UNIHDA: A UNIFIED AND VERSATILE FRAMEWORK FOR GENERALIZED HYBRID DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

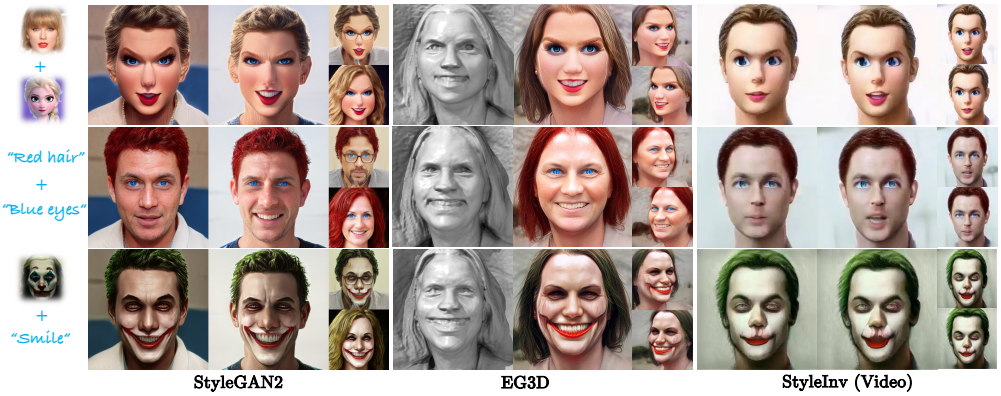

Figure 1: Given a pre-trained generator and multiple domains, UniHDA adapts the generator to a hybrid target domain that blends all characteristics at once and maintains robust cross-domain consistency. It supports both image and text modalities and is versatile to multiple generators.

ABSTRACT

Recently, generative domain adaptation has achieved remarkable progress, enabling us to adapt a pre-trained generator to a new target domain. However, existing methods are limited to a single target domain and single modality, either text-driven or image-driven. In this paper, we explore a novel task – *Generalized Hybrid Domain Adaptation*. Compared with conventional generative domain adaptation, it provides greater flexibility to adapt the generator to the hybrid of multiple target domains, with multi-modal references including one-shot image and zero-shot text prompt. Meanwhile, it is more challenging to represent the composition of multi-modal target domains and preserve the characteristics from the source domain. To address these issues, we propose UniHDA, a **unified** and **versatile** framework for generalized hybrid domain adaptation. Drawing inspiration from the interpolable latent space of StyleGAN, we find that a linear interpolation between domain shifts in CLIP's embedding space can also uncover favorable compositional capabilities for the adaptation. In light of this finding, we linearly interpolate the domain shifts from multiple target domains to achieve hybrid domain adaptation. To enhance **consistency** with the source domain, we further propose a novel cross-domain spatial structure (CSS) loss that maintains the detailed spatial structure between the source and target generator. Experiments show the adapted generator can synthesize realistic images with various attribute compositions and maintain robust consistency with the source domain. Additionally, UniHDA is generator-agnostic and versatile to multiple generators, *e.g.*, StyleGAN, EG3D, and video generators.

1 INTRODUCTION

Benefiting from the tremendous success of modern image generators (Karras et al., 2019; Brock et al., 2018; Vahdat et al., 2021; Rombach et al., 2022), generative domain adaptation has achieved remarkable progress. Typically, it aims to adapt a pre-trained generator to a new target domain while preserving the variation in the source domain. Depending on the modality of references, generative

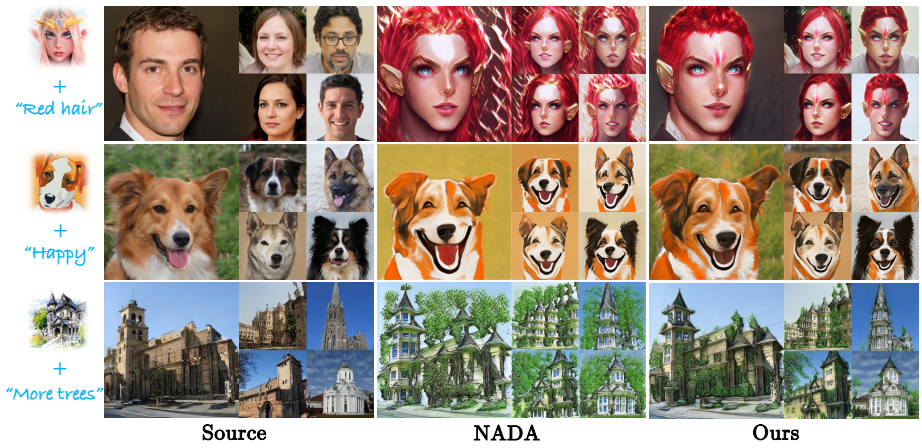

Figure 2: UniHDA succeeds in generalized hybrid domain adaptation and maintains robust consistency with the source domain. NADA (Gal et al., 2021), by cross-model interpolation, can somehow yield alike images of hybrid domains but with poor consistency. It overfits the limited references, which impedes the inheritance of the diversity in the source domain.

domain adaptation can be categorized into two schools: text-driven (Gal et al., 2021; Nitzan et al., 2023a; Liu et al., 2023; Song et al., 2022; Lei et al., 2023) and image-driven (Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022b; Xiao et al., 2022; Mondal et al.; Wu et al., 2023; Zhao et al., 2022a).

Despite their promising results for single modality and single target domain, they fail to adapt the generator to more practical scenarios like `smiling Joker` given the non-smiling `Joker` image and the `smile` text. For more general purposes, we explore a new task – *Generalized Hybrid Domain Adaptation*. It aims to adapt the generator to the hybrid of multiple target domains that integrates all attributes with multi-modal references including one-shot image and zero-shot text prompt (Fig. 1).

Compared with conventional generative domain adaptation, our task is more challenging in two aspects: (1) It is harder to represent the composition of multi-modal target domains. While cross-model interpolation technique (Gal et al., 2021) could somehow yield alike images of the hybrid domain, it doubles the model size and training time to train a separate model per domain. (2) With multiple target domains and very limited references from each domain, the generator is more prone to overfitting domain-specific attributes. This leads to a loss of consistency with the source domain, which impedes the inheritance of the diversity (Fig. 2).

To address these issues, we propose UniHDA, a **Uni**fied and versatile framework for Generalized **H**ybrid **D**omain **A**daptation. UniHDA facilitates the references of one-shot image and text prompt simultaneously and blends the attributes from target domains to create a hybrid domain. To enable multiple modalities, we leverage pre-trained CLIP (Radford et al., 2021) to project multi-modal references into a unified embedding space and represent the domain shift by the direction vector from the source embedding to the target embeddings.

To achieve hybrid domain adaptation, we draw inspiration from the compositional capabilities in the latent space of StyleGAN (Härkönen et al., 2020; Shen & Zhou, 2021; Xu et al., 2022). Specifically, we demonstrate a semantically meaningful linear interpolation between direction vectors in CLIP's embedding space can uncover favorable compositional capabilities (Fig. 3). In light of this intriguing finding, we linearly interpolate direction vectors of multiple target domains to obtain the direction vector corresponding to the hybrid domain that semantically integrates attributes from all target domains.

Furthermore, we introduce a novel Cross-domain Spatial Structure loss (CSS) to preserve the consistency between the source and target generator by maintaining detailed spatial structure information. Concretely, we leverage pre-trained Dino-ViT (Dosovitskiy et al., 2020; Oquab et al., 2023) to encode generated images into patch tokens with fine-grained spatial information. For cross-domain consistency, we maintain the correspondence between source and target tokens with contrastive learning (Oord et al., 2018). Equipped with CSS loss, UniHDA maintains robust consistency with the source domain as shown in Fig. 2.

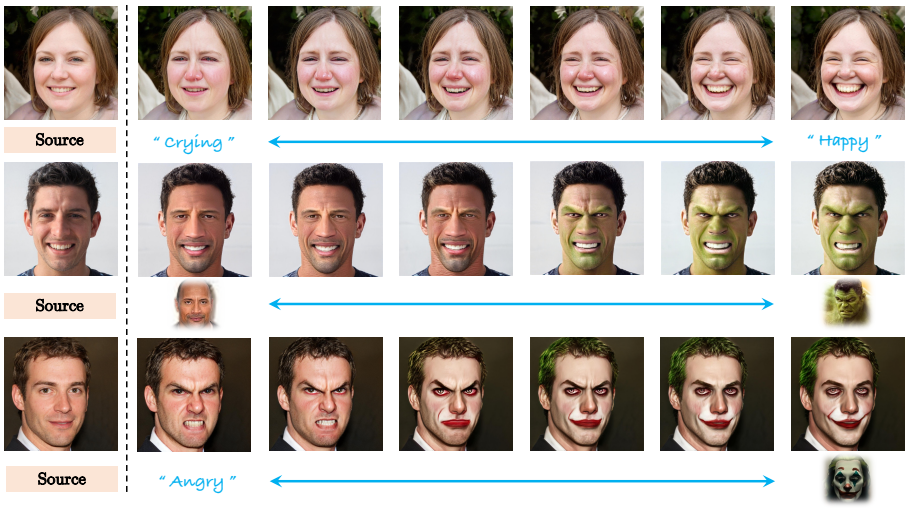

Figure 3: In CLIP's embedding space, we find direction vectors from source domain to target domains are linearly interpolable to represent the domain shift to the hybrid target domain. As shown above, the traversal portrays smooth transition between two target domains and shows favorable compositional capabilities to semantically integrates the attributes. Note that the coefficients for the right domain are respectively 0, 0.2, 0.4, 0.6, 0.8, and 1, while they are inverse for the left domain.

