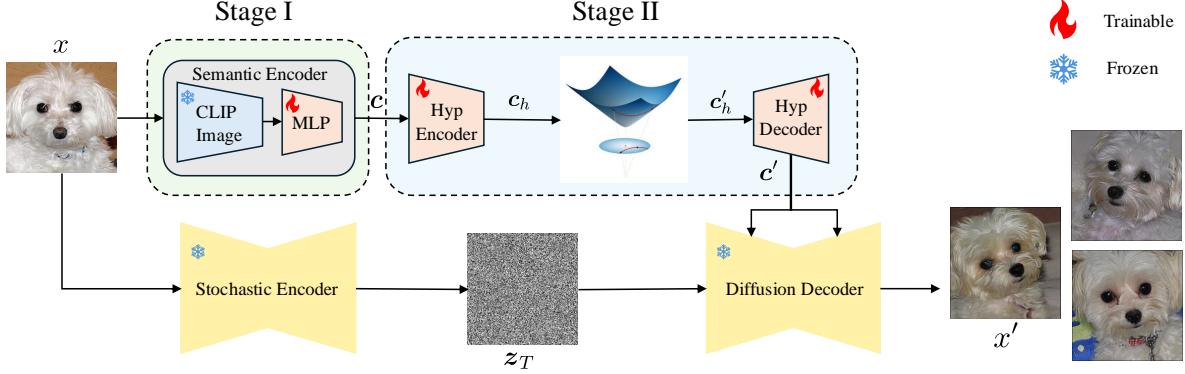


Figure 3. The overview of **HypDAE**. The hyperbolic autoencoder consists of a “semantic” encoder that maps the reference image to the semantic code ($x \rightarrow \mathbf{c}$), and a stable diffusion model that acts both as a “stochastic” encoder ($x \rightarrow \mathbf{z}_T$) and a diffusion decoder ($(\mathbf{c}', \mathbf{z}_T) \rightarrow x'$). Here, \mathbf{c} captures the high-level semantics while \mathbf{z}_T captures low-level stochastic variations; they can be decoded back with high fidelity. The “Hyp Encoder” is used to project the latent code from Euclidean space \mathbb{R}^n to hyperbolic space \mathbb{D}^n [6]. The semantic code can be edited as Fig. 2 shows and be used as the condition for the diffusion decoder to generate diverse new images with the same category. The VAE module of SD is omitted for better illustration.

input images to meaningful latent codes, serving as conditions for the pre-trained Stable Diffusion (SD). To enhance diversity and prevent direct replication, a content bottleneck and strong augmentation techniques are applied. 2) In Stage II, a hyperbolic encoder-decoder projects visual features from Stage I into hyperbolic space using supervised classification loss and then reconstructs them in Euclidean space. This process captures hierarchical relationships within the image corpus, facilitating the generation of diverse images by manipulating latent codes within the same category.

We design a two-stage model for two reasons: 1) jointly and end-to-end optimizing the whole pipeline is challenging for satisfactory results as multiple loss functions must be optimized for various image generation functionalities; 2) end-to-end training requires extensive labeled data, which is difficult to obtain. By staging the pipeline, we can leverage existing datasets collected for general vision tasks and avoid the need for new task-specific paired data.

Diffusion Autoencoders. In this stage, we aim to generate new images x' of the reference image x with rich stochastic details (e.g., hair, fur, color, etc.) while maintaining the identity and category of the image unchanged.

To address the limitations of GAN-based methods, we adopt the approach of DiffAE [47], leveraging a pre-trained SD model. The **SD decoder**, $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$, predicts noise based on timestep t and a high-level semantic condition \mathbf{c} , which is learned by a **semantic encoder** mapping input image x to a meaningful latent code \mathbf{c} . SD also functions as a **stochastic encoder** that captures a low-level “stochastic” subcode \mathbf{z}_T , as illustrated in Fig. 3.

For the original SD models, the condition \mathbf{c} is the given text and is usually processed by a pre-trained CLIP [48] text encoder, outputting 77 tokens. Hence, a naive solution is to directly replace it with CLIP image embeddings to capture high-level semantics. However, this naive solution

makes the model easy to remember instead of understanding the context information and copying the content, arriving at a trivial solution. We apply two tricks to avoid this: 1) **Strong data augmentations** \mathcal{A} (e.g., flip, rotation, blur, and elastic transform) on the reference image x to break down the connection with the source image and encourage identity-invariant feature learning; 2) **Content Bottleneck**. To promote diversity, we use only the CLIP image encoder’s class token. It compresses the reference image from spatial size $224 \times 224 \times 3$ to a vector of dimension 1×1024 for a compact representation, aligning it with the original CLIP text feature space of SD via additional fully connected layers (MLP) that inject features into the diffusion process through cross-attention. This mechanism forces the semantic encoder to focus on the main object while ignoring backgrounds and identity-irrelevant features. Thus, given reference image x , the condition is defined as: $\mathbf{c} = \text{MLP}(\text{CLIP}(\mathcal{A}(x)))$.

In this setup, noise is progressively added to \mathbf{z}_0 (x encoded by VAE) to produce a noisy latent \mathbf{z}_t , where t denotes the number of noise additions. To utilize the strong image prior learned by SD and CLIP, only the MLP is trainable during the training process. The UNet ϵ_θ learns to predict the added noise with the condition \mathbf{c} :

$$\mathcal{L}_{align} = \mathbb{E}_{\mathbf{z}_0, t, \mathbf{c}, \epsilon \sim \mathcal{N}(0, 1)} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2]. \quad (3)$$

Besides decoding, the SD model can also be used to encode an input image latent \mathbf{z}_0 to the stochastic subcode \mathbf{z}_T by running DDIM sampling process backward [53]:

$$\mathbf{z}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} \mathbf{z}_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c}). \quad (4)$$

For the definition of α_t and additional details, please refer to Sec. B of the supplementary material (SM).

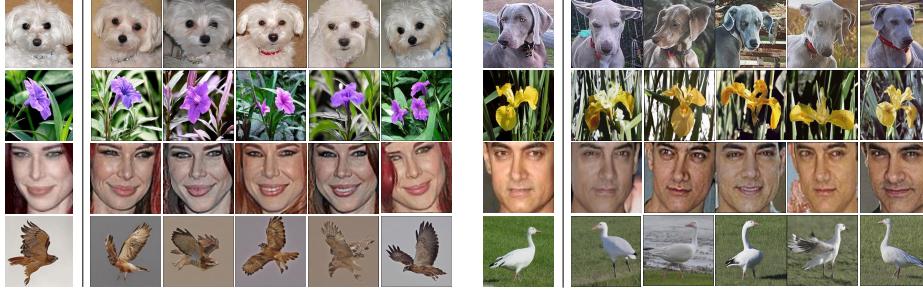


Figure 4. One-shot image generation from HypDAE on Animal Faces, Flowers, VGGFaces, and NABirds.

This process functions as a stochastic encoder, where z_T retains details not captured by the limited-capacity semantic representation c . By integrating both semantic and stochastic encoders, the diffusion autoencoder preserves full image details while providing a high-level representation c for downstream tasks. Notably, the stochastic encoder is omitted during training (Eq. (3)) and applied post-training to compute z_T , enabling control over image diversity.

Hyperbolic Image Encoder. Despite the category-invariant feature extraction in Stage I, the generated image variations remain limited. Thus, Stage II aims to capture hierarchical relationships within the image, enabling latent code editing to enhance diversity while preserving object identity.

To manipulate latent code in hyperbolic space, we need to define a bijective map from \mathbb{R}^n to \mathbb{D}_c^n to map Euclidean vectors to the hyperbolic space and vice versa. A manifold is a differentiable topological space that locally resembles the Euclidean space \mathbb{R}^n [34, 35]. For $x \in \mathbb{D}^n$, one can define the tangent space $T_x \mathbb{D}_c^n$ of \mathbb{D}_c^n at x as the first-order linear approximation of \mathbb{D}_c^n around x . Therefore, this bijective map can be performed by exponential and logarithmic maps. Specifically, the *exponential map* $\exp_x^c : T_x \mathbb{D}_c^n \cong \mathbb{R}^n \rightarrow \mathbb{D}_c^n$, maps from the tangent spaces into the manifold. While the *logarithmic map* $\log_x^c : \mathbb{D}_c^n \rightarrow T_x \mathbb{D}_c^n \cong \mathbb{R}^n$ is the reverse map of the exponential map.

We train a hyperbolic image encoder to map latent codes from Euclidean to hyperbolic space using exponential and logarithmic maps at the origin $\mathbf{0}$. Given a latent vector $c \in \mathbb{R}^{1 \times 1024}$ from Stage I in CLIP-space, a 5-layer single-head Transformer encoder E reduces its dimensionality to $\mathbb{R}^{1 \times 512}$ before mapping it to hyperbolic space via the exponential map. A hyperbolic feed-forward layer [16] then produces the hierarchical representation c_h , formulated as: $c_h = f^{\otimes c}(\exp_0^c(E(c)))$, where $f^{\otimes c}$ denotes the Möbius linear layer mapping Euclidean to hyperbolic space. To project c_h back to the CLIP image space and reconstruct the latent code c' , we apply a logarithmic map followed by a 30-layer single-head Transformer decoder D: $c' = D(\log_0^c(c_h))$, which is then fed into the cross-attention layer of the SD model to reconstruct the image x' .

Loss functions. To learn a semantic hierarchy in hyperbolic space, we minimize the distance between latent codes of

similar images while increasing the distance between dissimilar ones. For multi-class classification on the Poincaré disk (Sec. 3.1), we extend multinomial logistic regression (MLR) [16] to hyperbolic space by incorporating a linear layer and computing the **Hyperbolic Loss** via the *negative log-likelihood* (NLL):

$$\mathcal{L}_{\text{hyper}} = -\frac{1}{N} \sum_{n=1}^N \log(p_n), \quad (5)$$

where N is the batch size and p_n represents the predicted probability of the correct class.

As discussed in Sec. 3.1, distances in the Poincaré disk grow exponentially with radius. To minimize Eq. (5), fine-grained image embeddings are pushed toward the disk’s boundary to maximize inter-category distances, while abstract image embeddings—representing shared features across categories—remain near the center.

To ensure that the network projects back to the CLIP image space, we use L-2 distance **reconstruction loss** and **cosine similarity loss** to reconstruct c' as the same as the original c in CLIP image space:

$$\mathcal{L}_{\text{rec}}(c, c') = \|c - c'\|_2 + 1 - \cos(c, c'). \quad (6)$$

The **overall loss function** is:

$$\mathcal{L} = \mathcal{L}_{\text{hyper}} + \lambda \cdot \mathcal{L}_{\text{rec}}, \quad (7)$$

where λ is a trade-off adaptive parameter. This curated set of loss functions ensures the model learns the hierarchical representation and reconstructs images. More details can be found in Sec. A of the supplementary material.

3.3. Pseudo-Labeling

As discussed in Sec. 1, a key limitation of previous few-shot image generation methods is the reliance on labeled data, which is also evident in our earlier approach (*i.e.*, Eq. (5)). To mitigate this, we leverage a pre-trained CLIP model to predict the pseudo-label. Given an image and a predefined set of class names (*e.g.*, Golden Retriever, Spaniel), we utilize CLIP to compute semantic similarities between the image embedding and class embeddings. The pseudo-label is determined by selecting the class with the highest similarity score,

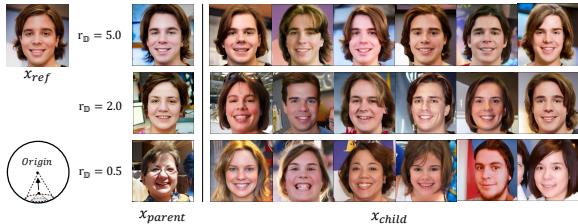


Figure 5. **Images with hierarchical semantic similarity generated by HypDAE** by sampling “child” images of their “parent” images, where r_D represents the hyperbolic distance from the “parent” images to the center of the Poincaré disk. As the reference image x moves from the edge to the center (from fine-grained/certain to abstract/ambiguous), the “children” of the reference image become more diverse and less similar to the reference image.

