# AIR: Rethinking Image-Text Offset Alignment in Multimodal Contrastive Representation Space

**Guimeng Liu** **Milad Abdollahzadeh** **Ngai-Man Cheung**

Temasek Laboratories, Singapore University of Technology and Design

## Abstract

Image-text offset alignment, which refers to the consistency between visual transformation directions and their corresponding textual transformation directions, has become an important relational property in multimodal contrastive representation spaces such as CLIP and SigLIP. This property plays a central role in text-driven visual transformation tasks, including generative model domain adaptation and text-guided object-centric image editing, where textual offsets are used to guide visual transformations. These methods fundamentally rely on the assumption that image and text offsets are well aligned, such that textual transformation directions can provide reliable guidance for corresponding visual transformations. However, the validity of this alignment assumption has not been systematically examined. In this work, we question this foundational assumption by conducting a comprehensive empirical analysis of image-text offset alignment in multimodal contrastive representation space. Our findings reveal not only noticeable offset misalignment but also a meaningful positive correlation between image-text offset misalignment and semantic concept distance across six large datasets and eight contrastive vision-language models. Based on this discovery, we propose Adaptation with Iterative Refinement (AIR), a method that iteratively refines text offsets through anchor sampling and our proposed concept description learning to mitigate the downstream effect of image-text offset misalignment and improve guidance accuracy. Comprehensive experiments on zero-shot generative model domain adaptation and text-guided object-centric image editing across 40 setups, including qualitative, quantitative, and user studies, consistently show that AIR enables state-of-the-art performance in these tasks. *Project Page:* https://guimeng-leo-liu.github.io/AIR-Rethinking-Image-Text-Offset-Alignment/

## 1 Introduction

Multimodal contrastive models, such as CLIP and SigLIP, learn a shared representation space by aligning paired image and text embeddings [1, 2]. Beyond this pairwise alignment, many downstream methods rely on a stronger relational property: for two semantic concepts in the shared representation space, the offset between their text embeddings is expected to align with the corresponding offset between their image embeddings. This consistency between textual and visual transformation directions is referred as *image-text offset alignment*. This property plays a central role in vision-language tasks that reason about semantic transformations, where text offsets serve as compact representations of semantic changes for guiding, retrieving, or evaluating visual transformations. Consequently, image-text offset alignment has been widely exploited in tasks including zero-shot generative model domain adaptation [3, 4, 5, 6], text-guided image editing [7, 8, 9, 10, 11], composed image retrieval [12, 13, 14], and direction-aware evaluation metrics [8, 15, 16]. Specifically, in zero-shot generative model domain adaptation (ZSGM) and text-guided object-centric image editing (TOIE), text offsets are used to guide corresponding image offsets toward desired visual