We conduct experiments for a wide range of source and target domains to validate the effectiveness of our method. Results demonstrate that the adapted generator can synthesize realistic images with various attribute compositions. We also show that UniHDA is agnostic to the type of generators, *e.g.*, StyleGAN (Karras et al., 2019; 2020; 2021), EG3D (Chan et al., 2022), video model (Wang et al., 2023), and Diffusion models (Ho et al., 2020; Kim et al., 2022a). Our contributions are as follows:

- We explore a novel task *Generalized Hybrid Domain Adaptation* that offers greater flexibility for hybrid target domain and multi-modal references. To enable it, we propose a **unified** and **versatile** framework which effectively accomplishes the adaptation and is versatile for various generators.
- We demonstrate strong compositional capabilities of direction vectors in CLIP's embedding space. Taking advantage of it, we propose to linearly interpolate the direction vectors for generalized hybrid domain adaptation.
- We propose a cross-domain spatial structure loss to maintain consistency with the source domain. It is conducted in generator-agnostic embedding space which is versatile for various generators. To our knowledge, it is the very first trial in generative domain adaptation.

2 Related Work

Text-driven Generative Domain Adaptation. Text-driven domain adaptation (Gal et al., 2021; Nitzan et al., 2023a; Liu et al., 2023; Alanov et al., 2022; Zhu et al., 2022; Lyu et al., 2023; Lei et al., 2023; Kim & Chun, 2023; Kim et al., 2023; Song et al., 2022) involves using a textual prompt to shift the domain of a pre-trained model toward a new domain. For example, Style-NADA (Gal et al., 2021) presents a local direction CLIP (Radford et al., 2021) loss to align the embeddings of the generated images and text. Based on Style-NADA, Domain Expansion (DE) (Nitzan et al., 2023b) proposes to expand the generator to jointly model multiple domains with texts.

Image-driven Generative Domain Adaptation. Image-driven generative domain adaptation (Mo et al., 2020; Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022b; Xiao et al., 2022; Mondal et al.; Wu et al., 2023; Zhao et al., 2022a; 2023; Zhang et al., 2022; Zhu et al., 2021; Alanov et al., 2023; Kim et al., 2022b) refers to the adaptation of a pre-trained image generator to a new target domain using a limited number of training images. Prior methods often integrate additional regularization terms to prevent overfitting. For instance, DiFa (Zhang et al., 2022) utilizes GAN inversion (Tov et al., 2021) to align the latent codes which helps inherit diversity. Although these works have made significant strides in generative domain adaptation, they heavily rely on the discriminator or generator, making it challenging to handle hybrid domain adaptation and extend to other generators.

Generative Hybrid Domain Adaptation. To achieve hybrid domain adaptation, Style-NADA (Gal et al., 2021) proposes to train a separate generative model per domain and combine their effects in test-time. However, it doubles the model size and training time. Domain Expansion (DE) (Nitzan et al., 2023b) expands the generator to jointly model multiple domains via decomposing latent space. However, it requires the source dataset for regularization, which significantly increases training time. Recently, FHDA (Li et al., 2023) proposes few-shot hybrid domain adaptation and introduces a directional subspace loss. Differently, we focus on multi-modal references including one-shot image, which offers greater flexibility.

Disentanglement in Generative Models. As observed in StyleGAN (Karras et al., 2019), the latent space is essentially a linear subspace. Recent works (Härkönen et al., 2020; Shen & Zhou, 2021; Xu et al., 2022; Shen et al., 2020; Wu et al., 2020; Patashnik et al., 2021; Voynov & Babenko, 2020; Spingarn-Eliezer et al., 2020) propose to find individual latent factors for image variations. Among them, SeFa (Shen & Zhou, 2021) computes the eigenvalues of the transformation matrix to find the latent directions. For diffusion models, DiffAE (Preechakul et al., 2022) explores the possibility of using DPMs for representation learning and extracts a decodable representation of an input image.

3 METHOD

3.1 GENERALIZED HYBRID DOMAIN ADAPTATION

We start with a pre-trained generator $G_{\mathcal{S}}$ (*e.g.*, StyleGAN (Karras et al., 2019; 2020; 2021) and Diffusion model (Ho et al., 2020; Song et al., 2020)), that maps from noise z to images in a source domain $\mathcal{S}$. Given a new target domain $\mathcal{T}$ referenced by texts (Gal et al., 2021; Zhang et al., 2022; Kwon et al., 2022; Kim et al., 2022a) or images (Mo et al., 2020; Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022b; Xiao et al., 2022; Mondal et al.), generative domain adaptation aims to adapt $G_{\mathcal{S}}$ to yield a target generator $G_{\mathcal{T}}$, which can generate images similar to domain $\mathcal{T}$.

Despite the promising results of existing methods, a major limitation of them is that they only support adaptation from the source domain to individual target domains and fail to directly adapt the generator to the hybrid domain which blends the characteristics of multiple domains. Furthermore, they fail with multi-modal adaptation driven by texts and images simultaneously.

For more general purposes, we explore generalized hybrid domain adaptation. Given N domains $\{\mathcal{T}_i\}_{i=1}^{N}$ with one-shot image $\{Y_i\}$ and M domains $\{\mathcal{T}_j\}_{j=1}^{M}$ with the text prompt $\{P_j\}$, it aims to adapt the source generator $G_{\mathcal{S}}$ to $G_{\mathcal{T}}$ that models the hybrid domain $\mathcal{T} = \{\mathcal{T}_i\} \cup \{\mathcal{T}_j\}$ and generates images with integrated characteristics. To the end, we introduce UniHDA, a unified and versatile framework for generalized hybrid domain adaptation (Fig. 4).

3.2 MULTI-MODAL DIRECTION LOSS

To enable multiple modalities, we leverage pre-trained CLIP (Radford et al., 2021) to encode text-image references into a unified semantic embedding space. Drawing inspiration from CLIP-based methods (Gal et al., 2021; Zhang et al., 2022; Kwon et al., 2022; Kim et al., 2022a), we represent the *domain shift* as the direction vector Δf_{dom} from the source embedding to the target embedding. For image reference Y_i and its CLIP embedding f_i, the *domain shift* is calculated by

$$\Delta f_{dom} = f_i - \overline{f_s}, \tag{1}$$

where $\overline{f_s}$ is the mean embedding of several samples generated by $G_{\mathcal{S}}$. For text prompt P_j and its CLIP embedding f_j,

$$\Delta f_{dom} = f_j - \widetilde{f}_s, \tag{2}$$

where $\widetilde{f}_s$ is the embedding of the source text prompt.

To adapt $G_{\mathcal{S}}$, we initialize a new generator $G_{\mathcal{T}}$ from $G_{\mathcal{S}}$ and finetune it by aligning the *sample-shift* direction Δf_{samp} with the *domain-shift* direction Δf_{dom}. Formally,

$$\begin{aligned} \Delta f_{samp} &= f_t - f_s, \\ \mathcal{L}_{direct} &= 1 - \frac{\Delta f_{samp} \cdot \Delta f_{dom}}{\|\Delta f_{samp}\| \, \|\Delta f_{dom}\|}, \end{aligned} \tag{3}$$

where f_s and f_t are the embeddings of samples generated by $G_{\mathcal{S}}$ and $G_{\mathcal{T}}$ with the same noise.

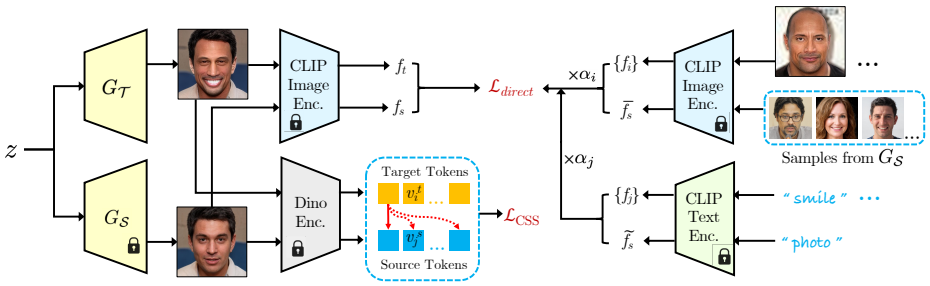

Figure 4: Overview of UniHDA with multi-modal direction loss $\mathcal{L}_{direct}$ and cross-domain spatial structure loss $\mathcal{L}_{\text{CSS}}$. Utilizing CLIP image and text encoder, $\mathcal{L}_{direct}$ encourages $G_{\mathcal{T}}$ to faithfully acquire domain-specific characteristics with multi-modal references. To facilitate diversity inherited from $G_{\mathcal{S}}$, $\mathcal{L}_{\text{CSS}}$ improves cross-domain consistency by maintaining detailed spatial structure information. The red solid line represents positive pairs, while the red dashed lines represent negative pairs.

3.3 Linear Composition of Direction Vectors

To achieve the hybrid domain adaptation, we draw inspiration from the compositional capabilities in the latent space of StyleGAN (Härkönen et al., 2020; Shen & Zhou, 2021; Xu et al., 2022). We illustrate that a linear interpolation between two direction vectors in the embedding space of CLIP, which is semantically meaningful, reveals promising compositional capabilities. As shown in Fig. 3, we can smoothly interpolate between two direction vectors calculated by distinct target prompts and source prompt "`photo`", resulting in a gradual adaptation toward the target domain.

In light of this intriguing finding, we employ linear interpolation on the direction vectors of multi-modal target domains, to derive the direction vector representing the hybrid domain that semantically integrates all attributes. For given domain coefficients $\{\alpha_i\}$ and $\{\alpha_j\}$, we obtain the direction vector

$$\Delta f_{dom} = \sum_{i=1}^{N} \alpha_i (f_i - \overline{f_s}) + \sum_{j=1}^{M} \alpha_j (f_j - \widetilde{f}_s), \tag{4}$$

which represents the *domain shift* between the hybrid domain and source domain. We then substitute Eq. (4) into Eq. (3) to adapt $G_{\mathcal{S}}$ to the hybrid domain.

3.4 Cross-domain Spatial Structure Loss

Albeit the direction loss achieves generalized hybrid domain adaptation, the adapted generator is prone to overfit domain-specific attributes. This exacerbates when it comes to image-image and image-text scenarios owing to the scarcity of the images. To address this issue, we introduce a novel cross-domain spatial structure loss (CSS) to enhance cross-domain consistency, ensuring the preservation of intricate spatial structural information between the source and target generator.