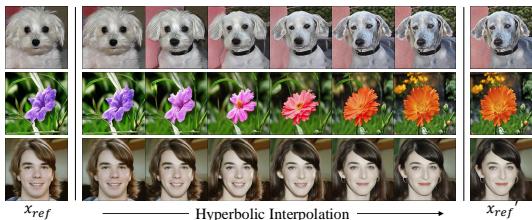


Figure 6. **Interpolations in hyperbolic space along the edge of the Poincaré disk** (with $r_D = 6.2126$) on four datasets. Zoom in to see the details.

which represents the closest semantic match to the image. This process enables label-free classification by leveraging CLIP’s cross-modal understanding capabilities. See the Sec. A in supplementary material for further details.

3.4. Hyperbolic Latent Editing

To generate diverse images with varying identity-irrelevant features, we edit hyperbolic latent codes via interpolation between two images or by applying random perturbations. In hyperbolic space, the shortest path between two points is defined by the geodesic under the induced distance (Eq. (2)). The geodesic between two embeddings, c_{hi} and c_{hj} , is denoted as $\gamma_{c_{hi} \rightarrow c_{hj}}(t)$. For perturbation-based generation, we first rescale the embedding c_h of a given image x_i to a fixed radius r_D , then sample a random vector c_{hj} from seen category embeddings, also constrained to r_D . The geodesic between them defines the perturbation direction, enabling diverse image generation. More details and formulas are provided in Sec. C of the appendix.

4. Experiment

4.1. Implementation Details

We choose Stable Diffusion V2.1 [50] as the base generator for the base generative model. During training, we set the image resolution to 512×512 . The dimension of the latent code in hyperbolic space is chosen to be 512. More details can be found in the Sec. A of the supplementary material.

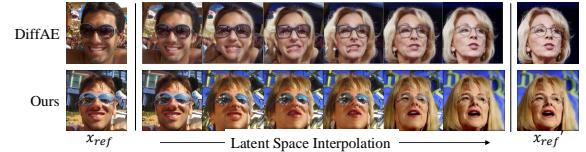


Figure 7. **Comparison of the latent space interpolations** between DiffAE [47] (Euclidean Space) and HypDAE (Hyperbolic Space).

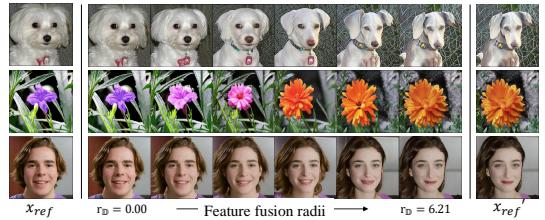


Figure 8. **Feature fusion from the edge to the center of the Poincaré disk** between two images on four datasets. Zoom in to see the details.

4.2. Datasets

We evaluate our method on Animal Faces [39], Flowers [43], VGGFaces [45], FFHQ [30], and NABirds [58] following the settings described in [14, 37]. Due to the low resolution (64×64) of the VGGFaces dataset, we use FFHQ [30] to fine-tune the model pre-trained on VGGFaces without supervision and visualize images of human faces with FFHQ.

4.3. Analysis of Hierarchical Feature Editing

We examine the properties of the hierarchical representations, focusing on how attribute levels correspond to latent code locations in hyperbolic space. As noted in Sec. 3.1, there is a continuum from fine-grained to abstract attributes, corresponding to points from the periphery to the center of the Poincaré disk. We define the hyperbolic radius r_D ² as the distance of a latent code to the disk’s center. To investigate the effect of r_D , we conduct experiments with various radii.

Hierarchical Image Sampling and Interpolation. Fig. 10 shows 2-D embeddings of Animal Faces in the Poincaré disk, visualized via UMAP [40]. By varying the radii of “parent” images, HypDAE controls the semantic diversity of generated images (Fig. 5). Results indicate that identity-relevant attributes change below $r_D \approx 2.0$, while identity-irrelevant attributes vary above $r_D \approx 5.0$, demonstrating that r_D correlates with attribute levels. As r_D decreases, identities become more abstract, resulting in greater diversity among “children” images, as seen in Fig. 12. Smooth interpolation between attributes is achievable in hyperbolic space (Fig. 6) without distortion, as shown in Fig. 7, confirming that HypDAE enables geodesic, hierarchical editing.

Hierarchical Feature Fusion. HypDAE also supports infinite semantic levels for attribute fusion. In Fig. 8, attributes from two images are combined at varying levels along the radius from the disk’s edge to the center, achieving smooth

²The radius of the Poincaré disk in our experiment is about 6.2126

Method	Settings	Flowers		Animal Faces		VGG Faces*		NA Birds	
		FID(\downarrow)	LPIPS(\uparrow)						
DAWSON [38]	3-shot	188.96	0.0583	208.68	0.0642	137.82	0.0769	181.97	0.1105
F2GAN [27]	3-shot	120.48	0.2172	117.74	0.1831	109.16	0.2125	126.15	0.2015
WaveGAN [61]	3-shot	42.17	0.3868	30.35	0.5076	4.96	0.3255	-	-
F2DGAN [66]	3-shot	38.26	0.4325	25.24	0.5463	<u>4.25</u>	0.3521	-	-
DeltaGAN [28]	1-shot	109.78	0.3912	89.81	0.4418	80.12	0.3146	96.79	0.5069
SAGE [14]	1-shot	43.52	0.4392	27.43	0.5448	34.97	0.3232	19.45	0.5880
HAE [37]	1-shot	50.10	0.4739	26.33	0.5636	35.93	0.5636	21.85	0.6034
LSO [65]	1-shot	35.87	0.4338	27.20	0.5382	4.15	0.3834	-	-
HypDAE (Euc)	1-shot	25.06	0.7420	20.72	0.7288	6.52	0.5429	8.40	0.7904
HypDAE (Real)	1-shot	23.96	<u>0.7595</u>	<u>14.31</u>	<u>0.7415</u>	6.25	0.5685	<u>7.64</u>	<u>0.7959</u>
HypDAE (Pseudo)	1-shot	<u>24.43</u>	0.7630	13.14	0.7431	5.96	<u>0.5560</u>	7.57	0.7966

Table 1. FID(\downarrow) and LPIPS(\uparrow) of images generated by different methods for unseen categories on four datasets. **Bold** indicates the best results and underline indicates the second best results. VGGFaces is marked with * because different methods report different numbers of unseen categories on this dataset (*e.g.* 552 in LoFGAN, 96 in DeltaGAN, 497 in L2GAN, and 572 in HypDAE, HAE, and SAGE). We do not compare with diffusion-based models for editing or customization purposes as none of them focus on our task in the same settings.

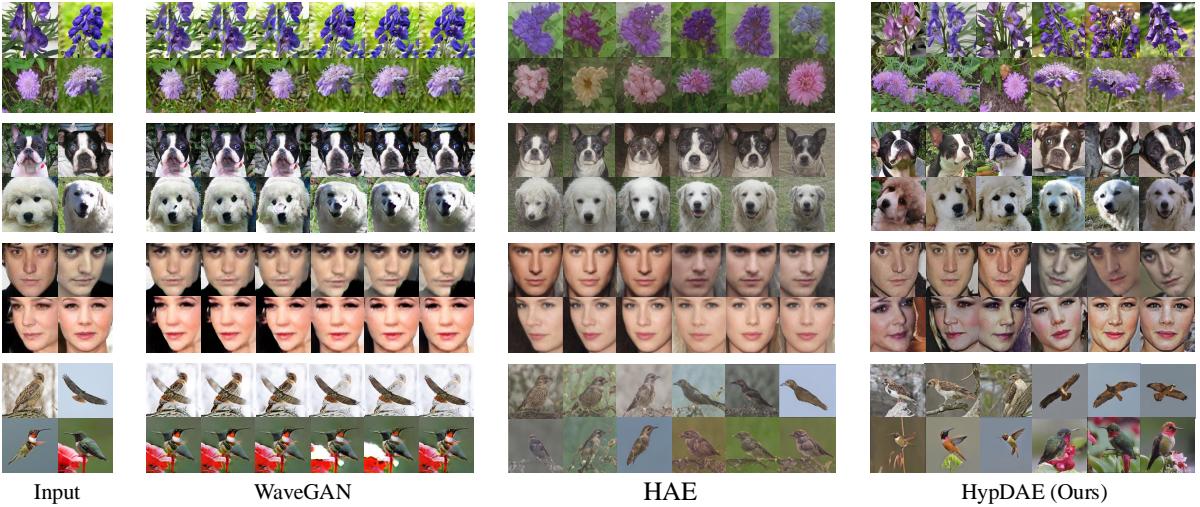


Figure 9. **Comparison between images generated by WaveGAN, HAE, and HypDAE** on Flowers, Animal Faces, VGGFaces and NABirds. Note: WaveGAN uses a 2-shot setting; HAE and HypDAE are both in a 1-shot setting. **Zoom in to see the details.**

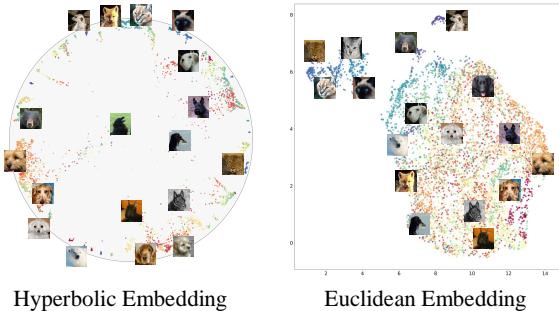


Figure 10. **UMAP visualization** of Hyperbolic and Euclidean 2-D embeddings of AnimalFaces dataset. Images with clear identities are clustered and positioned near the boundary, while ambiguous samples are located near the center in hyperbolic space.

fusion without the limitations of GAN-based generators, which typically support only finite fusion levels (*e.g.*, 18

levels in StyleGAN2’s \mathcal{W}^+ -space).

4.4. Ablation Study

Three ablation studies assess the impact of the stochastic encoder and hyperbolic embedding radius on few-shot image quality and diversity. As shown in Fig. 11, the stochastic encoder regulates the similarity between generated and reference images, where stronger encoding enhances similarity at the cost of diversity. Table 2 further demonstrates that the hyperbolic embedding radius $r_{\mathbb{D}}$ affects image quality and diversity—smaller radii increase diversity but may alter high-level attributes. The optimal trade-off is observed at $r_{\mathbb{D}} \approx 5.5$. Additionally, Tab. 3 reveals that identity shifts toward perturbed images when $r_{\mathbb{D}} < 4.5$, as measured by CLIP-S (CLIP similarity between generated images and reference images) and CLIP-P (CLIP similarity between gener-

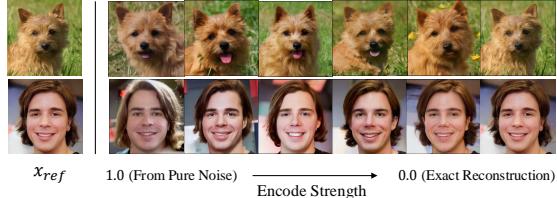


Figure 11. **Ablation study** on the influence of the encoding strength of the stochastic encoder on four datasets (strength equals 1 means x_0 is fully deconstructed, *i.e.*, x_T is a Gaussian noise).