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

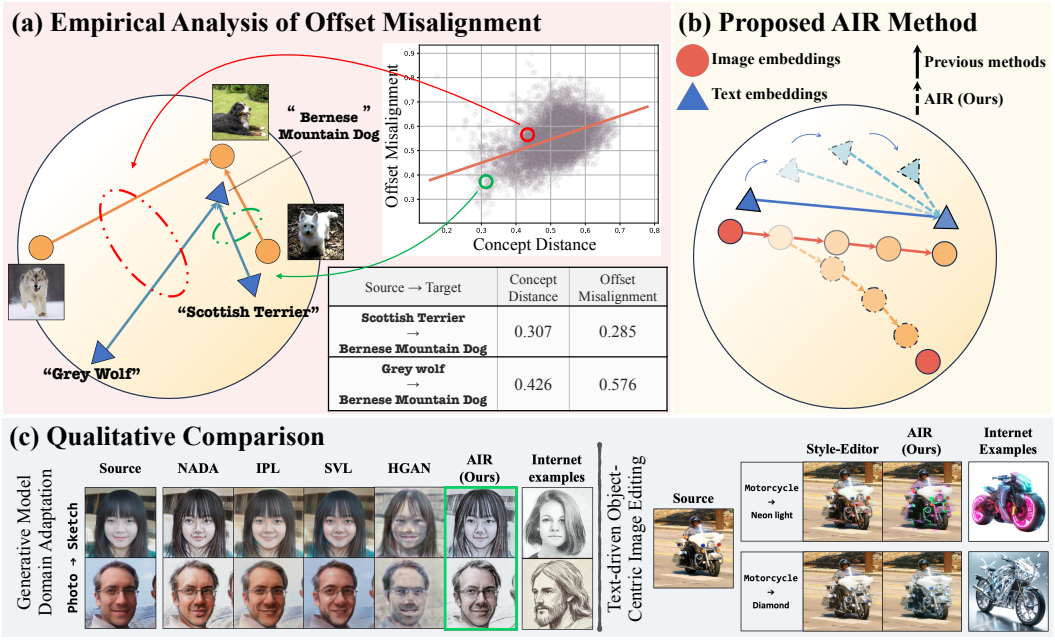



Figure 1: **Our contributions:** **(a)** We question the assumption of image-text offset alignment and perform a comprehensive analysis of offset alignment in multimodal contrastive representation space. **Our analysis reveals not only noticeable misalignment between image offsets (orange arrow) and text offsets (blue arrow), but also a meaningful positive correlation between offset misalignment and concept distance**. For example, in the ImageNet-1K dataset, the "Grey Wolf" is a more distant concept to the "Bernese Mountain Dog" (concept distance=0.426) than the "Scottish Terrier" (concept distance=0.307). Accordingly, "Grey Wolf" → "Bernese Mountain Dog" has higher offset misalignment than "Scottish Terrier" → "Bernese Mountain Dog" (0.576 vs 0.285). This misalignment is overlooked in existing uses of image-text offsets (Sec. 3). **(b)** Leveraging our discovery, we propose Adaptation with Iterative Refinement (AIR) to iteratively refine text offsets and mitigate the downstream effect of offset misalignment for text-driven visual transformations, which we refer as guidance misalignment (Sec. 4). **(c)** Our proposed AIR consistently reduces the guidance misalignment and enables SOTA performance in important text-guided generative modeling tasks, e.g., capturing rich target concepts' style and details across diverse setups. (see Sec. 5 and Sec. C.3 for detailed results).

transformations; in composed image retrieval, image and text offsets are composed to retrieve target images matching the intended semantic change; and in direction-aware evaluation metrics, the alignment between image and text offsets measures whether an observed visual change follows the textual instruction.

Despite their different formulations and objectives, these tasks fundamentally rely on the same underlying assumption: image and text offsets are sufficiently aligned in multimodal contrastive representation spaces, such that textual transformation directions can provide reliable guidance for corresponding visual transformations. *Critically, the validity of image-text offset alignment has not been systematically examined.*

In this work, we question this foundational assumption by conducting, to the best of our knowledge, the first systematic empirical analysis of image-text offset alignment within multimodal contrastive representation space. **Our primary findings** reveal not only noticeable offset misalignment, but also a meaningful positive correlation between image-text offset misalignment and semantic concept distance: for closely related concept pairs, misalignment tends to be smaller, whereas it increases as concepts become more distant (Fig. 1). This correlation is consistently observed across six large datasets and eight contrastive vision-language models.

Leveraging this discovery, we propose *Adaptation with Iterative Refinement (AIR)*, a framework designed to mitigate the downstream effect of image-text offset misalignment and enhance text-driven

visual transformations. We refer to the downstream effect of image-text offset misalignment in text-driven visual transformation as *guidance misalignment*. AIR iteratively samples intermediate transformed concepts as *anchors*, which lie closer to the target than the source in representation space (Sec. C.7), and refines text offsets with these anchors to reduce guidance misalignment and improve guidance accuracy. Since the textual descriptions of these anchor concepts are unknown, we introduce *concept description learning* to infer them dynamically. Our extensive results consistently demonstrate that AIR enables improved transformation quality and SOTA performance in two important text-guided generative modeling tasks: generative model domain adaptation and text-guided image editing. Our main contributions are summarized as follows:

- We question image-text offset alignment by conducting the first comprehensive empirical analysis on representative multimodal contrastive representation spaces. Our study reveals noticeable offset misalignment and a meaningful positive correlation with concept distance (Sec. 3).
- Based on our discovery, we introduce Adaptation with Iterative Refinement (AIR), a novel framework that leverages the discovered correlation to reduce guidance misalignment and improve transformation quality through iterative anchor sampling and a new concept description learning strategy to dynamically infer anchor's concept descriptions (Sec. 4).
- Our extensive experiments under two important tasks, ZSGM and TOIE, demonstrate consistent improvements in transformation quality, validated by qualitative, quantitative, and user study results, achieving state-of-the-art performance (Sec. 5 and Supp.).

**Remark:** Our discovery of image-text offset misalignment in multimodal contrastive space can be viewed as an analogy to text offset misalignment studies in unimodal text embedding space for natural language processing (NLP). In NLP, analogical reasoning [17, 18, 19, 20] relies on aligning offsets between word vectors, such as alignment between $E_v$("Man") - $E_v$("Woman"), and $E_v$("King") - $E_v$("Queen"), where $E_v$ denotes a text vector. Research indicates that the accuracy of analogical reasoning improves with similar, nearby concepts but decreases with growing distance [21, 22, 23, 24]. Similarly, our finding of a positive correlation between image-text offset misalignment and concept distance in multimodal contrastive model reveals a similar distance-dependent relationship in multimodal contrastive representation spaces.