Specifically, we leverage pre-trained Dino-ViT (Dosovitskiy et al., 2020; Oquab et al., 2023) to encode the generated images into patch tokens, containing detailed spatial structural information. Dino-ViT is self-supervised to focus on the distinction between subjects of the same class (Ruiz et al., 2023), which facilitates us in maintaining cross-domain consistency. Motivated by contrastive learning (Oord et al., 2018), we reduce distances between the positive token pairs at the same position and push away the negative token pairs at other positions by

$$\mathcal{L}_{\text{CSS}} = -\sum_{i} \log \frac{\exp(v_i^t \cdot v_i^s)}{\sum_j \exp(v_i^t \cdot v_j^s)}, \tag{5}$$

where v_i^t and v_j^s are the i-th and j-th tokens in the last layer of Dino-ViT from $G_{\mathcal{T}}$ and $G_{\mathcal{S}}$ respectively. The dot mark $\cdot$ represents dot product.

Overall, our training loss consists of two terms, i.e., the multi-modal direction loss $\mathcal{L}_{direct}$ to achieve generalized hybrid domain adaptation and the cross-domain spatial structure loss $\mathcal{L}_{\text{CSS}}$ to maintain cross-domain consistency:

$$\mathcal{L}_{overall} = \mathcal{L}_{direct} + \lambda \mathcal{L}_{\text{CSS}}. \tag{6}$$

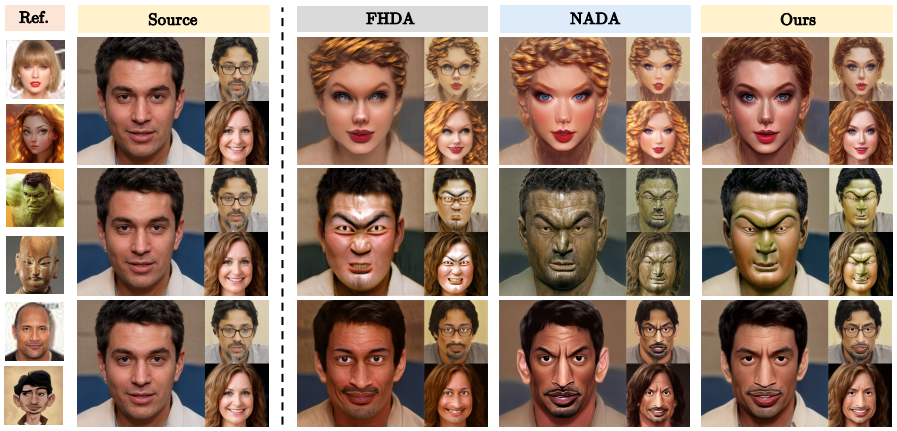

Figure 5: **Image-image** hybrid domain adaptation. We compare the results of FHDA (Li et al., 2023), NADA (Gal et al., 2021) and UniHDA (Ours) with the same noise. FHDA and NADA generate images with poor cross-domain consistency, leading to a limited diversity. In contrast, UniHDA alleviates overfitting and maintains strong cross-domain consistency.

Method	Taylor-Elena		Hulk-Wooden		Johnson-Comic		Average	
	CS-I (↑)	SCS (↑)	CS-I (↑)	SCS (↑)	CS-I (↑)	SCS (↑)	CS-I (↑)	SCS (↑)
FHDA	0.685	0.576	0.635	0.659	0.640	0.679	0.630	0.661
NADA	0.684	0.579	0.624	0.575	0.647	0.642	0.628	0.639
Ours	**0.699**	**0.738**	**0.649**	**0.707**	**0.656**	**0.764**	**0.642**	**0.769**

Table 1: Quantitative results for **image-image** domain adaptation. We present the results for cases in Fig. 5. To further demonstrate the robustness, we average the results for more cases in Appendix.

4 Experiments

4.1 Experimental Settings

Methodology. We demonstrate the versatility of UniHDA on generalized hybrid domain adaptation, *i.e.*, image-image, image-text, and text-text (in Appendix). To show the generator-agnostic nature of UniHDA, we apply it to three well-known generators, *i.e.*, StyleGAN2 (Karras et al., 2020), Diffusion model (Kim et al., 2022a), and EG3D (Chan et al., 2022). Following previous generative domain adaptation literatures (Gal et al., 2021; Zhang et al., 2022; Mo et al., 2020; Li et al., 2020; Nitzan et al., 2023b; Ojha et al., 2021; Zhao et al., 2022b; Xiao et al., 2022; Mondal et al.), we use StyleGAN2 for comparisons in most experiments.

Datasets. We conduct experiments for a wide range of source and target domains to validate the effectiveness of UniHDA. Following previous work, we consider FFHQ (Karras et al., 2019), AFHQ-Dog (Choi et al., 2020), and LSUN-Church (Yu et al., 2015) as the source domains. The resolutions of images in these datasets are respectively 1024, 512, and 256. We adapt the generator to diverse hybrid domains driven by the text prompt and one-shot image. To demonstrate the effect of the hybrid domain, *we set the domain coefficients in Eq.* (4) *as 0.5*. Unless stated otherwise, we use $\lambda = 5$ in Eq. (6) for all experiments.

Evaluation Metrics. One important aspect of evaluating generative domain adaptation is the preservation of domain-specific characteristics. Following (Ruiz et al., 2023), we use CLIP Score (CS-T and CS-I) for text-text and image-image adaptation respectively. Concretely, CS-T is measured by the average cosine similarity between the target prompt and generated images' embedding. CS-I is the average pairwise cosine similarity between CLIP embeddings of real and generated images. Here we use the average CS-T or CS-I of multiple domains. For image-text adaptation, we use the average of CS-T and CS-I as the metric (CS). Another important evaluation is the cross-domain consistency of the source domain. To measure it, we adopt the Structural Consistency Score (SCS) (Xiao et al., 2022) to evaluate the spatial structural consistency between the source and target generator.

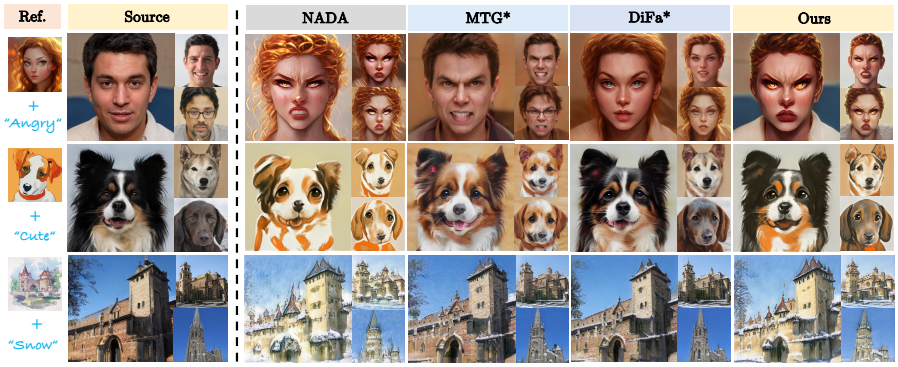

Figure 6: **Image-text** hybrid domain adaptation. We compared our method with previous methods (Gal et al., 2021; Zhu et al., 2021; Zhang et al., 2022), UniHDA well captures the attributes of hybrid target domain and maintains strong cross-domain consistency with source domain. $*$ indicates that MTG and DiFa support multi-modalities by interpolating model parameters with NADA.

Figure 7: Hybrid domain adaptation in 3D generator. To show the versatility of UniHDA, we apply it on the popular 3D-aware generator, EG3D (Chan et al., 2022).

4.2 Image-Image Hybrid Domain Adaptation

Fig. 5 shows the qualitative results for image-image adaptation, starting from the same source domain FFHQ (Karras et al., 2019) to the combinations of individual domains. As shown in the figure, FHDA (Li et al., 2023) suffers from severe model collapse and generates images with limited diversity due to the scarcity of image references. While NADA (Gal et al., 2021) mitigates overfitting to a certain extent, its cross-domain consistency remains poor, resulting in the generation of similar images. In contrast, UniHDA maintains strong consistency and effectively generates images with integrated characteristics.

Method	FFHQ		Dog		Church	
	CS (↑)	SCS (↑)	CS (↑)	SCS (↑)	CS (↑)	SCS (↑)
NADA	0.563	0.586	0.424	0.533	**0.414**	0.629
MTG	0.536	0.529	0.403	0.526	0.403	0.684
DiFa	0.548	0.681	0.413	0.683	0.407	0.711
Ours	**0.565**	**0.742**	**0.430**	**0.796**	**0.414**	**0.781**

Table 2: Quantitative results for **image-text** domain adaptation. We average the results for cases in Appendix.

We also quantitatively compare UniHDA with baselines. As shown in Tab. 1, ours clearly outperforms them. For CS-I, UniHDA significantly outperforms other methods, indicating that generated images effectively integrate multiple characteristics from distinct domains. Furthermore, UniHDA achieves better SCS, which effectively maintains cross-domain consistency compared with baselines.

4.3 Image-Text Hybrid Domain Adaptation

Fig. 6 shows the results of image-text adaptation, including FFHQ (Karras et al., 2019), AFHQ-Dog (Choi et al., 2020), and LSUN -Church (Yu et al., 2015). As depicted in Sec. 4.2, NADA is susceptible to overfitting, which retains poor cross-domain consistency. Besides, we interpolate NADA's parameters with MTG(Zhu et al., 2021) and DiFa (Zhang et al., 2022), which alleviates overfitting to some extent. However, they can't accurately capture the attributes of the hybrid target domain and still fail to maintain good consistency. In contrast, UniHDA well captures the attributes

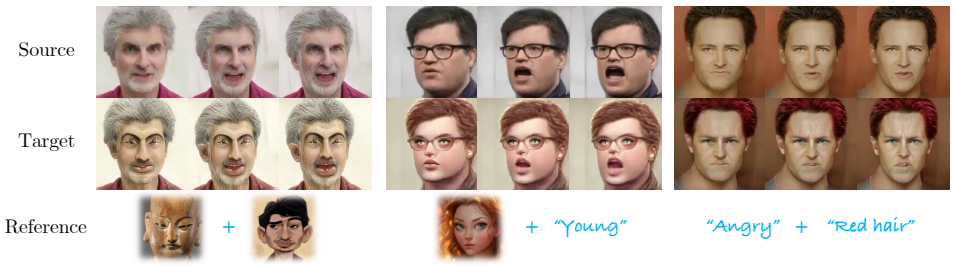

Figure 8: Hybrid domain adaptation in the video generator. Specifically, we apply it to StyleInv (Wang et al., 2023), an unconditional video generator to synthesize high-quality videos.