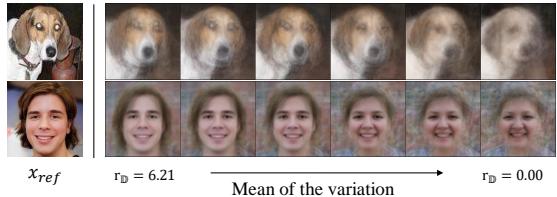


Figure 12. **The mean of 20 randomly sampled images** by moving the latent codes from the edge to the center of the Poincaré disk.



Figure 13. **Ablation study** of few shot image generation by **HypDAE(Hyp)** and **HypDAE(Euc)**.

ated images and perturbed images).

Euclidean versus Hyperbolic. To evaluate performance gains, we re-trained **HypDAE** (Euc) in Euclidean space using Eq. (5). As in Tab. 1 and Fig. 13, the hyperbolic space improves performance due to enhanced latent code disentanglement [17]. This is further validated by UMAP visualizations in Fig. 10. More details in Sec. E of the SM.

Pseudo versus Real Label. The accuracy of the pseudo labels predicted in Sec. 3.3 for the four datasets are: Animal Faces (48.9%), Flowers (78.6%), VGG Faces (41.5%), and NABirds (39.1%). We re-trained **HypDAE** (Pseudo) using the pseudo labels. Results in Tab. 1 indicate that **HypDAE** (Pseudo) outperforms **HypDAE** (Real) across most benchmarks, suggesting that noise in human-annotated labels may degrade hierarchical representation learning. This finding underscores that *manually labeled data is not essential for few-shot image generation*.

4.5. Few-Shot Image Generation

As described in Sec. 3, our approach enables few-shot image generation by varying the stochastic subcode z_T or shifting the latent code in a randomly chosen semantic direction within the category cluster. We set $r_D = 5.5$ and an encoding strength of 0.95 (*i.e.*, encoding 5% of the information) to achieve this, as illustrated in Fig. 4. We conduct three experiments demonstrating that **HypDAE** achieves promising

Hyp Radius	6.2	6.0	5.5	5.0	4.5	4.0	3.0
FID(\downarrow)	15.18	14.67	14.31	14.56	14.71	16.34	20.65
LPIPS(\uparrow)	0.7035	0.7278	0.7415	0.7643	0.7935	0.8378	0.8964

Table 2. **Ablation study** of different radii on Animal Faces.

Hyp Radius	6.2	6.0	5.5	5.0	4.5	4.0	3.0
CLIP-S	77.37	75.41	75.15	72.16	71.45	68.20	67.89
CLIP-P	69.62	69.89	72.35	72.86	74.25	76.34	77.00

Table 3. **Ablation study** of different radii on Animal Faces.

	WaveGAN	HAE	SAGE	HypDAE (Ours)
Quality (\uparrow)	1.34	2.67	2.54	3.45
Fidelity (\uparrow)	2.05	1.89	2.68	3.58
Diversity (\uparrow)	1.25	2.53	2.36	3.86

Table 4. **User study** on the comparison between our **HypDAE** and existing alternatives. “Quality”, “Fidelity”, and “Diversity” measure synthesis quality, object identity preservation, and object diversity. Each metric is rated from 1 (worst) to 4 (best).

few-shot generation, with additional examples in the SM.

Quantitative Comparison with State-of-the-Art. To assess fidelity and diversity, we compute FID [22] and LPIPS [64] following the one-shot settings in [37]. Results in Tab. 1 show significant improvements in FID and LPIPS scores on all four datasets except for the FID score on vggfaces due to the low resolution, affirming the advantages of our diffusion-based approach over previous GAN-based methods.

Qualitative Evaluation. We compare **HypDAE** qualitatively with WaveGAN [61] and HAE [37]. As shown in Fig. 9, **HypDAE** generates highly diverse images with fine-grained fidelity, such as detailed feather textures, which other methods fail to preserve. A user study further evaluates synthesis quality, identity preservation, and category-irrelevant diversity across these methods. As reported in Tab. 4, **HypDAE** consistently outperforms all baselines, with notable improvements in fidelity and diversity, attributed to its effective disentanglement of identity-relevant and irrelevant features. Additional details are provided in Sec. I of the SM.

5. Conclusion

In this work, we present **HypDAE**, the first diffusion-based method for editing hierarchical attributes in hyperbolic space. By learning the semantic hierarchy of images, our diffusion autoencoder enables continuous and flexible editing of hierarchical features. Experimental results show that **HypDAE** significantly outperforms existing GAN-based approaches without human-annotated labels, excelling in both random and customized few-shot generation. Furthermore, it facilitates hierarchical image generation with varying semantic similarity, advancing the state of the art in few-shot image generation. We hope this work inspires future research in this direction.

Acknowledgement. This work is supported by the Shun Hing Institute of Advanced Engineering (SHIAE) Fund (No. 8115074).

References

- [1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017. [1](#)
- [2] Sergey Bartunov and Dmitry P. Vetrov. Few-shot generative modelling with generative matching networks. In *AISTATS*, 2018. [1](#)
- [3] Silvère Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9): 2217–2229, 2013. [3](#)
- [4] Gary Béćigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *ICLR*, 2019. [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 *ICCV*, pages 22560–22570, 2023. [1, 2](#)
- [6] Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *NeurIPS*, page 4868–4879, 2019. [3, 4](#)
- [7] Wenhui Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *NeurIPS*, 2023. [3](#)
- [8] Louis Clouâtre and Marc Demers. Figr: Few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019. [1, 2](#)
- [9] Jiali Cui, Ying Nian Wu, and Tian Han. Learning joint latent space ebm prior model for multi-layer generator. In *CVPR*, pages 3603–3612, 2023. [2](#)
- [10] Giannis Daras and Alexandros G. Dimakis. Multiresolution textual inversion. In *NeurIPS Workshop*, 2022. [2](#)
- [11] Karan Desai, Maximilian Nickel, Tamay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. *PMLR*, 2023. [3](#)
- [12] Bhuvan Dhingra, Chris Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*, 2018. [2](#)
- [13] Guanqi Ding, Xinzhe Han, Shuhui Wang, Shuzhe Wu, Xin Jin, Dandan Tu, and Qingming Huang. Attribute group editing for reliable few-shot image generation. In *CVPR*, pages 11184–11193, 2022. [1, 2](#)
- [14] Guanqi Ding, Xinzhe Han, Shuhui Wang, Xin Jin, Dandan Tu, and Qingming Huang. Stable attribute group editing for reliable few-shot image generation. *arXiv preprint arXiv:2302.00179*, 2023. [1, 2, 6, 7, 12, 17](#)
- [15] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. [2](#)
- [16] Octavian-Eugen Ganea, Gary Béćigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, pages 5345–5355, 2018. [3, 5, 11](#)
- [17] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *CVPR*, 2023. [8](#)
- [18] Michael Gromov. Hyperbolic groups. In *Essays in group theory*, 1987. [3](#)
- [19] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. Lofgan: Fusing local representations for fewshot image generation. In *ICCV*, 2021. [1, 2](#)
- [20] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023. [2](#)
- [21] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *CVPR*, 2023. [3](#)
- [22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. [8](#)
- [23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NIPS Workshop*, 2022. [14](#)
- [24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. [1, 12](#)
- [25] Yan Hong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Deltagan: Towards diverse few-shot image generation with sample-specific delta. In *CVPR*, 2020. [1](#)
- [26] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Matchnggan: Matching-based few-shot image generation. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020. [1, 2](#)
- [27] Yan Hong, Li Niu, Jianfu Zhang, Weijie Zhao, Chen Fu, and Liqing Zhang. F2gan: Fusing-and-filling gan for few-shot image generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 2535–2543. Association for Computing Machinery, 2020. [1, 2, 7](#)
- [28] Yan Hong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Deltagan: Towards diverse few-shot image generation with sample-specific delta. In *ECCV*, 2022. [1, 2, 7](#)
- [29] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Few-shot image generation using discrete content representation. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 2796–2804, New York, NY, USA, 2022. Association for Computing Machinery. [1, 2](#)
- [30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4217–4228, 2019. [6, 11, 12](#)
- [31] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, pages 6417–6427, 2020. [2, 3, 13](#)
- [32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. [12](#)
- [33] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. [1, 2](#)
- [34] John M Lee. Riemannian manifolds: an introduction to curvature. *Springer Science & Business Media*, 176, 2006. [5](#)

- [35] John M Lee. *Introduction to Smooth Manifolds*. Springer, 2013. 5
- [36] Zhiying Leng, Tolga Birdal, Xiaohui Liang, and Federico Tombari. Hypersdfusion: Bridging hierarchical structures in language and geometry for enhanced 3d text2shape generation. In *CVPR*, pages 19691–19700, 2024. 3
- [37] Lingxiao Li, Yi Zhang, and Shuhui Wang. The euclidean space is evil: Hyperbolic attribute editing for few-shot image generation. In *ICCV*, pages 22714–22724, 2023. 1, 2, 3, 6, 7, 8, 11, 12, 17
- [38] Weixin Liang, Zixuan Liu, and Can Liu. Dawson: A domain adaptive few shot generation framework. *arXiv preprint arXiv:2001.00576*, 2020. 1, 2, 7
- [39] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *ICCV*, 2019. 6, 11, 12, 15
- [40] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018. 6
- [41] Maximillian Nickel and Douwe Kiela. Generative visual manipulation on the natural image manifold. In *ECCV*, 2017. 2, 3
- [42] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, 2018. 3
- [43] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 6, 11, 12, 15
- [44] Jiwoong Park, Junho Cho, Hyung Jin Chang, and Jin Young Choi. Unsupervised hyperbolic representation learning via message passing auto-encoders. In *CVPR*, pages 5512–5522, 2021. 3
- [45] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 6, 11, 12
- [46] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 15
- [47] Konpat Preechakul, Nattanat Chathee, Suttisak Wizadwongs, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 2, 4, 6
- [48] 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 *ICML*, 2021. 2, 4
- [49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022. 1
- [50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 6, 11
- [51] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 1, 2
- [52] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *CVPR*, 2024. 3
- [53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 4
- [54] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1
- [55] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. In *CVPR*, pages 12602–12612, 2021. 3
- [56] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv:2205.16007*, 2022. 14
- [57] Alexandru Tifrea, Gary Bécigneul, and OctavianEugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *ICLR*, 2019. 2, 3
- [58] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *CVPR*, 2015. 6, 11, 12, 15
- [59] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. Pt+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2
- [60] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023. 11
- [61] Mengping Yang, Zhe Wang, Ziqiu Chi, and Wenyi Feng. Wavegan: Frequency-aware gan for high-fidelity few-shot image generation. In *ECCV*, pages 1–17. Springer, 2022. 1, 2, 7, 8, 11, 17
- [62] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ippo-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2
- [63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, pages 3813–3824, 2023. 14
- [64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 8
- [65] Chenxi Zheng, Bangzhen Liu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Where is my spot? few-shot image generation via latent subspace optimization. In *CVPR*, pages 3272–3281, 2023. 2, 7
- [66] Yingbo Zhou, Yutong Ye, Pengyu Zhang, Xian Wei, and Mingsong Chen. Exact fusion via feature distribution matching for few-shot image generation. In *CVPR*, pages 8383–8392, 2024. 1, 7

Supplementary Material

Overview

This appendix is organized as follows:

Sec. 6 gives more implementation details of **HypDAE**.
Sec 3.2 & Sec 4.1

Sec. 7 gives detailed explanation of diffusion models.