## 2 Preliminaries: Image-Text Offset Alignment

**Image-Text Offset Alignment.** Let $E_I$ and $E_T$ denote the image and text encoders of a multimodal contrastive model, respectively. For two concepts $\alpha$ and $\beta$, with corresponding visuals $I_\alpha, I_\beta$ and textual descriptions $T_\alpha, T_\beta$, we define the image offset and text offset as: $\Delta I_{\alpha\to\beta} = E_I(I_\beta) - E_I(I_\alpha)$, and $\Delta T_{\alpha\to\beta} = E_T(T_\beta) - E_T(T_\alpha)$. Image-text offset alignment refers to the consistency between $\Delta I_{\alpha\to\beta}$ and $\Delta T_{\alpha\to\beta}$ in multimodal contrastive representation space. A common assumption is that the text offset between two concepts aligns with the corresponding image offset.

Image-text offset alignment is used in different vision-language tasks. In *text-driven visual transformation tasks* including zero-shot generative model adaptation and text-guided image editing, text offsets are used as optimization guidance for producing visual transformations. In composed image retrieval, image and text offsets are composed to retrieve target images matching the intended semantic change. In evaluation metrics, the alignment between image and text offsets is used to measure whether an observed visual change follows a textual instruction. Despite these different uses, these tasks share the same underlying assumption that image and text offsets are well aligned in multimodal contrastive representation space.

**Directional Loss.** Specifically, in text-driven visual transformation tasks, image-text offset alignment is typically operationalized through directional loss formulations. Given a source visual $I_\mathcal{S}$, a transformed visual $I_t$, and textual descriptions of the source and target concepts, denoted by $T_\mathcal{S}$ and $T_\mathcal{T}$, respectively, text-driven visual transformation methods compute a text offset from the source concept to the target concept and optimize the transformation so that the image offset aligns with this text offset:

$$\begin{aligned} \mathcal{L}_{direction} &= 1 - \cos(\Delta I_{\mathcal{S}\to t}, \Delta T_{\mathcal{S}\to\mathcal{T}}), \\ \text{where } \Delta I_{\mathcal{S}\to t} &= E_I(I_t) - E_I(I_\mathcal{S}), \\ \text{and } \Delta T_{\mathcal{S}\to\mathcal{T}} &= E_T(T_\mathcal{T}) - E_T(T_\mathcal{S}). \end{aligned} \tag{1}$$

Here, $\cos(x, y) = x \cdot y/(|x||y|)$ denotes cosine similarity. The directional loss is minimized when the offset between the source and transformed images is aligned with the offset between the source and target text embeddings. As such, directional loss fundamentally assumes that textual offsets faithfully capture the desired semantic concept transformation and can therefore provide reliable guidance for corresponding visual transformations.