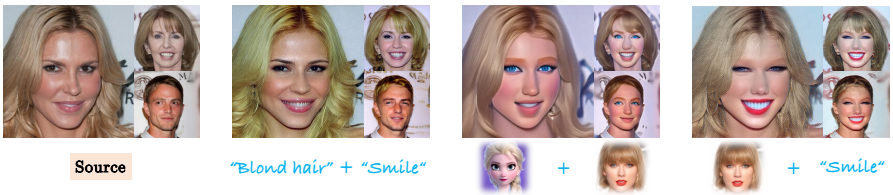

Figure 9: Results of UniHDA with DiffusionCLIP (Kim et al., 2022a), which demonstrate UniHDA is agnostic to the type of generator, allowing for broader application on diffusion models.

and achieves robust consistency in all scenarios. As shown in Tab. 2, we also compare UniHDA with the baselines quantitatively. Consistent with qualitative results in Fig. 6, ours clearly outperforms the baselines. Additionally, we also conduct the user study in the Appendix.

4.4 Generalization on other generative models

In this section, we demonstrate UniHDA is agnostic to the type of generative models and can easily generalize to other generators, *e.g.*, EG3D (Chan et al., 2022), StyleInv (Wang et al., 2023), and DiffusionCLIP (Kim et al., 2022a). For EG3D, a popular 3D-aware image generation method, we replace the discrimination loss with our framework for hybrid domain adaptation. As shown in Fig. 7, the results effectively integrate the attributes and preserve the characters and poses of the source domain. For video adaptation, we conduct experiments on StyleInv (Wang et al., 2023), an unconditional video generator. Fig. 8 verifies the generalization to synthesize high-quality videos. For DiffusionCLIP, we replace the training objective of DiffusionCLIP with our proposed $\mathcal{L}_{direct}$ and $\mathcal{L}_{\text{CSS}}$. As shown in Fig. 9, the results integrate the characteristics from multiple target domains and maintain robust consistency with the source domain. More results are included in the Appendix.

4.5 Comparison with Text-to-Image Generators

Recent text-to-image generators like IP-Adapter (Ye et al., 2023) could synthesize promising results with provided image and text prompt. However, the objective of UniHDA is to generate images with attributes of the generalized hybrid target domains while maintaining considerable diversity from the source domain, which can be applicable in scenarios like *data collection*. As shown in Fig. 10, while IP-Adapter can produce images with multiple attributes, their diversity often diminishes when generating large quantities of images. This is due to their inability to retain the distribution of the original domain, which makes it impractical for *data collection*.

4.6 Comparison with Image Editing

Image editing could somehow yield alike images with the attribute of the target domain. However, UniHDA holds several key differences: (1) Existing editing methods (Patashnik et al., 2021; Duan et al., 2023; Lyu et al., 2023) are based on inversion techniques, which inherently involve the loss of information. Multi-attribute editing exacerbates this problem. As shown in Fig. 11, after multiple edits, the resulting images show poor consistency with the source domain. However, our results maintain strong consistency. (2) UniHDA can adapt generators to more composite and expansive domains, which offers greater flexibility. However, during the editing, subsequent edits often overwrite the previously modified attributes, making it difficult to generate images with a

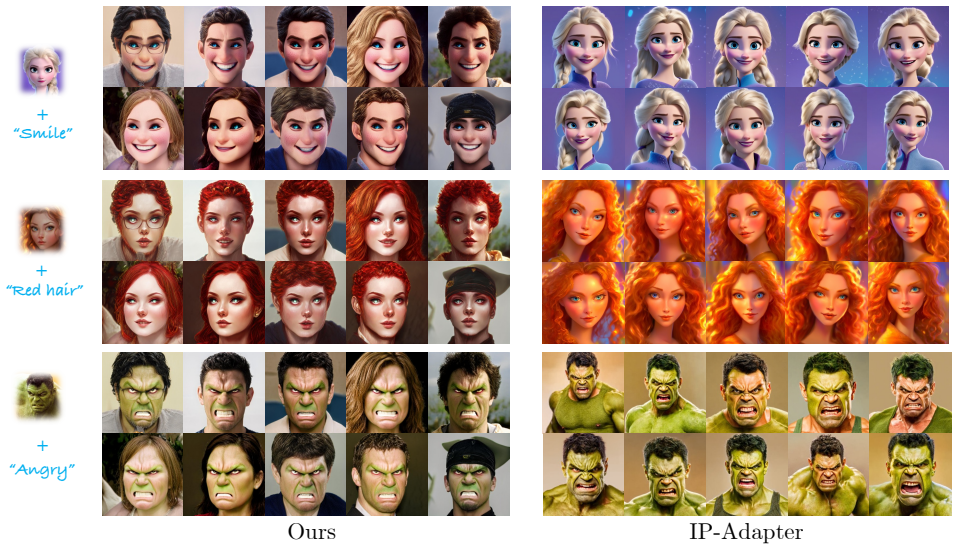

Figure 10: Comparison between UniHDA and IP-Adapter (Ye et al., 2023). Given a single image reference, We adopt the IP-Adapter to generate the results with a fixed text prompt and different random seeds. IP-Adapter tends to overfit the single reference and lose the diversity. Differently, UniHDA maintains robust consistency with the source domain and presents compelling diversity.

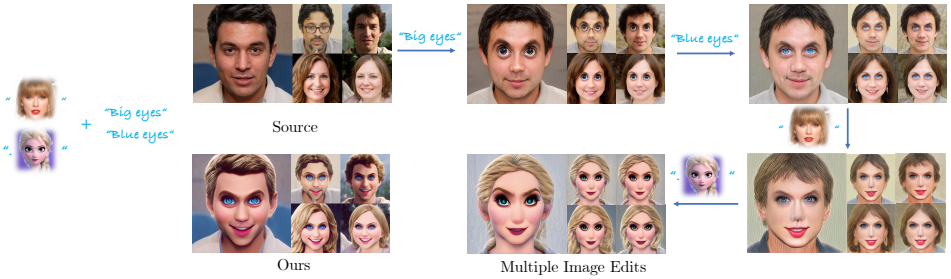

Figure 11: Comparison between UniHDA and image editing. We use StyleCLIP (Patashnik et al., 2021) to perform multiple edits on images for hybrid attributes, which presents two issues: (1) Multiple inversions exacerbate information loss, leading to decreased consistency. (2) Subsequent edits may overwrite the effects of previous edits, as observed that the big eyes attribute disappears after the blue eyes edit.

combination of hybrid attributes as shown in Fig. 11. (3) UniHDA is generator-agnostic and versatile to multiple generators, including 3D generators and video generators. However, 2D image editing is challenging to apply to 3D images or videos because it is difficult to maintain robust multi-view or temporal consistency. (4) The objective of generative domain adaptation is to generate images with attributes of the target domain while maintaining considerable diversity, which can be applicable in scenarios like *data collection*. Image editing, on the other hand, requires the original image as input, rendering it impractical for such applications.

4.7 Results of Incompatible Domain Adaptation

Typically, the attributes of the target domain and the source domain are complementary. Even in cases where conflicts arise, our UniHDA can maintain robust consistency with the source domain while acquiring the attributes of the target domain. To verify this, we additionally conduct experiments for hybrid domain adaptation on incompatible domains, *i.e.*, from cat to rabbit. As shown in Fig. 12, we start from AFHQ (Choi et al., 2020) cat to incompatible domains. Although there is a conflict between the reference image and the source domain, we can observe that the results still effectively integrate the attributes of the corresponding domain and maintain robust consistency with the source domain.

Figure 12: Hybrid domain adaptation from AFHQ-Cat to incompatible domains.

Figure 13: Ablation of our proposed $\mathcal{L}_{\text{CSS}}$ on hybrid domain adaptation, which significantly alleviates overfitting and improves cross-domain consistency. λ is the coefficient of $\mathcal{L}_{\text{CSS}}$.

4.8 Ablation of CSS Loss

We conduct the ablation study to evaluate the effects of our proposed CSS loss. As shown in Fig. 13, the results without $\mathcal{L}_{\text{CSS}}$ suffer from overfitting and have very limited cross-domain consistency, *e.g.*, distorted backgrounds in Row 1 and 3. Benefiting from $\mathcal{L}_{\text{CSS}}$, the generated images maintain consistency with the source images in terms of spatial structure, thereby inheriting the diversity from the source domain. Besides, we conduct the quantitative ablation in the Appendix. There exists a trade-off between adaptation to the target domain and preserving the characteristics from the source domain. We can adjust the coefficient λ based on the desired effect.

5 Conclusion & Limitation

In this paper, we explore a new task, generalized hybrid domain adaptation, and propose UniHDA, a unified and versatile framework to enable it. For the hybrid domain, we demonstrate the compositional capabilities of direction vectors in CLIP's embedding space and linearly interpolate direction vectors of multiple target domains. We also propose a new cross-domain spatial structure loss to improve consistency, which is conducted in generator-agnostic space and versatile for various generators. We believe our work is an important step towards generative domain adaptation, since we have demonstrated the source generator can be effectively adapted to a hybrid domain with multi-modal references and maintain robust cross-domain consistency. Our code will be made public.

While UniHDA effectively realizes generalized hybrid domain adaptation, it also has limitations. To encode both image and text into a shared embedding space, we utilize pre-trained CLIP during training time, which might bring potential bias for some domains. Nevertheless, we believe that the exploration of the novel task is significant for future work and solutions could be integrated into UniHDA to eliminate the bias.