Sec. 8 provides the mathematical formulae used in hyperbolic neural networks. Sec 3.2 & Sec 3.3

Sec. 9 shows more results of the ablation study of **HypDAE**. Sec 4.3

Sec. 10 shows more comparisons between the latent manipulation in hyperbolic and Euclidean space. Sec 4.3

Sec. 11 provides examples to show the exceptional out-of-distribution few-shot image generation ability. Sec 4.4

Sec. 12 shows the images generated with different radii in the Poincaré disk. Sec 4.3

Sec. 13 compares the images generated by state-of-the-art few-shot image generation method, *i.e.* WaveGAN [61], HAE [37] and our methods **HypDAE**. Sec 4.4

Sec. 14 gives more details of the user study we conducted. Sec 4.4

Sec. 15 gives more examples generated by **HypDAE**. Sec 4.4

6. Implementation Details and Analysis

Stage I. As mentioned in Sec 3.2, this stage does not require class labels for the images. To promote diversity, we use only the CLIP image encoder’s class token (dimension 1×1024) for a compact representation, aligning it with the CLIP text feature space via a 5-layer fully connected MLP following the same settings in [60] that inject features into the diffusion process through cross-attention to replace the text feature in the original stable diffusion model.

We choose Stable Diffusion V2.1 [50] as the base generator for the base generative model. We set the image resolution to 512×512 . We choose the Adam optimizer and set the learning rate as $1e-5$. During the training process, the pre-trained CLIP image encoder and SD V2.1 models are frozen, only the Transformer block for aligning features is trainable. Since the SD model is loaded during the training process, we use $2 \times$ NVIDIA A800 (80GB) GPUs for training, and the batch size is selected as 24 for each GPU. We train about $1e5$ steps to get the model to converge on each dataset.

Stage II. This stage is the only stage that requires the class labels for given images to learn the hierarchical representation. Although class labels are required in Stage II, the model only needs a small number of labeled data for pre-training and pseudo labels can be predicted by CLIP as

shown in Sec. 6. Furthermore, we show exceptional out-of-distribution generation ability in Sec. 11. For the hyperbolic encoder mentioned in Sec 3.2, we use a single-head 5-layer Transformer block to reduce the dimensionality of the Euclidean latent vector c from 1×1024 to 1×512 , which is then mapped to hyperbolic space via an exponential map. A hyperbolic feed-forward layer [16] produces the final hierarchical representation $z_{\mathbb{D}}$:

$$c_h = f^{\otimes_c}(\exp_0^c(E(c))), \quad (8)$$

where E is the Transformer encoder and f^{\otimes_c} is the Möbius translation of feed-forward layer f as the map from Euclidean space to hyperbolic space, denoted as *Möbius linear layer*. In order to perform multi-class classification on the Poincaré disk defined in Sec 3.1, one needs to generalize multinomial logistic regression (MLR) to the Poincaré disk defined in [16]. An extra linear layer needs to be trained for the classification, and the details on how to compute softmax probability in hyperbolic space are shown in Sec. 8. As mentioned in Sec 3.1, the distance between points grows exponentially with their radius in the Poincaré disk. In order to minimize Eq. (5) in the main paper, the latent codes of fine-grained images will be pushed to the edge of the ball to maximize the distances between different categories while the embedding of abstract images (images have common features from many categories) will be located near the center of the ball. Since hyperbolic space is continuous and differentiable, we are able to optimize Eq. (5) with stochastic gradient descent, which learns the hierarchy of the images.

Then we train a Transformer decoder to project the hyperbolic latent code back to the CLIP image space with exact reconstruction. In practice, this is achieved by firstly applying a logarithmic map followed by a Transformer decoder D :

$$c' = D(\log_0^c(c_h)). \quad (9)$$

and c' will be fed into the cross-attention layer of the stable diffusion model to reconstruct the image x' . We use a single-head 30-layer Transformer block as the Transformer decoder for Animal Faces [39], VGGFaces [45], FFHQ [30], and NABirds [58] since these datasets are relatively large. Therefore, a deeper network is needed to reconstruct the latent representation of these large datasets. However, for the Flowers dataset [43], the number of images is less than 10 thousand, which is not enough to train a deep neural network. As a consequence, we use a single-head 5-layer Transformer block as the Transformer decoder for Flowers which works well.

Fine-tuning on FFHQ. As we mentioned in Sec 4.2 in the main paper, we learn the hierarchy of human faces by training Stage II with VGGFaces first. However, we visualize the human faces with the FFHQ dataset. Note that the FFHQ dataset has no class labels. Therefore, we first use

the VGGFaces dataset to learn a good prior of hierarchy among human faces images with supervision, then fine-tune the model with the reconstruction loss \mathcal{L}_{rec} *only* to teach the model how to reconstruct images with high resolution but maintaining the hierarchical representation prior. The results show the great potential of our model to be fine-tuned on large-scale dataset without supervision.

In Stage II, only the CLIP image encoder is loaded during the training process. Besides, the CLIP image encoder is frozen, and only the lightweight Transformer encoder and decoder are trainable. We use $1 \times$ NVIDIA RTX 4090 (24GB) GPU for training, and the batch size is selected as 256. The λ in Eq. (7) in the main paper is selected as 0.1. We choose the AdamW [32] optimizer and set the learning rate as $1e-3$. A linear learning rate scheduler is used with a step size equal to 5000, with a multiplier $\gamma = 0.5$. We train about 1e5 steps to get the model to converge on each dataset.

In addition, as a remark, we choose the largest radius as 6 in most of our experiments as in hyperbolic space since any vector asymptotically lying on the surface unit N -sphere will have a hyperbolic length of approximately $r = 6.2126$, which can be directly calculated by Eq. (2).

Although training our model requires considerable computing resources as mentioned before, the runtime cost and resources required for the inference stage are affordable. Our model can inference on a single NVIDIA RTX 4090 GPU (24GB) thanks to our multi-stage training/inference since one does not need to load all models simultaneously.

Pseudo-Labeling. For Flowers, Animal Faces and NABirds, we utilize the CLIP ViT-B/32 model. Given an image (x), we extract its embedding using the CLIP image encoder. Similarly, we compute embeddings for a predefined set of class names (y_i) using the CLIP text encoder. The cosine similarity between the image embedding and each class embedding is computed as:

$$\text{sim}(x, y_i) = \frac{f(x) \cdot g(y_i)}{|f(x)| |g(y_i)|}, \quad (10)$$

where ($f(\cdot)$) and ($g(\cdot)$) denote the CLIP image and text encoders, respectively. The softmax function is applied to convert similarity scores into probabilities. The pseudo-label (y) is assigned as the class with the highest probability:

$$y = \operatorname{argmax}_j \frac{\exp(\text{sim}(x, y_j))}{\sum_j \exp(\text{sim}(x, y_j))}. \quad (11)$$

For the VGGFaces dataset, we employ the DeepFace framework with the VGG-Face architecture. DeepFace predicts pseudo-labels by comparing the face embedding of an image with embeddings of known identities in the dataset. Specifically, we construct a reference database by selecting one image per class, and each input image is assigned to the class with the highest similarity score. This approach en-

sures robust pseudo-labeling by leveraging DeepFace’s face recognition capabilities.

Dataset Settings We evaluate our method on Animal Faces [39], Flowers [43], VGGFaces [45], FFHQ [30], and NABirds [58] following the settings described in [14, 37].

Animal Faces. We randomly select 119 categories as seen for training and leave 30 as unseen categories for evaluation.

Flowers. The Flowers [43] dataset is split into 85 seen categories for training and 17 unseen categories for evaluation.

VGGFaces. For VGGFaces [45], we randomly select 1802 categories for training and 572 for evaluation.

NABirds. For NABirds [58], 444 categories are selected for training and 111 for evaluation.

FFHQ. Due to the low resolution (64×64) of the VGGFces dataset, we use FFHQ [30] to fine-tune the model pre-trained on VGGFaces without supervision and visualize images of human faces with FFHQ.

7. Additional Background - Diffusion Models

Diffusion Denoising Probabilistic Models (DDPM) [24] are generative latent variable models that aim to model a distribution $p_\theta(x_0)$ that approximates the data distribution $q(x_0)$ and easy to sample from. DDPMs model a “forward process” in the space of x_0 from data to noise. This is called “forward” due to its procedure progressing from x_0 to x_T . Note that this process is a Markov chain starting from x_0 , where we gradually add noise to the data to generate the latent variables $x_1, \dots, x_T \in X$. The sequence of latent variables, therefore, follows $q(x_1, \dots, x_t | x_0) = \prod_{i=1}^t q(x_t | x_{t-1})$, where a step in the forward process is defined as a Gaussian transition $q(x_t | x_{t-1}) := N(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)$ parameterized by a schedule $\beta_0, \dots, \beta_T \in (0, 1)$. When T is large enough, the last noise vector x_T nearly follows an isotropic Gaussian distribution.

An interesting property of the forward process is that one can express the latent variable x_t directly as the following linear combination of noise and x_0 without sampling intermediate latent vectors:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} w, \quad w \sim N(0, I), \quad (12)$$

where $\alpha_t := \prod_{i=1}^t (1 - \beta_i)$.

To sample from the distribution $q(x_0)$, we define the dual “reverse process” $p(x_{t-1} | x_t)$ from isotropic Gaussian noise x_T to data by sampling the posteriors $q(x_{t-1} | x_t)$. Since the intractable reverse process $q(x_{t-1} | x_t)$ depends on the unknown data distribution $q(x_0)$, we approximate it with a parameterized Gaussian transition network $p_\theta(x_{t-1} | x_t) := N(x_{t-1} | \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$. The $\mu_\theta(x_t, t)$ can be replaced [24] by predicting the noise $\epsilon_\theta(x_t, t)$ added to x_0 using equation 12.

8. Hyperbolic Neural Networks

For hyperbolic spaces, since the metric is different from Euclidean space, the corresponding calculation operators also differ from Euclidean space. In this section, we start by defining two basic operations: Möbius addition and Möbius scalar multiplication [31], given fixed curvature c .

For any given vectors $x, y \in \mathbb{H}^n$, the *Möbius addition* is defined by:

$$x \oplus_c y = \frac{(1 - 2c\langle x, y \rangle - c\|y\|_2^2)x + (1 + c\|x\|_2^2)y}{1 - 2c\langle x, y \rangle + c^2\|x\|_2^2\|y\|_2^2}, \quad (13)$$

where $\|\cdot\|$ denotes the 2-norm of the vector, and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product of the vectors.

Similarly, the *Möbius scalar multiplication* of a scalar r and a given vector $x \in \mathbb{H}^n$ is defined by:

$$r \otimes_c x = \tan_c(r \tan_c^{-1}(\|x\|_2)) \frac{x}{\|x\|_2}. \quad (14)$$

We also would like to give explicit forms of the exponential map and the logarithmic map which are used in our model to achieve the translation between hyperbolic space and Euclidean space as mentioned in Sec 3.2.