## 3 A Closer Look at Image-Text Offset Misalignment

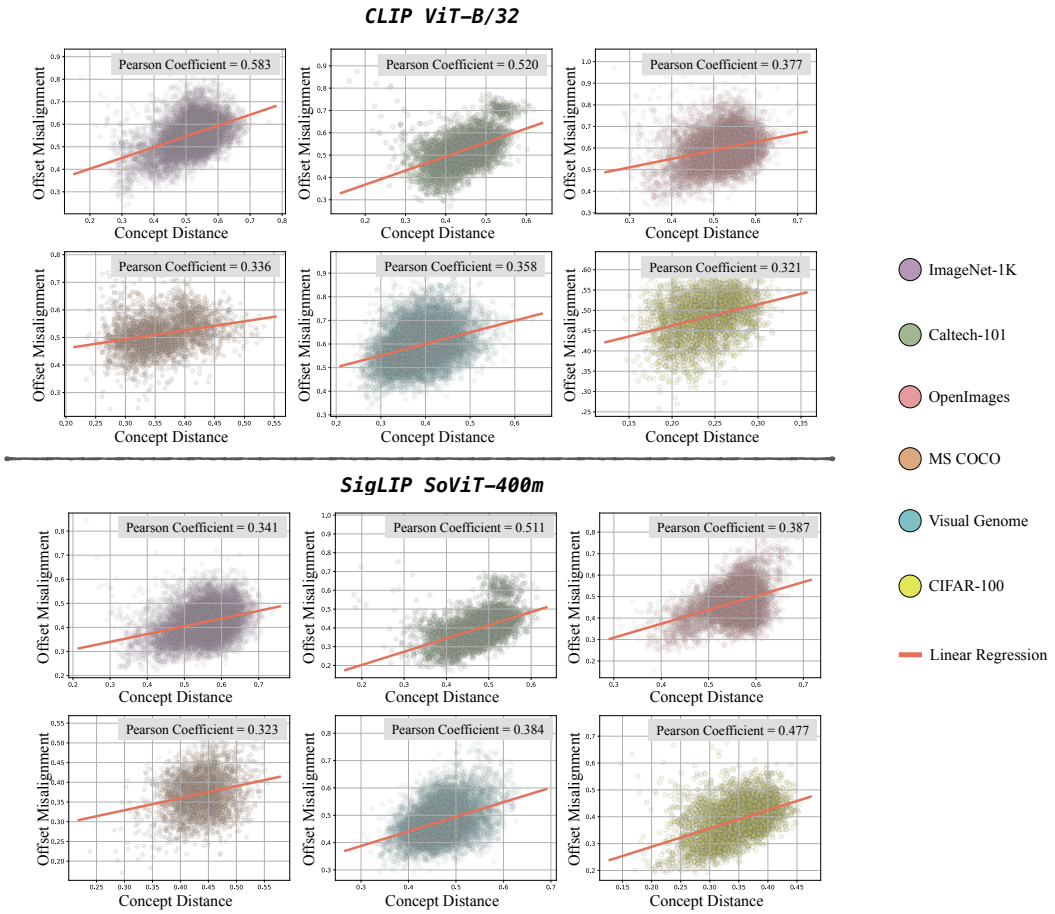



Figure 2: **Empirical analysis of offset misalignment.** We plot the offset misalignment (Eq. 2) vs concept distance in CLIP ViT-B/32 and SigLIP SoViT-400m for $N = 10,000$ text-image pairs sampled from 6 large publicly available datasets (details in Sec. A.1 ; total 120, 000 text-image pairs ). Our results show that there is a meaningful correlation (measured by Pearson coefficient [25]) between offset misalignment and concept distance for datasets with different distributions, i.e., close concepts has less offset misalignment. **Furthermore, we have consistent observations for other multimodal contrastive representation spaces, see Sec. A.2.**

Previous works assume that, for two semantic concepts $\alpha$ and $\beta$, the image offset $\Delta I_{\alpha\to\beta}$ and the corresponding text offset $\Delta T_{\alpha\to\beta}$ are well aligned in multimodal contrastive representation spaces (e.g., CLIP and SigLIP). This assumption serves as a foundational principle underlying several vision-language tasks discussed in Sec. 2, and is directly operationalized through directional loss formulations in text-driven visual transformation tasks. Despite its widespread adoption, we identify two fundamental limitations of this image-text offset alignment assumption:

- Multimodal contrastive models are trained with contrastive loss to maximize cosine similarity between corresponding image-text pairs, *i.e.,* maximize $\cos(E_I(I_\alpha), E_T(T_\alpha))$ for concept $\alpha$ (*e.g.*, cat), or maximize $\cos(E_I(I_\beta), E_T(T_\beta))$ for concept $\beta$ (*e.g.*, dog). Note that the degree of

alignment of image offset $\Delta I_{\alpha\to\beta}$ and text offset $\Delta T_{\alpha\to\beta}$ in multimodal contrastive space is not studied in the literature.

- In addition, this degree of alignment between $\Delta I_{\alpha\to\beta}$ and $\Delta T_{\alpha\to\beta}$ may vary based on the distance between two concepts pair $\alpha$ and $\beta$.

In this section, we take a closer look at this degree of offset alignment between two different modalities. Inspired by offset misalignment in NLP, we conduct an empirical study on large public datasets to analyze the offset misalignment between image and text modalities in multimodal contrastive space. Our analysis suggests that **there is misalignment between $\Delta I_{\alpha\to\beta}$ and $\Delta T_{\alpha\to\beta}$, and this misalignment tends to increase as concepts $\alpha$ and $\beta$ become more distant**.

### 3.1 Empirical Analysis of Image-Text Offset Misalignment

We conduct an empirical experiment on public datasets to evaluate the degree of alignment between image and text offsets. We randomly sample two classes for each dataset as a pair of concept $(\alpha, \beta)$. Then, the images within each class are used alongside the related concept textual description (*e.g.,* label) of each class to measure offset misalignment $\mathcal{M}(\alpha, \beta)$:

$$\begin{aligned} \mathcal{M}(\alpha,\beta) &= 1 - \cos(\Delta I_{\alpha\to\beta}, \Delta T_{\alpha\to\beta}), \\ \text{where } \Delta I_{\alpha\to\beta} &= E_I(\overline{I_\beta}) - E_I(\overline{I_\alpha}), \\ \text{and } \Delta T_{\alpha\to\beta} &= E_T(T_\beta) - E_T(T_\alpha) \end{aligned} \tag{2}$$

where $E_I(\overline{I_\alpha})$ is the average embedding of all images of the class (concept) $\alpha$ in multimodal contrastive space. Note we compute offset misalignment at the concept-distribution level rather than the instance level, since instance-specific visual variations may introduce semantic differences that are not described by the class-level text offset. To measure the distance between two concepts (concept distance) denoted by $\mathcal{D}(\alpha, \beta)$, we use cosine distance between images of two classes, *i.e.*, $\mathcal{D}(\alpha, \beta) = 1 - \cos(E_I(\overline{I_\beta}), E_I(\overline{I_\alpha}))$. We repeat this for all pairs of concepts in each dataset. Then, we plot $\mathcal{M}(\alpha, \beta)$ against $\mathcal{D}(\alpha, \beta)$ for each pair of concepts.