6 Ethics Statement

Our main objective in this work is to empower novice users to generate visual content creatively and flexibly. However, the broad adoption of such technology brings up ethical issues related to privacy, misinformation, and potential misuse. We strongly support the responsible development and deployment of tools to detect biases and malicious use cases, highlighting the need for ethical standards to guarantee their safe and ethical use in the field of computer vision.

7 Reproducibility Statement

We make the following efforts to ensure the reproducibility of UniHDA: (1) Our training and inference codes together with the trained model weights will be publicly available. (2) We provide the details of the human evaluation setups in the appendix (Appendix A.5). (3) We provide training details in the appendix (Appendix A.8), which is easy to follow.

References

Aibek Alanov, Vadim Titov, and Dmitry P Vetrov. Hyperdomainnet: Universal domain adaptation for generative adversarial networks. *Advances in Neural Information Processing Systems*, 35: 29414–29426, 2022.

Aibek Alanov, Vadim Titov, Maksim Nakhodnov, and Dmitry Vetrov. Styledomain: Efficient and lightweight parameterizations of stylegan for one-shot and few-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2184–2194, October 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yuxuan Duan, Li Niu, Yan Hong, and Liqing Zhang. Weditgan: Few-shot image generation via latent space relocation. *arXiv preprint arXiv:2305.06671*, 2023.

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in Neural Information Processing Systems*, 33:9841–9850, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems,Neural Information Processing Systems*, Jan 2020.

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

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 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.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14203–14213, 2023.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022a.

Gwanghyun Kim, Ji Ha Jang, and Se Young Chun. Podia-3d: Domain adaptation of 3d generative model across large domain gap using pose-preserved text-to-image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22603–22612, 2023.

Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022b.

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

Biwen Lei, Kai Yu, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. Diffusiongan3d: Boosting text-guided 3d generation and domain adaption by combining 3d gans and diffusion priors. *arXiv preprint arXiv:2312.16837*, 2023.

Hengjia Li, Yang Liu, Linxuan Xia, Yuqi Lin, Tu Zheng, Zheng Yang, Wenxiao Wang, Xiaohui Zhong, Xiaobo Ren, and Xiaofei He. Few-shot hybrid domain adaptation of image generators. *arXiv preprint arXiv:2310.19378*, 2023.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4804–4814, 2022.

Yijun Li, Richard Zhang, Jingwan (Cynthia) Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15885–15896. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b6d767d2f8ed5d21a44b0e5886680cb9-Paper.pdf.

Zhenhuan Liu, Liang Li, Jiayu Xiao, Zheng-Jun Zha, and Qingming Huang. Text-driven generative domain adaptation with spectral consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7019–7029, October 2023.

Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6894–6903, 2023.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *CVPR AI for Content Creation Workshop*, 2020.

Arnab Kumar Mondal, Piyush Tiwary, Parag Singla, and AP Prathosh. Few-shot cross-domain image generation via inference-time latent-code learning. In *The Eleventh International Conference on Learning Representations*.

Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, and Eli Shechtman. Domain expansion of image generators. 2023a.

Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, and Eli Shechtman. Domain expansion of image generators. *arXiv preprint arXiv:2301.05225*, 2023b.

Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10743–10752, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1532–1540, 2021.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv: Learning,arXiv: Learning*, Oct 2020.

Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Diffusion guided domain adaptation of image generators. *arXiv preprint arXiv:2212.04473*, 2022.

Nurit Spingarn-Eliezer, Ron Banner, and Tomer Michaeli. Gan" steerability" without optimization. *arXiv preprint arXiv:2012.05328*, 2020.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021.

Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. *arXiv preprint arXiv:2002.03754*, 2020.

Yuhan Wang, Liming Jiang, and Chen Change Loy. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22851–22861, 2023.

Yi Wu, Ziqiang Li, Chaoyue Wang, Heliang Zheng, Shanshan Zhao, Bin Li, and Dacheng Ta. Domain re-modulation for few-shot generative domain adaptation. *arXiv preprint arXiv:2302.02550*, 2023.

Zongze Wu, Dani Lischinski, and Eli Shechtman. StyleSpace analysis: Disentangled controls for StyleGAN image generation. *arXiv:2011.12799*, 2020.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11204–11213, 2022.

Guodong Xu, Yuenan Hou, Ziwei Liu, and Chen Change Loy. Mind the gap in distilling stylegans. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pp. 423–439. Springer, 2022.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Yabo Zhang, mingshuai Yao, Yuxiang Wei, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. Towards diverse and faithful one-shot adaption of generative adversarial networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=IXoHxXIGpyV.

Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Man Cheung. Few-shot image generation via adaptation-aware kernel modulation. *Advances in Neural Information Processing Systems*, 35:19427–19440, 2022a.

Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9140–9150, 2022b.

Yunqing Zhao, Chao Du, Milad Abdollahzadeh, Tianyu Pang, Min Lin, Shuicheng Yan, and Ngai-Man Cheung. Exploring incompatible knowledge transfer in few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7380–7391, 2023.

Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. *arXiv preprint arXiv:2110.08398*, 2021.

Yiming Zhu, Hongyu Liu, Yibing Song, Ziyang Yuan, Xintong Han, Chun Yuan, Qifeng Chen, and Jue Wang. One model to edit them all: Free-form text-driven image manipulation with semantic modulations. *Advances in Neural Information Processing Systems*, 35:25146–25159, 2022.

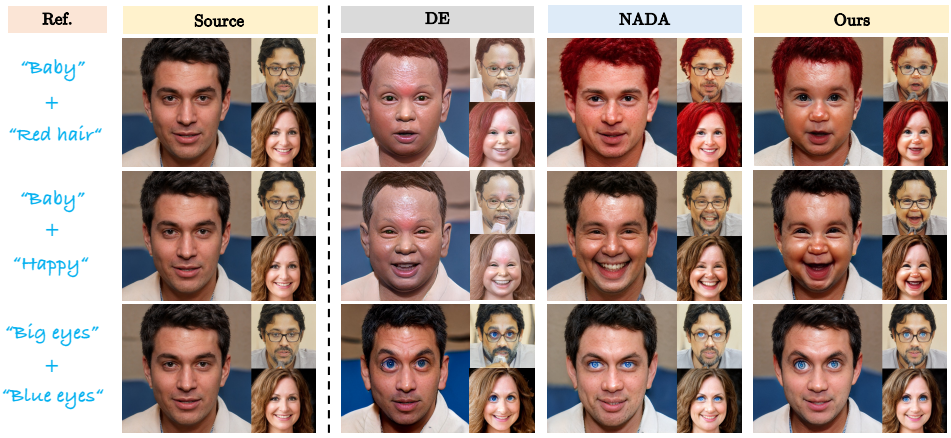

Figure 14: **Text-text** hybrid domain adaptation. We compare the results of DE (Nitzan et al., 2023b), NADA (Gal et al., 2021) and UniHDA (Ours) with the same noise. UniHDA exhibits desirable performance to acquire characteristics from hybrid target domain and maintain robust cross-domain consistency.

Method	`Baby-Red hair`		`Baby-Happy`		`Big-Blue eyes`		Average	
	CS-T (↑)	SCS (↑)	CS-T (↑)	SCS (↑)	CS-T (↑)	SCS (↑)	CS-T (↑)	SCS (↑)
DE	0.163	0.638	0.160	0.580	0.195	0.662	0.167	0.634
NADA	0.179	0.661	0.170	0.642	0.186	0.731	0.159	0.552
Ours	**0.186**	**0.744**	**0.175**	**0.757**	**0.197**	**0.765**	**0.176**	**0.707**

Table 3: Quantitative results for **text-text** domain adaptation. We present the quantitative results corresponding to each case in Fig. 14. Similar to Tab. 1, we average the results for cases in Appendix.

A Appendix

In this appendix, we begin to conduct the experiments on text-text hybrid domain adaptation in Appendix A.1. And we compare with existing methods in terms of efficiency in Appendix A.2. Then we show more ablation of CSS loss in Appendix A.3 and Appendix A.4, including quantitative results and the effect of the encoder in CSS loss. Additionally, we report the user study in Appendix A.5. Then we show more qualitative results *e.g.*, more domains in Appendix A.6, and additional results in Appendix A.7. Finally, we provide more implement details in Appendix A.8 and the potential bias in Appendix A.9.

A.1 Text-text Hybrid Domain Adaptation

Fig. 14 shows the qualitative results for text-text adaptation. Since the adaptation is conducted solely along one projection direction of the latent code, Domain Expansion (DE) (Nitzan et al., 2023b), does not fully capture the characteristics of the target domain, *e.g.*, `baby` (Row 1 and Row 2). Furthermore, DE does not maintain robust consistency, *e.g.*, the chin of the person in the upper-right corner and background artifacts in Row 3. The problem of NADA (Gal et al., 2021) is overfitting. Hard-to-learn characteristics, *e.g.*, `baby` (Row 1 and Row 2) and `big eyes` (Row 3) may be overshadowed by other overfitted ones. In contrast, UniHDA (Ours) exhibits desirable performance to generate images with integrated characteristics while maintaining robust consistency with the source domain.

Similar to Sec. 4.2, we also compare UniHDA with the baselines quantitatively. As shown in Tab. 3, ours clearly outperforms the baselines, which are consistent with qualitative results in Fig. 14. We achieve better CS-I and SCS, indicating that generated images effectively integrate domain-specific attributes and preserve primary characteristics of the source domain.

Method	Modality	Model Amount	2-domain		10-domain	
			size (↓)	time (↓)	size (↓)	time (↓)
NADA	Multi	N	48M	4min	240M	20min
MTG*	Multi	N	48M	4min	240M	20min
DiFa*	Multi	N	48M	4min	240M	20min
DE†	Text	1	24M	20h	24M	20h
FHDA	Image	1	24M	3min	24M	3min
Ours	Multi	1	**24M**	**2min**	**24M**	**2min**

Table 4: Comparison with previous methods. $*$ indicates MTG and DiFa support multi-modalities by interpolating model parameters with NADA. † means DE needs source dataset (e.g., FFHQ) that significantly increases training time.