The *exponential map* $\exp_x^c : T_x \mathbb{D}_c^n \cong \mathbb{R}^n \rightarrow \mathbb{D}_c^n$, that maps from the tangent spaces into the manifold, is given by

$$\exp_x^c(v) := x \oplus_c \left(\tanh\left(\sqrt{c} \frac{\lambda_x^c \|v\|}{2}\right) \frac{v}{\sqrt{c} \|v\|} \right). \quad (15)$$

The *logarithmic map* $\log_x^c(y) : \mathbb{D}_c^n \rightarrow T_x \mathbb{D}_c^n \cong \mathbb{R}^n$ is given by

$$\log_x^c(y) := \frac{2}{\sqrt{c} \lambda_x^c} \operatorname{arctanh}\left(\sqrt{c} \| -x \oplus_c y \| \right) \frac{-x \oplus_c y}{\| -x \oplus_c y \|}. \quad (16)$$

We also provide the formula to calculate the softmax probability in hyperbolic space used in Eq. (5) in the main paper: Given K classes and $k \in \{1, \dots, K\}$, $p_k \in \mathbb{D}_c^n$, $a_k \in T_{p_k} \mathbb{D}_c^n \setminus \{\mathbf{0}\}$:

$$p(y = k | x) \propto \exp\left(\frac{\lambda_{p_k}^c \|a_k\|}{\sqrt{c}} \sinh^{-1}\left(\frac{2\sqrt{c} \langle -p_k \oplus_c x, a_k \rangle}{(1 - c \| -p_k \oplus_c x \|^2) \|a_k\|}\right)\right), \quad \forall x \in \mathbb{D}_c^n, \quad (17)$$

where \oplus_c denotes the Möbius addition defined in Eq. (13) with fixed sectional curvature of the space, denoted by c .

Hierarchical Data Sampling. As illustrated in Fig. 14, as the latent code z_{ref} of the reference image x_{ref} moves from the edge to the center of the Poincaré disk, the control of the

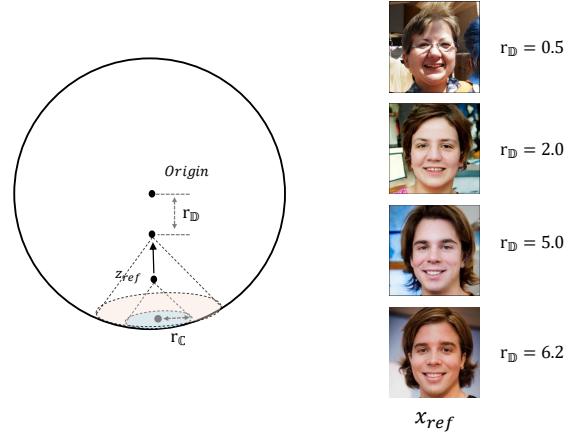


Figure 14. **The illustration of hierarchical data sampling in hyperbolic space**

identity of the sampled images becomes weaker and weaker. The identity of the generated images becomes more ambiguous. To conduct hierarchical image sampling in hyperbolic space, one can move the latent code z_{ref} of the reference image x_{ref} towards the origin of the Poincaré disk. Then, sampling latent points among the “children” of the rescaled reference images. In practice, the semantic diversity of the generated images can be controlled either by setting different values of r_D , then calculating r_C based on r_D , or by setting different values of r_C directly.

Recall that, in hyperbolic space, the shortest path with the induced distance between two points is given by the geodesic defined in Eq. (2) in the main paper. The geodesic equation between two embeddings z_{Di} and z_{Dj} , denoted by $\gamma_{z_{Di} \rightarrow z_{Dj}}(t)$, is given by

$$\gamma_{z_{Di} \rightarrow z_{Dj}}(t) = z_{Di} \oplus_c t \otimes_c ((-z_{Di}) \oplus_c z_{Dj}), \quad t \in [0, 1], \quad (18)$$

where \oplus_c denotes the Möbius addition with aforementioned sectional curvature c . Therefore, for hierarchical data sampling, we can first define the value of r_C , then sample random data points in hyperbolic space. If the distance between the sampled data point and z_{ref} is equal to or shorter than the value, we accept the data point. Otherwise, we move the latent codes along the geodesic between the sampled data point and z_{ref} until the distance is within the scope we define. We show more hierarchical image sampling examples in Sec. 11.

9. Ablation Study

There are a few hyperparameters of **HypDAE** that control the generation quality and diversity. We conduct ablation studies on each of them in this section.

Hyperbolic Radius. By varying the radii of “parent” images, **HypDAE** controls the semantic diversity of gener-

ated images (Fig. 24), where the “parent” images can be viewed as the image with the average attributes of its children. The quantitative results of the Flowers dataset are presented in Tab. 5. We can see that the diversity increases as the radius becomes smaller, therefore, the value of LPIPS increases accordingly. However, changing too many attributes changes the identity or category of the given images, therefore, the FID decreases when the radius is smaller than 5.5. In practice, we select 5.5 as the radius of the parent images for few-shot image generation.

Classifier-free Guidance. To achieve the trade-off between identity preservation and image harmonization, we find that classifier-free sampling strategy [23] is a powerful tool. Previous work [56] found that the classifier-free guidance is actually the combination of both prior and posterior constraints. In our experiments, we follow the settings in [63].

$$\epsilon_{\text{prd}} = \epsilon_{\text{uc}} + s (\epsilon_c - \epsilon_{\text{uc}}), \quad (19)$$

where ϵ_{prd} , ϵ_{uc} , ϵ_c , s are the model’s final output, unconditional output, conditional output, and a user-specified weight, respectively. The visualizations are shown in Fig. 16, and quantitative results for the Flowers dataset are presented in Tab. 6. Consistent with findings in Tab. 5, diversity, measured by LPIPS, increases as the cfg scale grows. However, excessive cfg scaling can alter the identity or category of the input images, leading to a decline in FID when the cfg scale exceeds 1.3. Based on these results, we select a cfg scale of 1.3 to achieve optimal few-shot image generation with a balance between fidelity and diversity.

Encoding Strength of the Stochastic Encoder. As described in Sec. 3.2, the encoding strength of the stochastic encoder determines the extent of information encoded from the given images. For instance, attributes such as rough posture, color, and style are encoded during the early steps of the diffusion process. A higher encoding strength deconstructs more information from the input images, while a lower encoding strength retains more original information. An encoding strength of 1 implies full deconstruction, where the initial latent of the denoising process is Gaussian noise. Conversely, an encoding strength of 0 results in exact reconstruction without information loss.

While lower encoding strength preserves the identity and style of the input images, it reduces diversity. This trade-off is visualized in Fig. 17, and quantitative results for the Flowers dataset are presented in Tab. 7. Consistent with Tab. 5, diversity, measured by LPIPS, increases with higher encoding strength, while excessive encoding strength can cause changes in identity or category. This is reflected in a decrease in FID when encoding strength exceeds 0.95 (*i.e.*, 5% of the information is encoded). Based on these findings, we set the encoding strength of the stochastic encoder to 0.95 to achieve reliable few-shot image generation.

Hyperparameter Ablation. To validate the robustness of

Hyp Radius	6.2	6.0	5.5	5.0	4.5
FID(\downarrow)	27.89	24.67	23.96	24.89	26.63
LPIPS(\uparrow)	0.7585	0.7589	0.7595	0.7643	0.7725

Table 5. **Ablation study** of different radii on Flowers.

CFG	1.0	1.1	1.3	1.5	1.7
FID(\downarrow)	25.52	24.89	23.96	26.04	25.63
LPIPS(\uparrow)	0.7391	0.7534	0.7595	0.7660	0.7737

Table 6. **Ablation study** of the influence of CFG on Flowers.

Strength	1.0	0.98	0.95	0.9	0.8
FID(\downarrow)	28.97	24.59	23.96	24.94	26.48
LPIPS(\uparrow)	0.7631	0.7606	0.7595	0.7585	0.7375

Table 7. **Ablation study** of the influence of encoding strength on Flowers.

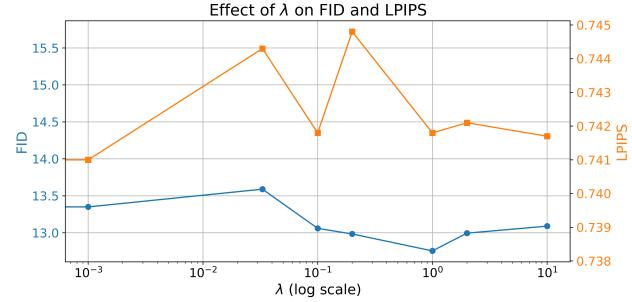


Figure 15. **Ablation Study** of trade-off adaptive hyperparameter λ in the loss function.

the trade-off parameter in Eq. (7), we rewrite Eq. (7) as: $\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{hyper}} + \mathcal{L}_{\text{rec}}$, ablate the loss to analyze this trade-off. As shown in Fig. 15, increasing λ improves semantic consistency (lower FID) while reducing diversity (lower LPIPS), validating the controllability introduced by the hyperbolic component.

10. Comparison with Euclidean space

In this section, we present a detailed comparison of different latent spaces, as shown in Fig. 18. Compared to classical Euclidean space, hyperbolic space enables smoother transitions between two given images. In hyperbolic space, identity-irrelevant features transition first, followed by a gradual change in identity-relevant features. In contrast, Euclidean space exhibits simultaneous changes in both identity-relevant and identity-irrelevant features, leading to less structured transitions.

These results confirm that our method effectively learns hierarchical representations in hyperbolic space, enabling few-shot image generation by selectively modify-

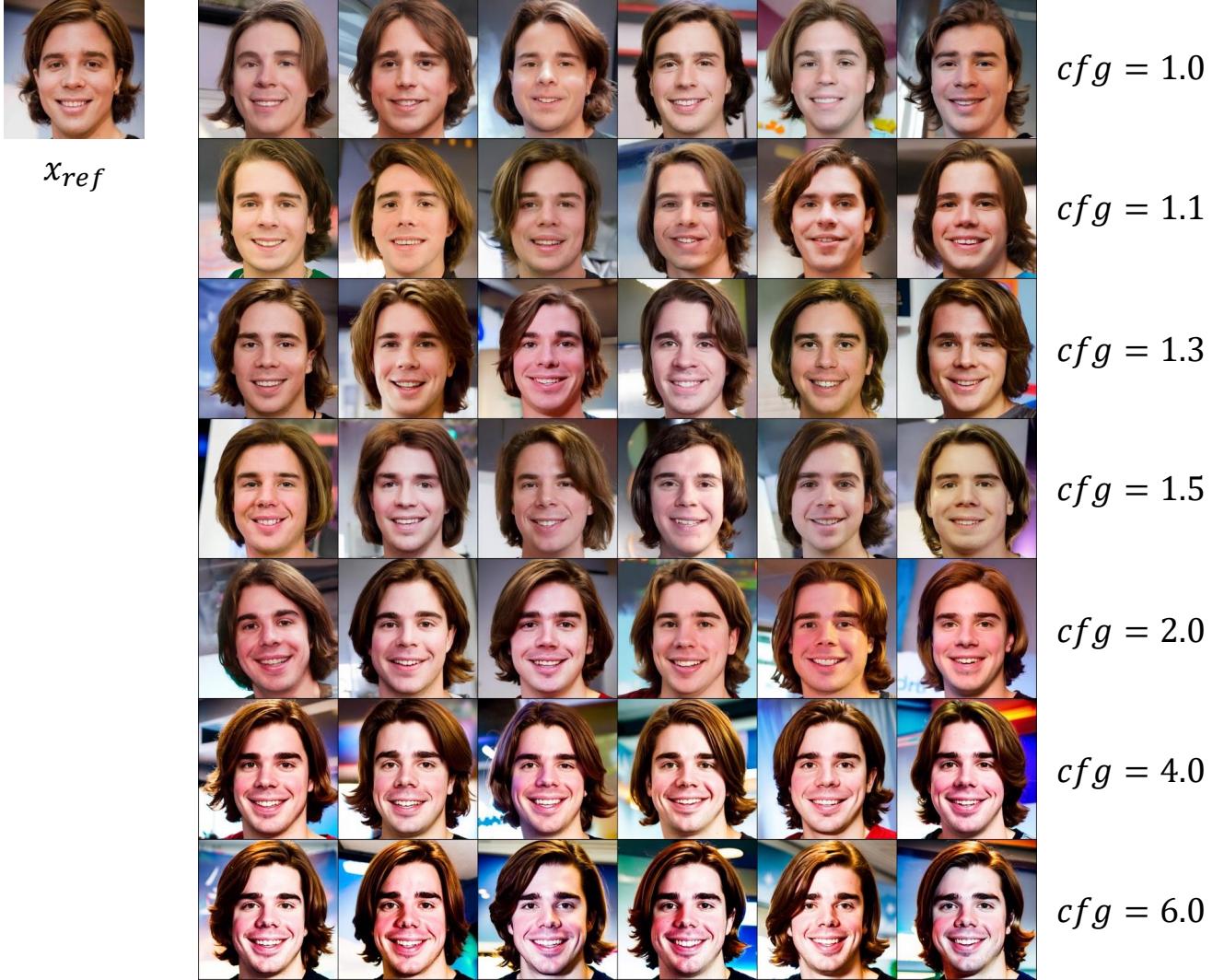


Figure 16. **Ablation study** on the influence of classifier free guidance.

ing category-irrelevant features—a capability that Euclidean space cannot achieve. Additional interpolation results, provided in Fig. 19, demonstrate that smooth and distortion-free transitions are achievable in hyperbolic space. These findings highlight that **HypDAE** enables precise geodesic and hierarchical control during editing, offering a significant advantage over traditional approaches.