**Experimental Setup.** We conduct experiment across 8 multimodal contrastive models, including CLIP ViT-B/32 [1], CLIP RN50x64, CLIP ConvNext-L, SigLIP ViT-L/16-256 [2], SigLIP SoViT-400m, EVA-CLIP-8B [26], DFN5B-CLIP [27], and MobileCLIP-S2 [28]. We use 6 large and multi-class datasets that are publicly available, including ImageNet [29], Caltech-101 [30], OpenImages [31], COCO [32], Visual Genome [33], and CIFAR-100 [34] (details in Sec. A.1).

**Results.** Fig. 2 shows the offset misalignment against the concept distance in CLIP ViT-B/32 and SigLIP SoViT-400m for 6 public datasets. As shown in the plots, for all datasets, apart from their different distributions and characteristics, there is a positive correlation between offset misalignment and concept distance. Particularly, if two concepts $\alpha$ and $\beta$ are distant, there is a higher misalignment between image offset $\Delta I_{\alpha\to\beta}$ and corresponding text offset $\Delta T_{\alpha\to\beta}$. This means that given $I_\alpha$, $T_\alpha$ and $T_\beta$, it is sub-optimal to align $\Delta I_{\alpha\to\beta}$ and $\Delta T_{\alpha\to\beta}$ to find $I_\beta$. On the other hand, if two concepts $\alpha$ and $\beta$ are closer, potentially, it is more accurate to align $\Delta I_{\alpha\to\beta}$ and $\Delta T_{\alpha\to\beta}$ to find $I_\beta$. We have consistent observations for other multimodal contrastive representation spaces, see Sec. A.2.

> ***Finding:*** Offset misalignment between image and text modalities in multimodal contrastive representation space correlates positively with concept distance.

## 4 Methodology: Adaptation with Iterative Refinement

Our analysis in Sec. 3 suggests that closer concepts tend to have less offset misalignment in multimodal contrastive representation space. When applied to text-driven visual transformation, this offset misalignment manifests as a mismatch between the text-offset guidance and the desired visual transformation; we refer to this task-specific mismatch as guidance misalignment. Leveraging this finding, we propose *Adaptation with Iterative Refinement (AIR)*, aiming to reduce guidance misalignment by refining text-offset guidance over multiple shorter steps. AIR is particularly suitable for text-driven visual transformation tasks, such as zero-shot generative model adaptation and text-guided object-centric image editing, since these tasks are optimized iteratively under directional-loss

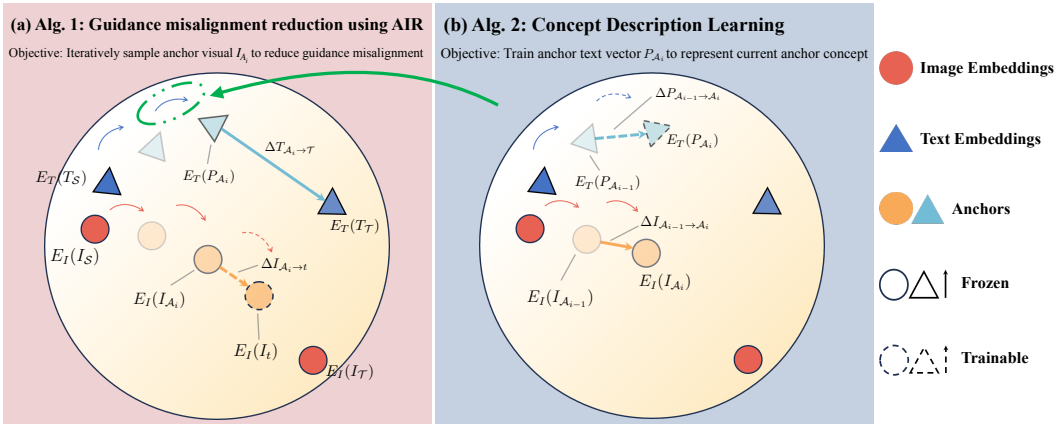



Figure 3: **Illustration of the proposed AIR method:** (a) Guidance misalignment reduction using Adaptation with Iterative Refinement (AIR) in text-driven visual transformation tasks (Sec. 4.1). (b) The proposed concept description learning method to learn anchor concept textual description $P_{\mathcal{A}_i}$ for the anchor $\mathcal{A}_i$ (Sec. 4.2). *Symbol descriptions*: $E_T$/$E_I$: multimodal contrastive text/image encoder; $I_\mathcal{S}$/$I_t$/$I_\mathcal{A}$: source/transformed/anchor visual; $T_\mathcal{S}$/$T_\mathcal{T}$/$P_\mathcal{A}$: source/target/anchor textual descriptions; $\Delta I$/$\Delta T$/$\Delta P$: image/text/anchor text offset.

guidance; reducing guidance misalignment at each step can therefore improve the overall transformation quality (see Sec. C.9).