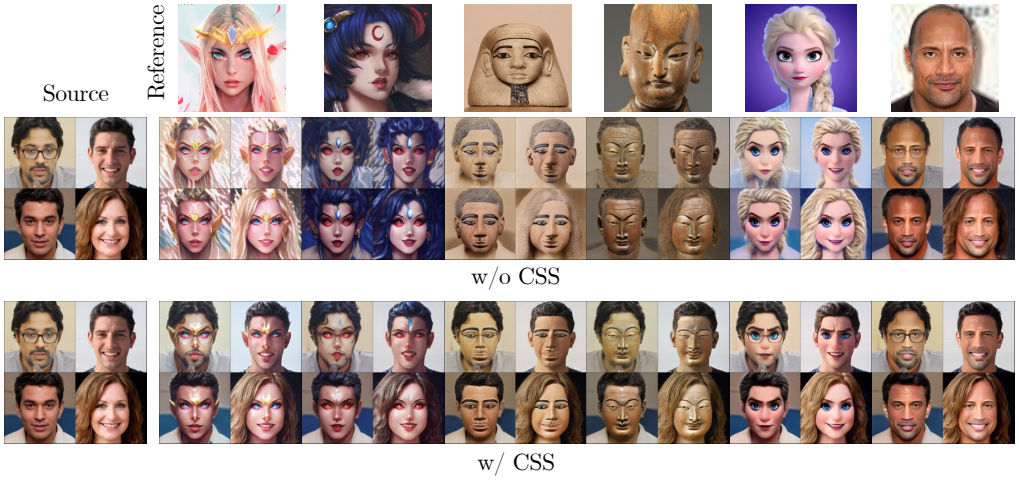

Figure 15: More qualitative results to verify the effectiveness of our proposed $\mathcal{L}_{\text{CSS}}$.

A.2 Comparison with Existing Methods

In addition to generation quality, UniHDA also surpasses existing methods in terms of efficiency, *e.g.*, model size and training time as shown in Tab. 4. NADA, MTG, and DiFa trains a separate generative model per domain and interpolates their parameters in test-time, which necessitates multiple times the model size and training time. Although DE avoids cross-model interpolation, it heavily relies on the large source dataset for regularization during training process, resulting in a significant increase in training time. In contrast, UniHDA circumvents these issues, which enables the adaptation within single generator in only two minutes.

Furthermore, DE relies on the semantic latent space of the generator (*e.g.*, StyleGAN (Karras et al., 2019) and DiffAE (Preechakul et al., 2022)) for hybrid domain adaptation, limiting its applicability to a broader range of generators. MTG and DiFa utilize GAN inversion, which restricts the applicability to generators similar to StyleGAN. Conversely, UniHDA is not constrained by the type of generators, allowing for its broader application across various generators.

A.3 More Ablation of CSS Loss

As depicted in Sec. 4.8 of the main paper, our proposed $\mathcal{L}_{\text{CSS}}$ significantly alleviates overfitting and improves cross-domain consistency. Results with $\mathcal{L}_{\text{CSS}}$ achieve better SCS score, indicating that they maintain stronger consistency with the source domain. Additionally, we show more qualitative results in Fig. 15 to verify the effectiveness of UniHDA. We also conduct the quantitative ablation in Tab. 5, which is consistent with the qualitative results.

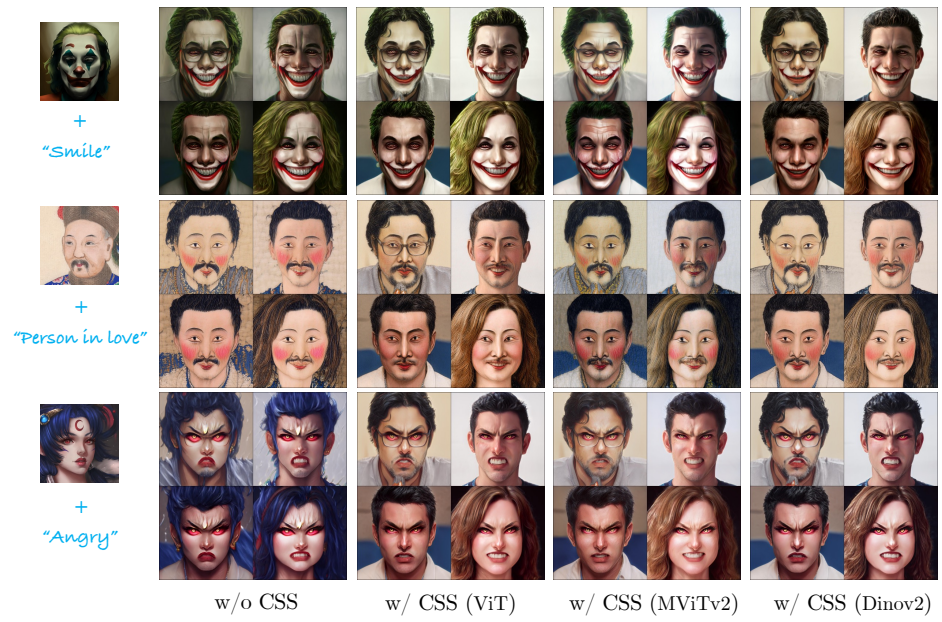

Figure 16: Ablation of different pre-trained encoders for CSS on hybrid domain adaptation.

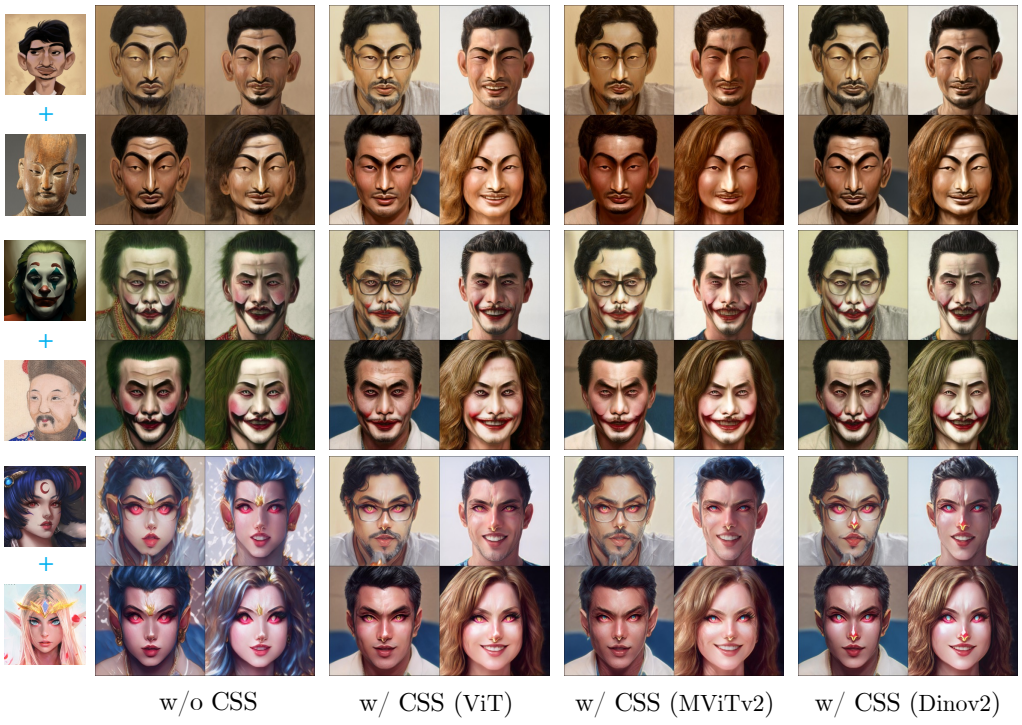

Figure 17: Effect of different pre-trained image encoders for CSS on image-image hybrid domain adaptation.

A.4 Ablation of Encoder for CSS

We conduct experiments on pre-trained ViT (Dosovitskiy et al., 2020), MViTv2 (Li et al., 2022), and Dinov2 to explore the impact of different image encoders for CSS. As shown in Fig. 16, Fig. 17 and

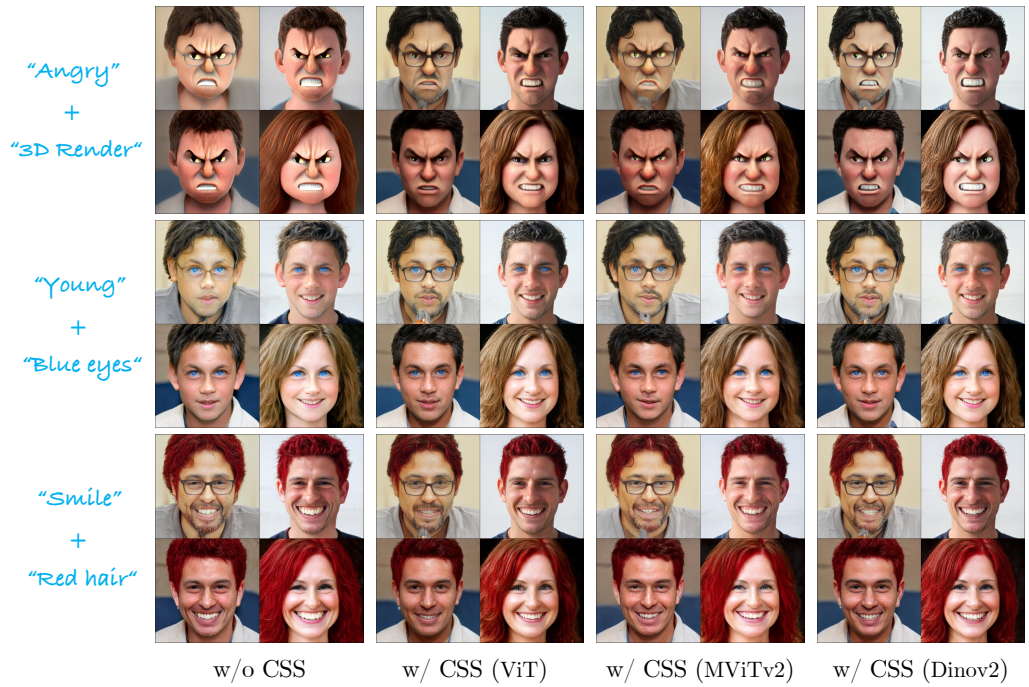

Figure 18: Effect of different pre-trained image encoders for CSS on text-text hybrid domain adaptation.