11. Out-of-distribution Few-shot Image Generation

In Sec 4.2 of the main paper, we mentioned we fine-tuned the model trained with VGGFaces using the FFHQ dataset. The model shows exceptional out-of-distribution generalization ability on the FFHQ dataset. To further verify the

OOD generalization ability of **HypDAE**, we select two images for Animal Faces [39], Flowers [43], and NABirds [58] datasets with three styles from DomainNet [46] including “painting”, “sketch”, “quick draw”, and “clipart” styles where are model never seen during the training stage. The OOD style transfer can be done by slightly increasing the encoding strength of the stochastic encoder to capture more style information of the given new images. The results in Fig. 20 Fig. 21 Fig. 22 Fig. 23 show that our proposed method has exceptional OOD generalization ability even for new domains with a big gap from the original domain. Although our model still generates images with some real detail for the style “clipart”, the performance in other styles is satisfying. Such an OOD generalization ability is significantly better than any of the previous works.



Figure 17. **Ablation study** on the influence of the encoding strength of the stochastic encoder on FFHQ (strength equals 1 means x_0 is fully deconstructed, *i.e.*, x_T is a Gaussian noise).

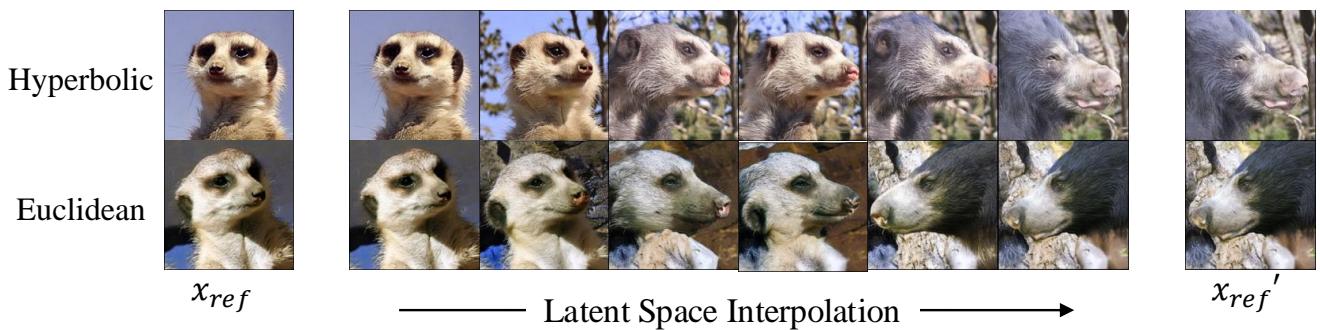


Figure 18. **Comparison of interpolation in hyperbolic space and Euclidean space** on Animal Faces dataset.

12. Hierarchical Image Generation

In this section, we provide additional examples of images generated by **HypDAE** at varying radii in the Poincaré disk.

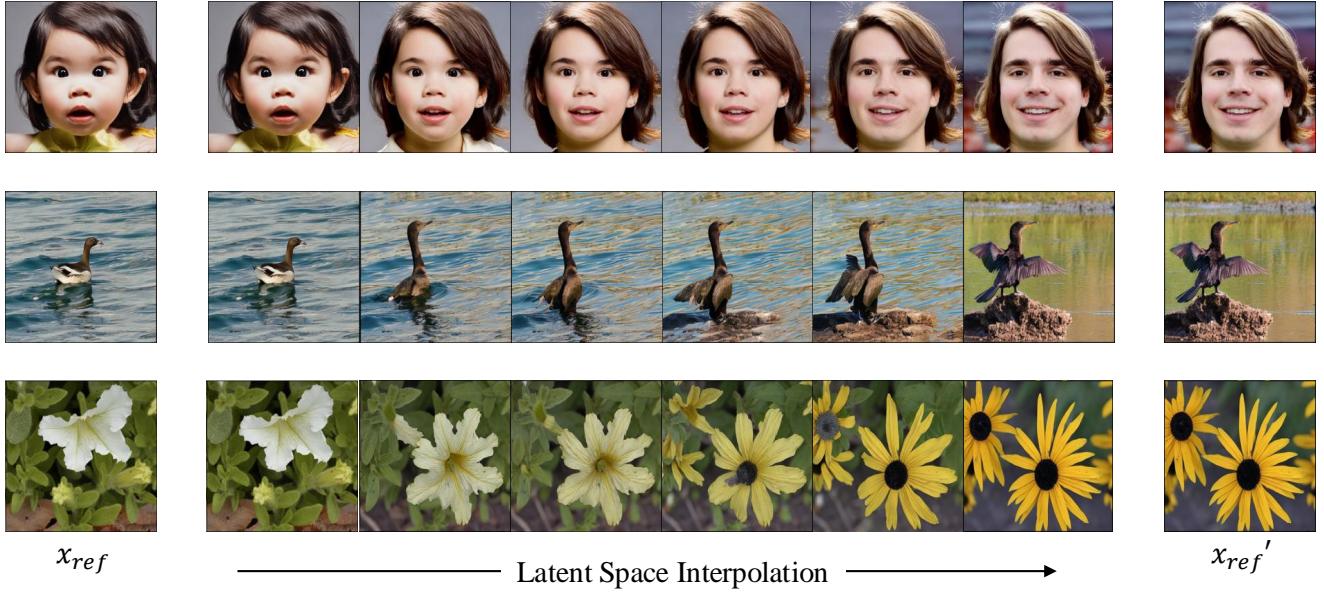


Figure 19. **More results of interpolation in hyperbolic space** on FFHQ, NABirds, and Flowers datasets.

As illustrated in Fig. 24, Fig. 25, and Fig. 26, high-level, category-relevant attributes remain unchanged when the radius is large, allowing for the generation of diverse images within the same category. Conversely, as the hyperbolic radius $r_{\mathbb{D}}$ decreases, the generated images become more abstract and semantically diverse. Moving closer to the center of the Poincaré disk results in the gradual loss of fine-grained details and changes to higher-level attributes.

For the few-shot image generation task, larger radii are optimal as they allow for the modification of category-irrelevant attributes while preserving the category identity. However, **HypDAE** is not limited to few-shot image generation and shows significant potential for other downstream applications. For example, **HypDAE** can generate a diverse set of feline images from a single cat image. This is achieved by scaling the latent code to a smaller radius in hyperbolic space and introducing random perturbations to approximate the average latent code for various feline categories. Finally, fine-grained and diverse feline images are generated by moving these average codes outward to larger radii, representing their "children" in the hierarchical space.

13. Comparison with State-of-the-art Few-shot Image Generation Method

We compare images generated by state-of-the-art methods, including WaveGAN [61], HAE [37], and our proposed method, across four datasets. As shown in Fig. 27, WaveGAN produces high-fidelity images, but the diversity is limited (*i.e.*, blending features from two input images with-

out significant variation). HAE improves diversity but suffers from low fidelity and quality, with missing details and changes in category or identity compared to the original images. In contrast, **HypDAE** achieves an excellent balance between maintaining identity and enhancing diversity while delivering significantly higher image quality than other methods. These results highlight the potential of **HypDAE** for broader applications in future downstream tasks.

14. User Study

As mentioned, we conducted an extensive user study with a fully randomized survey. Results are shown in the main text. Specifically, we compared **HypDAE** with three other models WaveGAN [61], HAE [37], and SAGE [14]:

1. We randomly chose 5 images from four datasets, and for each image, we then generated 3 variants in 1-shot setting (WaveGAN used 2-shot setting), respectively. Overall, there were 20 original images and 60 generated variants in total.
2. For each sample of each model, we present one masked background image, a reference object, and the generated image to annotators. We then shuffled the orders for all images.
3. We recruited 30 volunteers from diverse backgrounds and provided detailed guidelines and templates for evaluation. Annotators rated the images on a scale of 1 to 4 across three criteria: "Fidelity", "Quality", and "Diversity". "Fidelity" evaluates identity preservation, while "Quality" assesses quality of the images (*e.g.*, details of the im-

age). “Diversity” measures variation among generated proposals to discourage “copy-paste” style outputs.

The user-study interface is shown in Fig. 30.

15. Additional Examples Generated by HypDAE

Finally, we provide more examples generated by **HypDAE** in Fig. 28 and Fig. 29 for four datasets. The results show that our method achieves a balance between the quality and diversity of the generated images which significantly outperforms previous methods.