Specifically, even though the concept distance between source $\mathcal{S}$ and target $\mathcal{T}$ is fixed, text-driven visual transformation are inherently iterative: the visual state progressively moves closer to the target concept over the course of optimization (see Sec. C.7). Therefore, AIR is based on the idea that intermediate transformed concept can be used as an **anchor** (denoted by $I_\mathcal{A}$). AIR computes directional guidance from this anchor to the target. This anchor is updated iteratively during the transformation as the transformed visual state moves closer to the target concept. Because the anchor-to-target concept distance is smaller than the original source-to-target distance, our analysis suggests that the corresponding guidance misalignment is smaller, leading to more accurate directional guidance and improved transformation quality (see Sec. C.9). One challenge of using $I_\mathcal{A}$ for directional loss is that the corresponding textual description $P_\mathcal{A}$ that describes this anchor is unknown. In what follows, we first introduce AIR in Sec. 4.1, and then the concept description learning to infer $P_\mathcal{A}$ in Sec. 4.2.

### 4.1 Adaptation with Iterative Refinement (AIR)

In text-driven visual transformation tasks, existing methods rely on a fixed text offset in Eq. 1 to guide the transformed visual state toward the target concept throughout optimization. In contrast, AIR adopts an iterative refinement strategy that progressively updates the textual guidance during optimization. Specifically, we first optimize the transformation for $t_{\text{thresh}}$ iterations using the current directional loss, allowing the transformed visual state to move closer to target concept. AIR then adaptively refines the textual offset to reduce guidance misalignment for subsequent transformation.

Then, in each $t_{int}$ interval of transformation, we sample the current transformed visual state as a new anchor.[1] For zero-shot GAN and diffusion adaptation, images are sampled from the current transformed generator and encoded by CLIP to obtain the anchor visual representations. For object-centric editing, the current transformed image is directly encoded by CLIP as the visual anchor. We denote the $i^{th}$ sampled anchor by $I_{\mathcal{A}_i}$. **To reduce guidance misalignment and provide more accurate guidance, AIR uses the anchor $\mathcal{A}_i$ instead of the original source $\mathcal{S}$ for computing directional loss.** The proposed AIR scheme is illustrated in Fig. 3 (a). The image offset with anchor $\mathcal{A}_i$ is computed between the current transformed visual $I_t$ and the anchor visual $I_{\mathcal{A}_i}$: $\Delta I_{\mathcal{A}_i \to t} = E_I(I_t) - E_I(I_{\mathcal{A}_i})$. Assuming that the anchor is described by the textual description $P_{\mathcal{A}_i}$ (details of inferring $P_{\mathcal{A}_i}$ are discussed in Sec. 4.2), the anchor-to-target text offset is calculated as: $\Delta T_{\mathcal{A}_i \to \mathcal{T}} = E_T(T_\mathcal{T}) - E_T(P_{\mathcal{A}_i})$. Finally, the adaptive loss $\mathcal{L}_{adaptive}$ is computed by aligning the image and text

[1] We use the same settings of $t_{thresh}$ and $t_{int}$ across all 32 experiment setups.

offsets from anchor $\mathcal{A}_i$ to target $\mathcal{T}$:

$$\mathcal{L}_{adaptive} = 1 - \cos(\Delta I_{\mathcal{A}_i \to t}, \Delta T_{\mathcal{A}_i \to \mathcal{T}}). \tag{3}$$

By replacing the original source-to-target offset with a closer anchor-to-target offset, $\mathcal{L}_{adaptive}$ reduces the expected guidance misalignment and provides more accurate guidance for the remaining transformation. We empirically find that adding this adaptive loss to $\mathcal{L}_{direction}$ results in more stable optimization. The pseudo-code can be found in Sec. B.1.

### 4.2 Learning Anchor Concept Descriptions

Here, we explain the details of the proposed method for learning textual description $P_{\mathcal{A}_i}$ that describes $i^{th}$ anchor concept $\mathcal{A}_i$. Inspired by [35, 36], we define anchor textual description $P_{\mathcal{A}_i} \in \mathbb{R}^{(M+1)\times d}$ as combination of $M$ learnable tokens $[V]^i_j \in \mathbb{R}^d$ and a label token $Y_{\mathcal{A}_i} \in \mathbb{R}^d$:

$$P_{\mathcal{A}_i} = [V]^i_1[V]^i_2 \dots [V]^i_M[Y_{\mathcal{A}_i}] \tag{4}$$

Following [37], we learn the anchor textual description $P_{\mathcal{A}_i}$ by aligning text offset to the corresponding image offset. Here, the image offset is calculated between the current and previous anchors: $\Delta I_{\mathcal{A}_{i-1}\to\mathcal{A}_i} = E_I(I_{\mathcal{A}_i}) - E_I(I_{\mathcal{A}_{i-1}})$. Similarly, the anchor text offset is calculated as follows: $\Delta P_{\mathcal{A}_{i-1}\to\mathcal{A}_i} = E_T(P_{\mathcal{A}_i}) - E_T(P_{\mathcal{A}_{i-1}})$. Note that the only trainable parameter is the unknown anchor textual description $P_{\mathcal{A}_i}$ which is learned by aligning image and anchor text offsets:

$$\mathcal{L}_{align} = 1 - \cos(\Delta I_{\mathcal{A}_{i-1}\to\mathcal{A}_i}, \Delta P_{\mathcal{A}_{i-1}\to\mathcal{A}_i}) \tag{5}$$

The proposed concept description learning approach is shown in Fig. 3 (b). We remark that $P_{\mathcal{A}_i}$ is the tokenized anchor textual description before the multimodal contrastive text encoder, and for simplicity, we slightly abuse the notation and use $E_T(P_{\mathcal{A}_i})$ to show multimodal contrastive text embedding for anchor $\mathcal{A}_i$.

**Remark:** Given that offset misalignment is less for closer concepts, we propose to use the previous anchor $\mathcal{A}_{i-1}$ as the source for learning the textual description of the current anchor $\mathcal{A}_i$. Since consecutive anchors are closer, the offset alignment used for learning $P_{\mathcal{A}_i}$ is expected to be more accurate.

**Regularizer:** We further propose to use the interpolation between tokenized source and target concept descriptions as anchor label, *i.e.*, $Y_{\mathcal{A}_i} = (1 - p_i)Y_{\mathcal{S}} + p_i Y_{\mathcal{T}}$, with $p_i$ denoting the proportion of the transformation progress until anchor $\mathcal{A}_i$. The label token acts as a regularizer during concept description learning (see motivation in Sec. C.5.2).

Our experimental results in Sec. 5.4 shows that these design choices lead to better transformation quality with AIR compared to directly learning anchor textual descriptions from the anchor visuals. More details of the method are summarized in the pseudo-code in Sec. B.1.

## 5 Experiments

In this section, we evaluate AIR on two text-driven visual transformation tasks: zero-shot generative model domain adaptation and text-guided image editing. We first test whether AIR reduces guidance misalignment where ground-truth visual offsets are available, and then evaluate whether refined guidance improves transformation quality. We conclude with ablations on AIR's design choices.

### 5.1 AIR on Zero-Shot Generative Model Adaptation

**Experimental setup.** We evaluate AIR on ZSGM in both GAN and diffusion settings. Following prior ZSGM methods [3, 4, 5, 6], we adapt StyleGAN2-ADA models pretrained on FFHQ, AFHQ-Dog, and Church, as well as Guided Diffusion models pretrained on FFHQ and AFHQ-Dog. We compare AIR with the SoTA ZSGM methods: NADA [3], IPL [4], SVL [5], and HGAN [6]. All methods use the same CLIP for guidance. Dataset, metric, hyperparameter, and implementation details are provided in Sec. C.

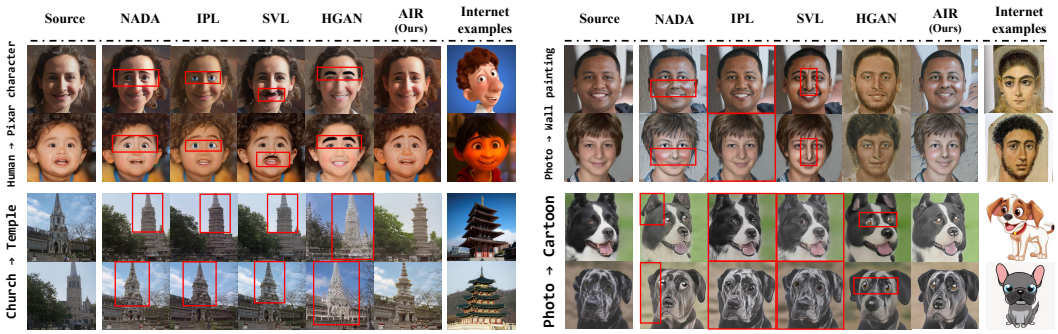



Figure 4: **Qualitative comparison for zero-shot generative model adaptation.** Degraded images or regions are highlighted with red boxes. NADA and HGAN often introduce undesirable changes, such as thick eyebrows in Human → Pixar character and red cheeks in Photo → Wall painting. IPL and SVL often under-adapt, yielding images that lack target-domain features or styles, particularly for transformations requiring substantial feature changes, such as Church → Temple and Photo → Cartoon. AIR avoids these artifacts and more faithfully captures target-domain styles, as illustrated by Pixar character, Wall painting, Temple, and Cartoon. Additional qualitative results are provided in Sec. C.3.1. (Best viewed in color and zoomed in.)