λ	FFHQ (I-I)		FFHQ (T-T)		FFHQ (T-I)		Dog (T-I)		Church (T-I)	
	SCS (↑)	CS-I (↑)	SCS (↑)	CS-T (↑)	SCS (↑)	CS (↑)	SCS (↑)	CS (↑)	SCS (↑)	CS (↑)
0	0.502	0.639	0.520	0.170	0.562	0.557	0.491	**0.430**	0.604	0.411
3	0.681	0.638	0.683	0.171	0.694	0.556	0.787	0.428	0.706	0.413
5	**0.769**	0.642	0.707	**0.176**	0.742	**0.565**	0.796	**0.430**	**0.781**	**0.414**
10	0.762	**0.655**	**0.756**	0.170	**0.773**	0.514	**0.798**	0.425	0.778	0.410

Table 5: Quantitative ablation for our proposed $\mathcal{L}_{\text{CSS}}$. λ is the coefficient of $\mathcal{L}_{\text{CSS}}$. There exists a trade-off between adaptation to the target domain and preserving the characteristics from the source domain. We can adjust λ based on the desired effect.

Fig. 18, we can observe that all of them improve the consistency with source domain compared with the baseline approach. Furthermore, they exhibit a similar qualitative style, which demonstrates that our CSS is agnostic to different pre-trained image encoders.

Method	Fidel.	Diver.	Corr.
vs. NADA (I-I)	85.2	90.6	76.0
vs. NADA (T-T)	81.4	84.2	80.8
vs. NADA (T-I)	84.6	85.8	78.6

Table 6: User study for fidelity, diversity, and reference correspondence (image or text) in hybrid domain adaptation. The value (%) represents the percentage of users who favor the images generated by our method over NADA.

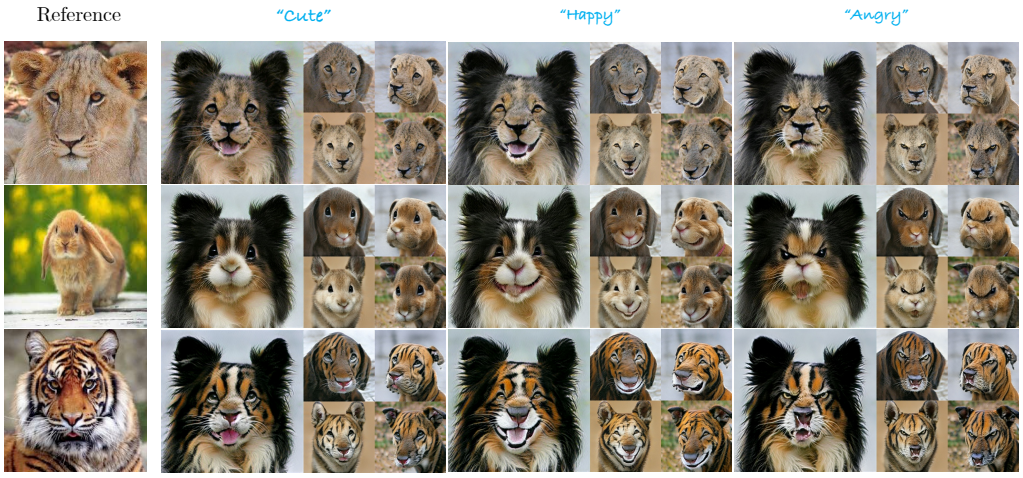

Figure 19: Hybrid domain adaptation from AFHQ-Dog to incompatible domains, *i.e.*, `lion`, `rabbit`, and `tiger`.

Figure 20: The results of hybrid domain adaptation from FFHQ to the hybrid of more domains.

A.5 User Study

We conduct user study in Tab. 6 to compare UniHDA with NADA. Specifically, we provide users with the target text or image, the source image, and adapted images. Then we ask them to choose the better image for fidelity, diversity and correspondence. For each case, we generate 1000 samples and randomly assign 200 samples to 30 users. The results indicate that UniHDA surpasses NADA in terms of fidelity, diversity and reference correspondence.

A.6 More Qualitative Results

We apply UniHDA to adapt the generator on FFHQ (Karras et al., 2019) to more hybrid domains, *i.e.*, text-text, image-image, and image-text, as well as AFHQ (Choi et al., 2020) `dog` to incompatible domains. As shown in Fig. 20, Fig. 21, Fig. 22, Fig. 23, and Fig. 19. UniHDA successfully generates images with integrated characteristics from multiple target domains and maintains robust consistency with the source domain. Besides, we showcase more results of hybrid domain adaptation from AFHQ-Dog and LSUN-Church (Yu et al., 2015) in Fig. 24 and Fig. 25.

A.7 More Results for DiffusionCLIP and EG3D

To demonstrate the versatility of UniHDA, we apply it on DiffusionCLIP and EG3D in Sec. 4.4 in the main paper. As shown in Fig. 26 and Fig. 27, we showcase more results including image-image, text-text, and image-text. All results achieve hybrid domain adaptation and preserve strong cross-domain consistency.

Figure 21: More results of image-image hybrid domain adaptation. The source image is in the top-left corner, and the first row and column consist of training images.

A.8 Implement Details

Following the setting of previous generative domain adaptation methods (Gal et al., 2021; Nitzan et al., 2023b), we utilize the batch size of 4 and ADAM Optimizer with a learning rate of 0.002 for all experiments during training. A training session typically requires 300 iterations in 2 minutes, which significantly reduces training time compared with adversarial methods for generative domain adaptation. Note that we conduct all experiments on a single NVIDIA RTX 4090 GPU. The code will be open source.

For experiments on FFHQ, we generate images with 1024×1024 resolution. As for AFHQ-Dog and LSUN-Church, we operate on 512×512 and 256×256 resolution images respectively.

A.9 Potential biases of CLIP

As depicted in NADA, CLIP may introduce textual bias and ambiguity in some specific domains. For example, the text 'Nurse' tends to convert the individuals to females, as shown in Fig. 28.

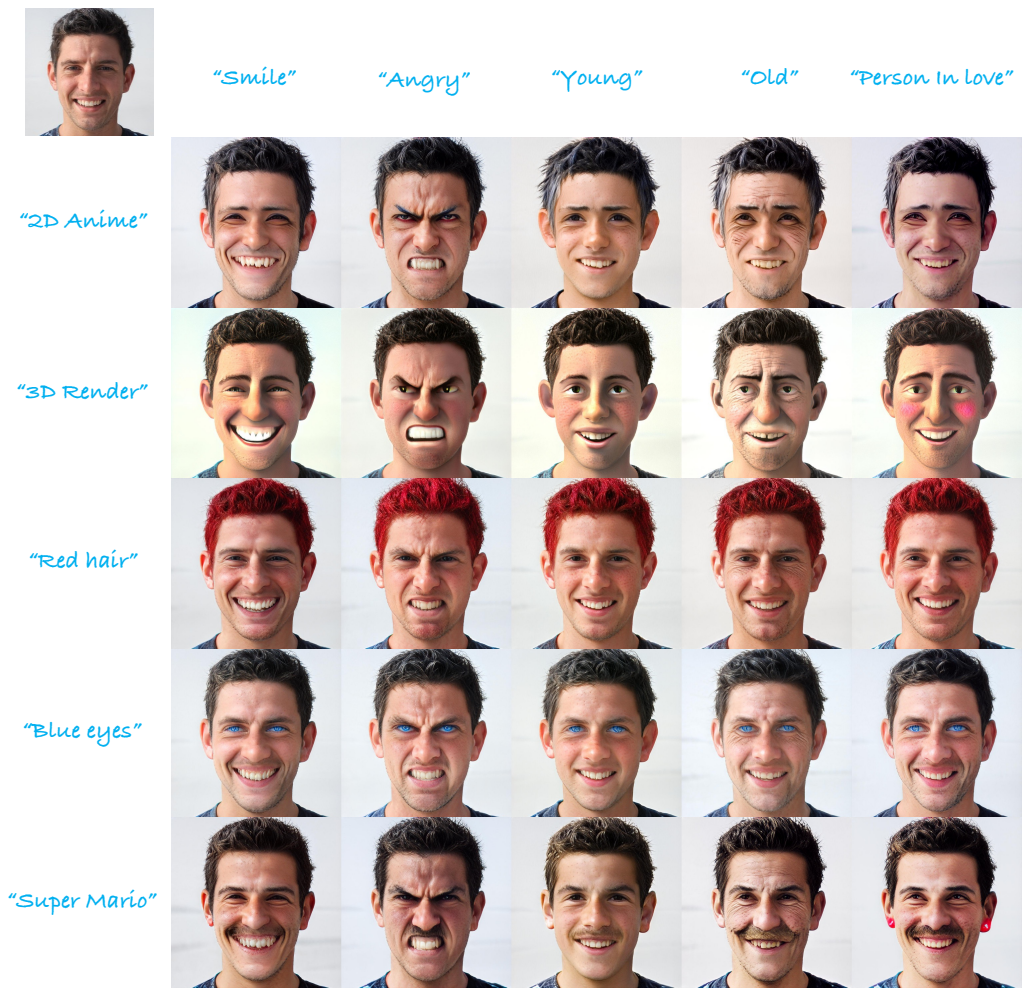

Figure 22: More results of text-text hybrid domain adaptation. The source image is in the top-left corner, and the first row and column consist of text prompts.