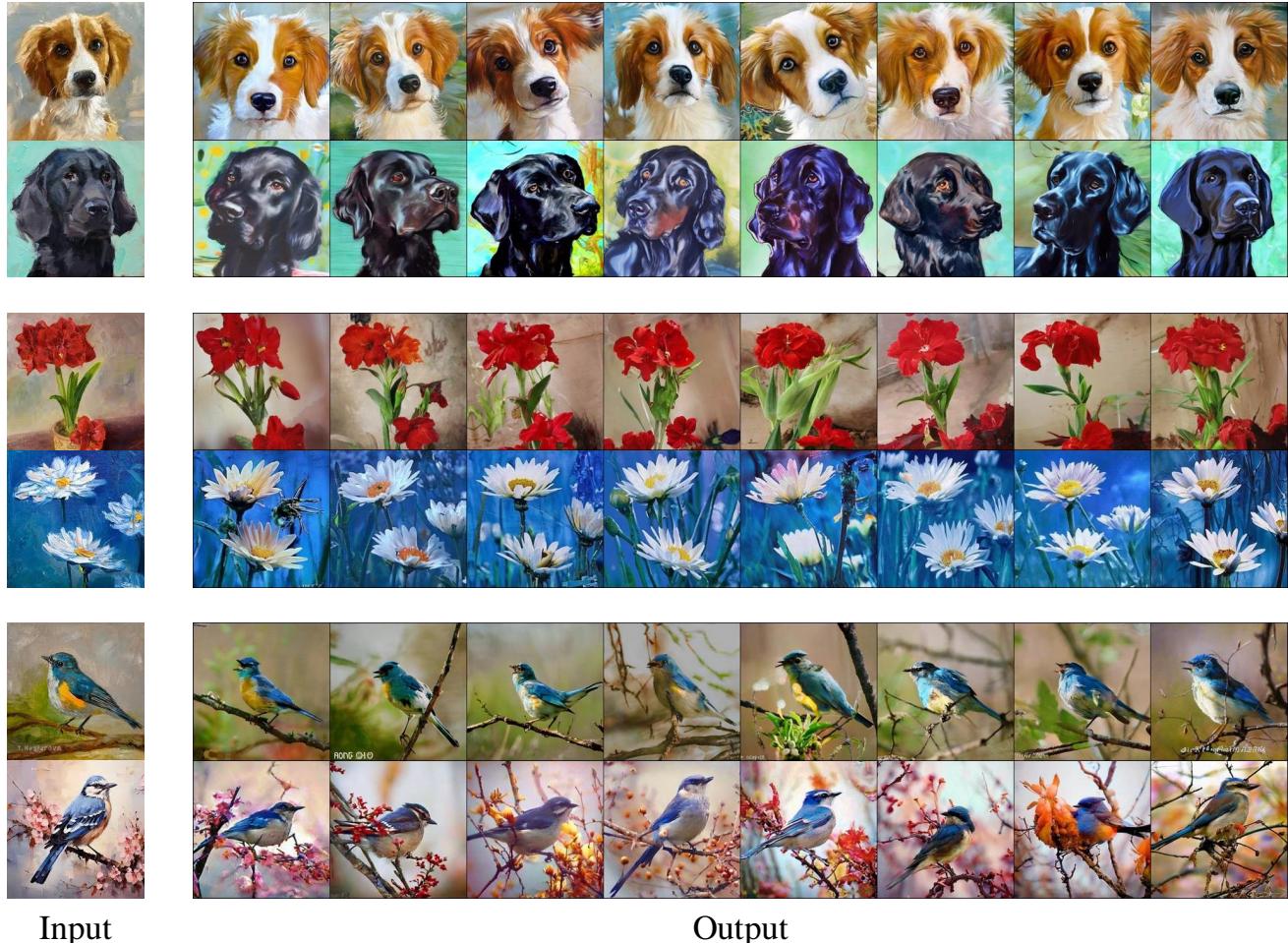


Figure 20. Few-shot image generation on **out-of-distribution examples in painting style**.

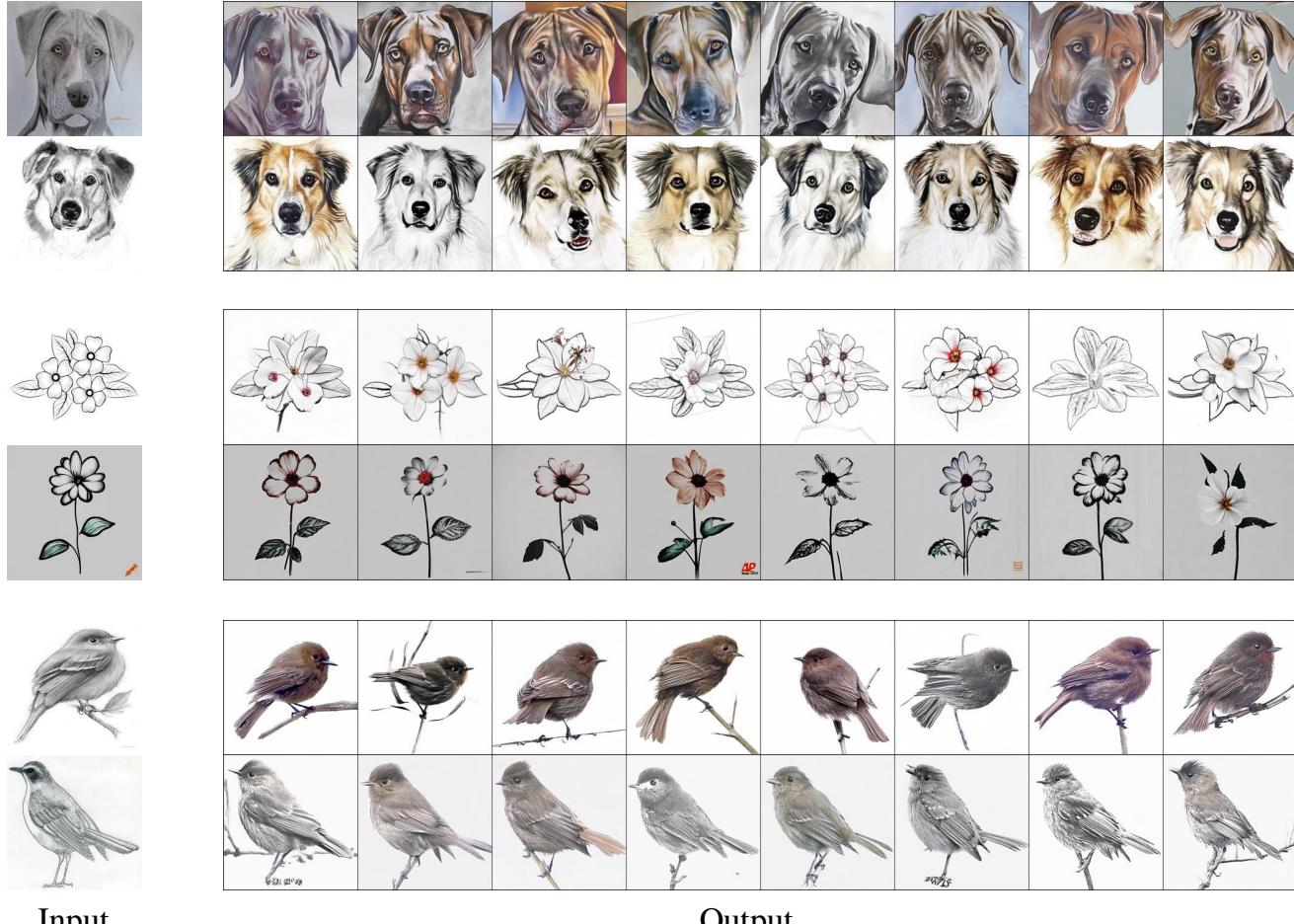


Figure 21. Few-shot image generation on **out-of-distribution examples in sketch style**.

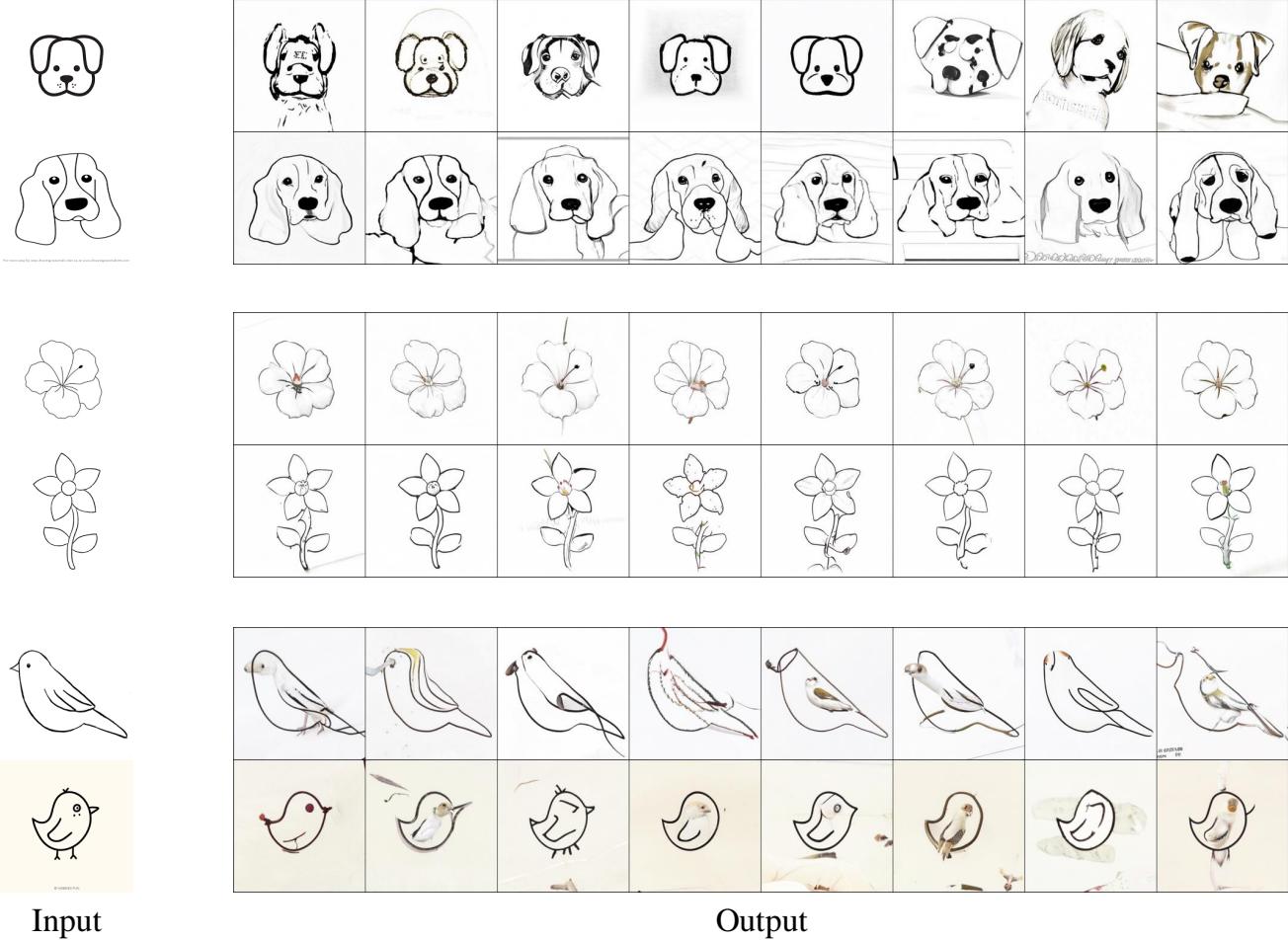
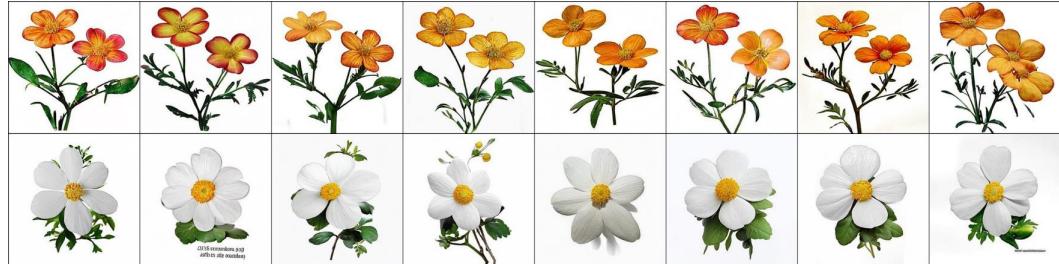


Figure 22. Few-shot image generation on **out-of-distribution examples in quick draw style**.



Input

Output

Figure 23. Few-shot image generation on **out-of-distribution examples in clipart style**.

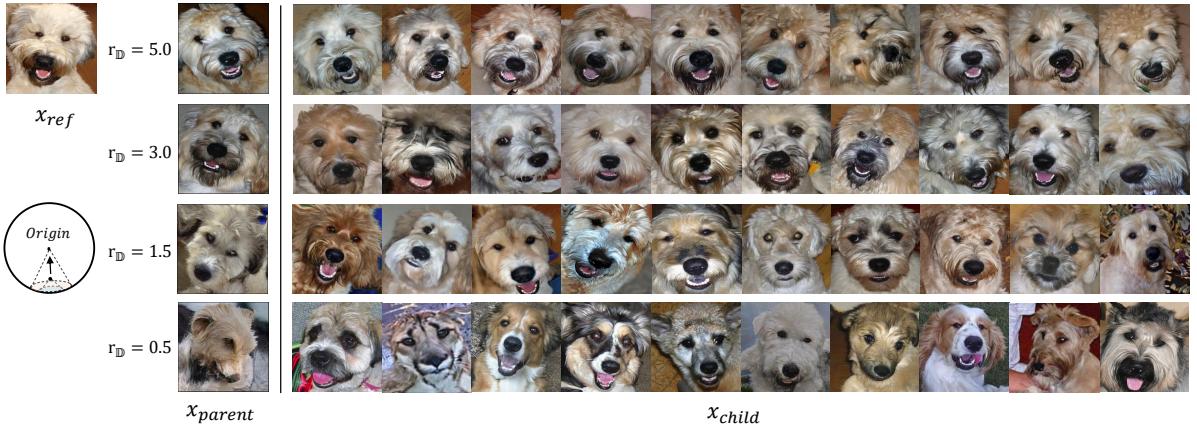


Figure 24. **Images with hierarchical semantic similarity generated by HypDAE.**



Figure 25. **Images with hierarchical semantic similarity generated by HypDAE.**

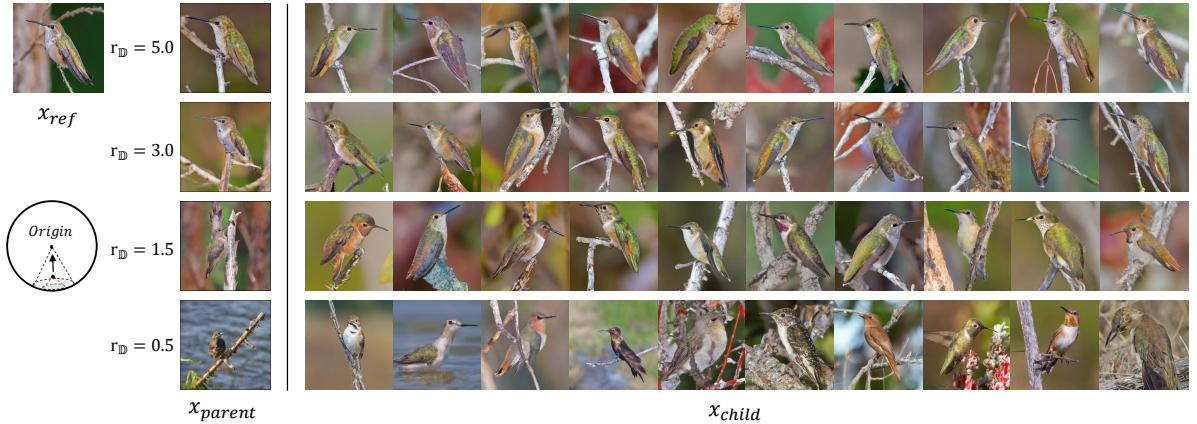


Figure 26. **Images with hierarchical semantic similarity generated by HypDAE.**



Input

WaveGAN

HAE

HypDAE

Figure 27. **More comparison between images generated by WaveGAN, HAE, and HypDAE on Flowers, Animal Faces, VGGFaces, and NABirds.** Note: WaveGAN uses a 2-shot setting; HAE and HypDAE are both in a 1-shot setting. **Zoom in to see the details.**

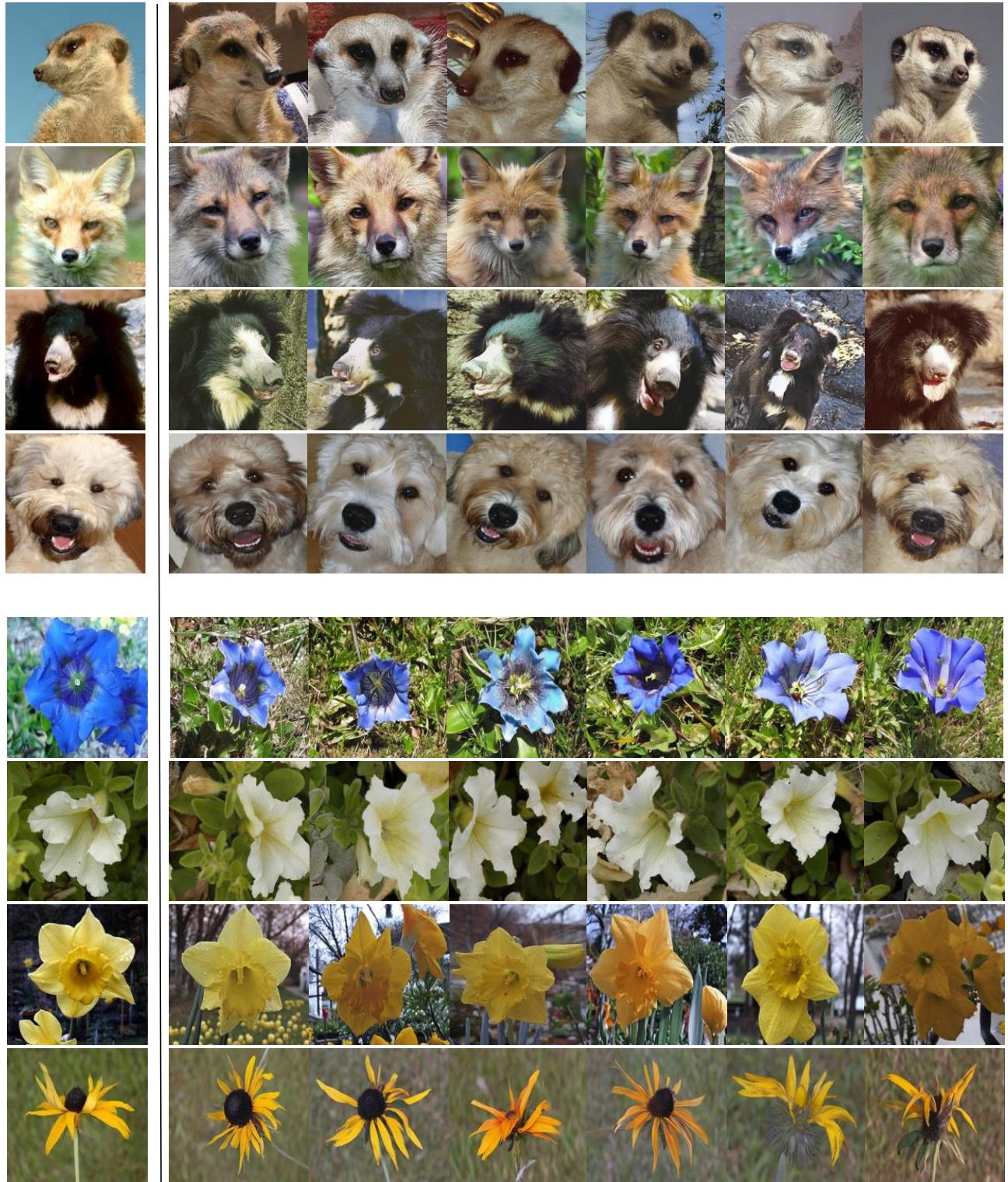
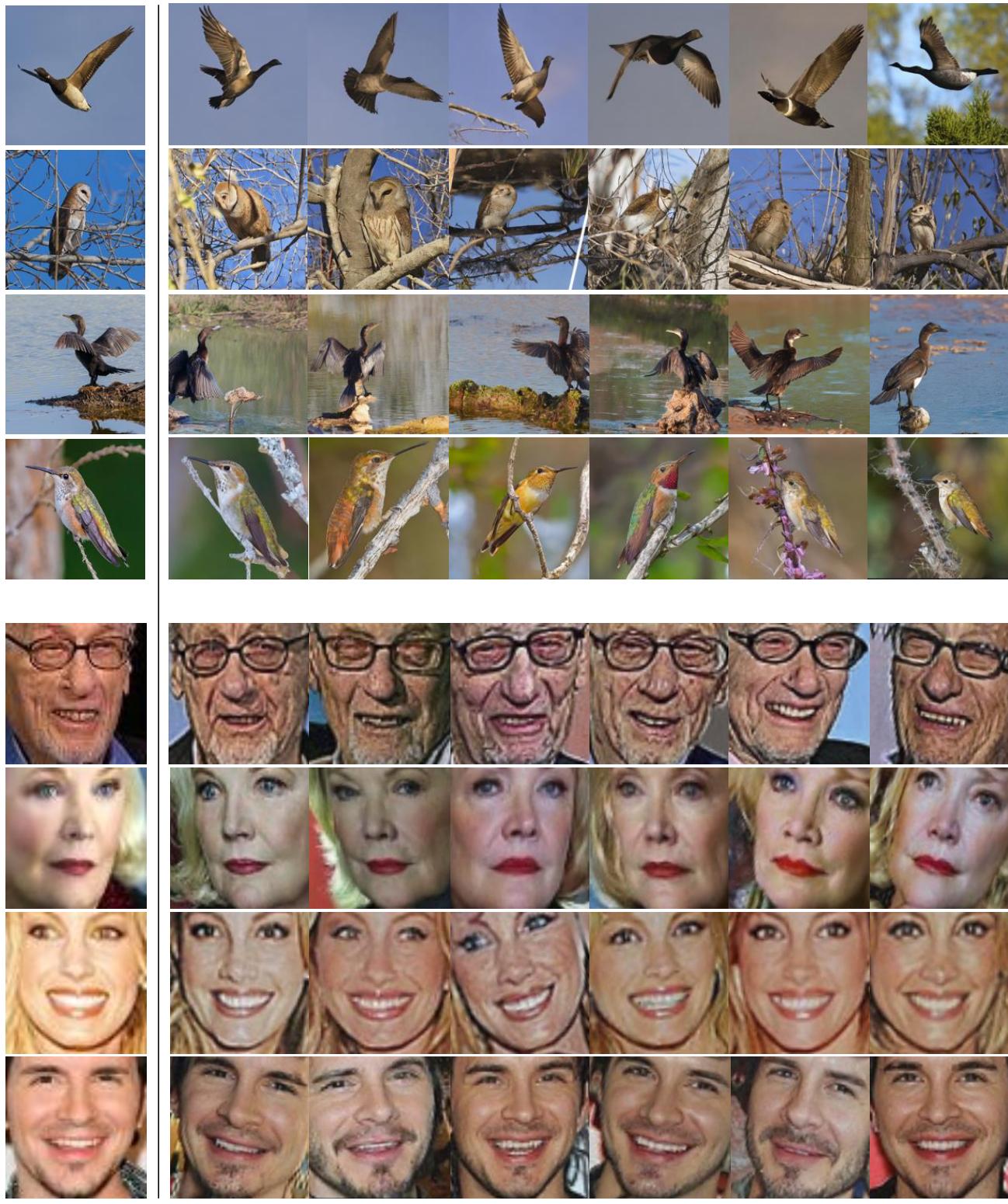


Figure 28. **More examples generated by HypDAE on Animal Faces and Flowers.**



Input

Output

Figure 29. **More examples generated by HypDAE on NABirds and VGGFaces.**

You are given a reference image to generate diverse new images that belong to the same category/identity as the reference image.

Your task is to rate the generated image from 1 (worst) to 4 (best) concerning

- 1) **Fidelity:** If the generated image preserves its original input category/identity
- 2) **Quality:** If the quality of the generated images is good (with details and looks like real images)
- 3) **Diversity:** If the generated images have novel views or poses

Problem 1: Input the score for Fidelity
Problem 2: Input the score for Quality
Problem 3: Input the score for Diversity

0	✓
0	✓
0	✓

INPUT



OUTPUT



Figure 30. The illustration of the user study interface.