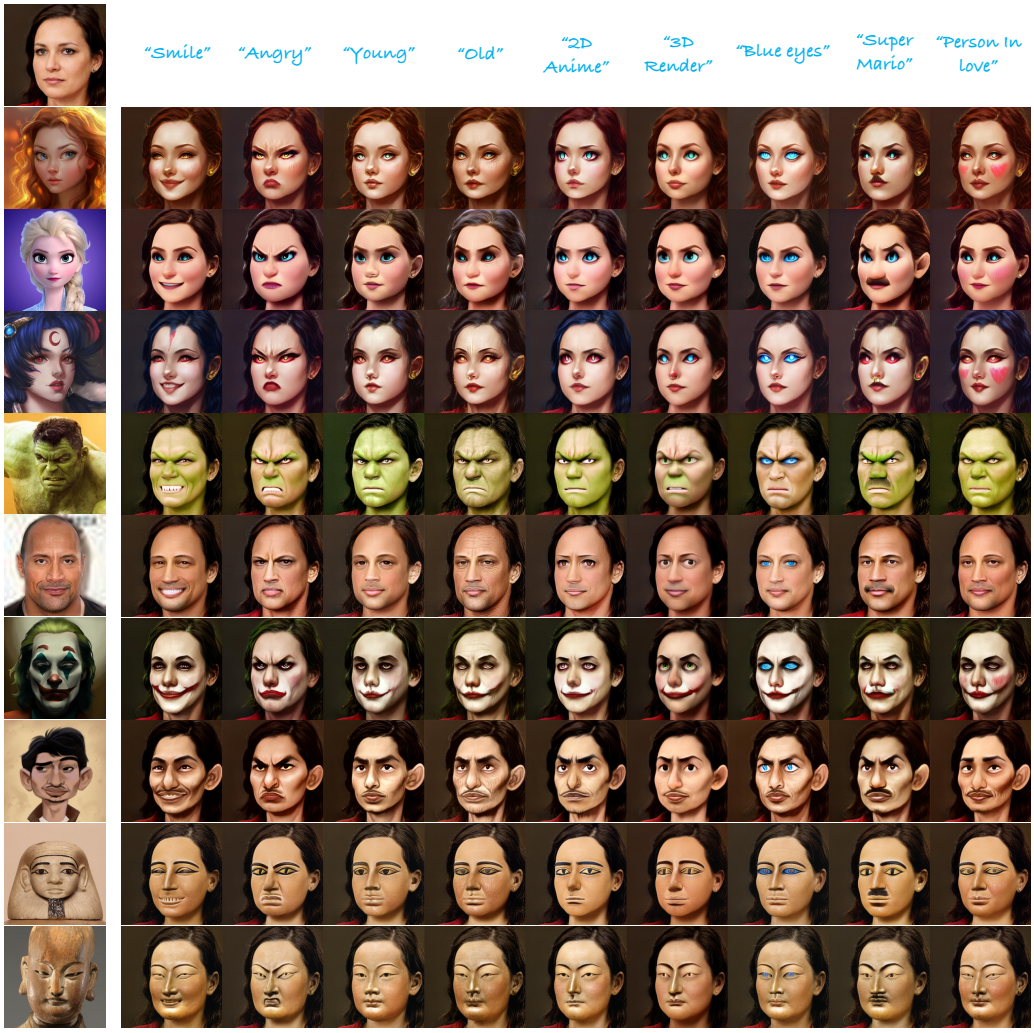

Figure 23: More results of image-text hybrid domain adaptation. The source image is in the top-left corner. The first row and column consist of training images and text prompts respectively.

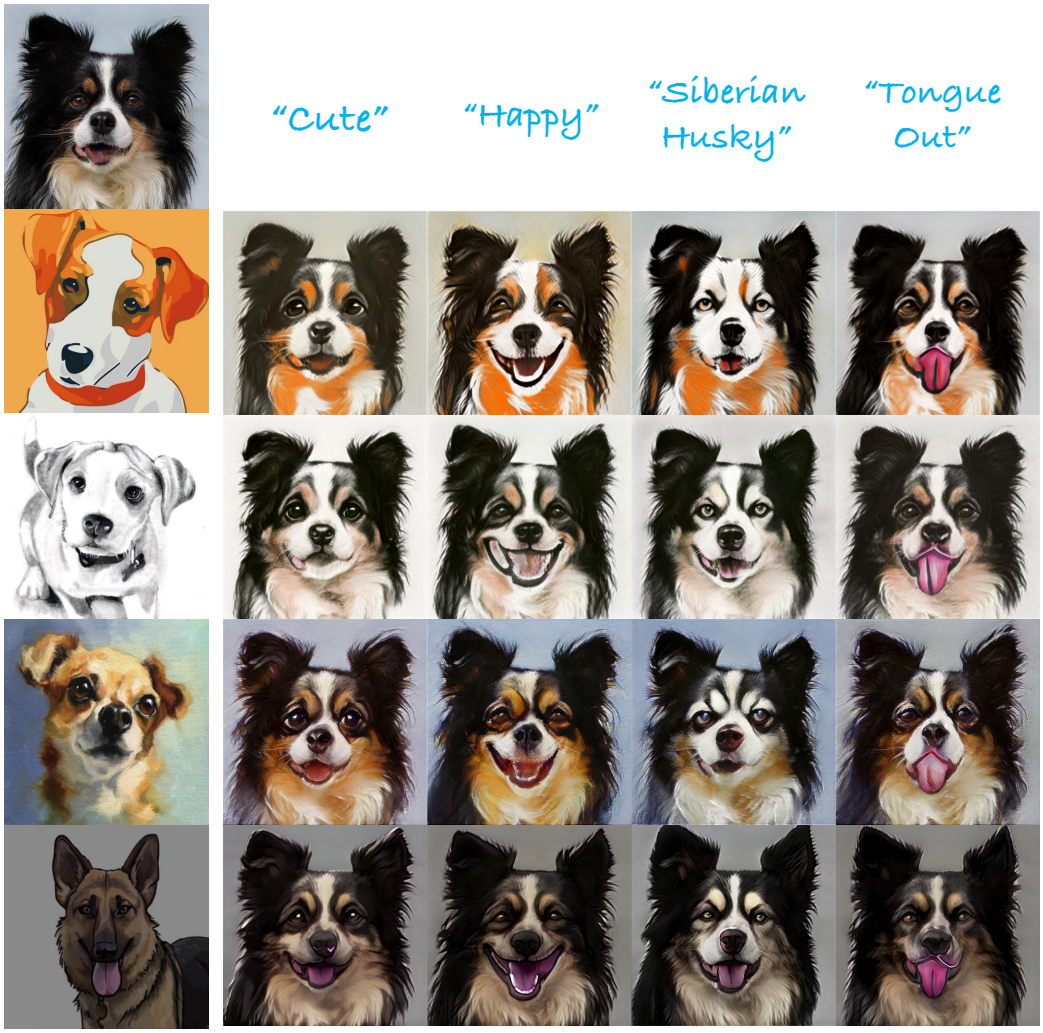

Figure 24: More results of image-text hybrid domain adaptation on AFHQ-Dog.

Figure 25: More results of image-text hybrid domain adaptation on LSUN-Church.

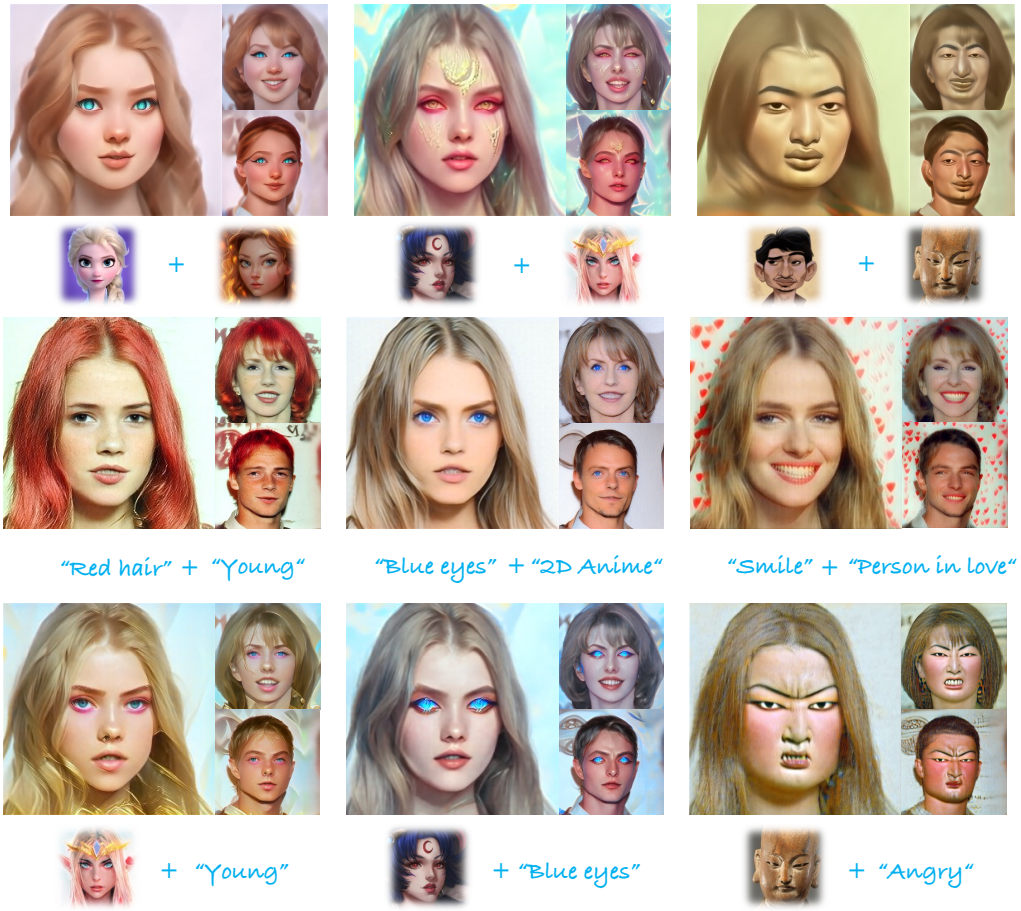

Figure 26: More results of UniHDA with DiffusionCLIP.

Figure 27: More results of UniHDA with EG3D.

Figure 28: Textual bias and ambiguity introduced by CLIP. We use 'Nurse' as the target domain and CLIP's learned biases manifest in the new domain, which converts the individuals to females.