Table 1: AIR reduces guidance misalignment in downstream text-driven visual transformation task. We report guidance misalignment computed between adaptation directions and ground-truth visual offsets. IPL and SVL report multiple directions.

| Transformation Setting | Guidance Misalignment ↓ | | | |
|---|---|---|---|---|
| | NADA | IPL | SVL | AIR |
| Human → Baby | 0.68± 0.01 | 0.70± 0.07 | 0.92± 0.04 | **0.49**± 0.06 |
| Dog → Cat | 0.54± 0.01 | 0.66± 0.07 | 0.59± 0.03 | **0.26**± 0.11 |

### 5.1.1 AIR Reduces Guidance Misalignment

We first verify whether AIR reduces guidance misalignment during zero-shot adaptation. For each adaptation, we compare the directional guidance used by each method with a ground-truth visual offset computed from real target-domain images. For AIR, we measure the last anchor-to-target direction, matching its refined guidance formulation. Tab. 1 shows that AIR consistently produces lower offset misalignment, confirming that iterative anchor refinement provides more accurate guidance. Detailed computation is provided in Sec. C.

### 5.1.2 Improved Directional Guidance Leads to Better Adaptation

We next assess how reduced offset misalignment improves downstream task's performance, i.e., adaptation quality. Qualitative results in Figs. 1 and 4 show that NADA and HGAN often introduce undesirable feature changes, whereas IPL and SVL may under-adapt, yielding samples that lack target-domain features or styles. AIR avoids these common failures and more faithfully captures the target domain; see the figure captions for details.

We present quantitative results in Tab. 2 and Tab. 3. AIR consistently improves target-domain quality measured by CLIP Distance and FID, while maintaining competitive diversity measured by Intra-LPIPS. We additionally report comparisons of SigLIP and DINOv2 Distances in Sec. C.3.1, and human evaluation in Sec. C.8. The consistent improvements across CLIP, SigLIP, and DINOv2 Distances, FID, and human evaluation provide complementary evidence that AIR improves adaptation quality. See detailed metric definitions in Sec. C.2.

## 5.2 AIR on Text-Guided Object-Centric Image Editing

**Experimental setup.** We evaluate AIR on TOIE following the SoTA Style-Editor [11]. AIR refines Style-Editor's directional guidance using anchor-based text offsets while leaving the rest of the editing framework unchanged. We evaluate diverse source images and target text descriptions spanning

Table 2: Quantitative results of zero-shot GAN adaptation. For fair comparison, all methods use the same CLIP for guidance. FID is reported only for `Baby` and `Cat`, which have sufficient target-domain samples. AIR substantially improves adaptation quality (e.g., average CLIP Distance improvements ranging from 9% to 25%, and average FID improvements ranging from a minimum of 16%), while also maintaining competitive diversity measured by Intra-LPIPS. More qualitative results are shown in Fig. 1, Fig. 4, and Sec. C.3.1.

| Pre-trained Dataset | Adaptation | CLIP Distance (↓) | | | | | Intra-LPIPS (↑) | | | | | FID (↓) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NADA | IPL | SVL | HGAN | AIR | NADA | IPL | SVL | HGAN | AIR | NADA | IPL | SVL | HGAN | AIR |
| FFHQ | Human → Baby | 0.3327 | 0.3562 | 0.3838 | 0.3596 | **0.3325** | 0.4474 | 0.4518 | 0.4506 | 0.4110 | **0.4520** | 68.35 | 68.48 | 158.76 | 123.55 | **62.13** |
| | Human → Pixar | 0.2335 | 0.2343 | 0.4224 | 0.2418 | **0.2213** | **0.4759** | 0.4488 | 0.4618 | 0.4013 | 0.4717 | - | - | - | - | - |
| | Human → Werewolf | 0.3467 | 0.3200 | 0.3998 | 0.3424 | **0.2431** | 0.4301 | 0.4387 | 0.4316 | 0.4395 | **0.4410** | - | - | - | - | - |
| | Photo → Wall painting | 0.4382 | 0.4898 | 0.4952 | **0.3747** | 0.4306 | 0.4217 | 0.4320 | 0.4332 | 0.4208 | **0.4381** | - | - | - | - | - |
| | Photo → Sketch | 0.3606 | 0.3955 | 0.4092 | 0.3327 | **0.3126** | 0.4190 | 0.4292 | **0.4476** | 0.4354 | 0.4257 | - | - | - | - | - |
| | Photo → Watercolor | 0.3548 | 0.3621 | 0.3639 | 0.3865 | **0.3376** | 0.4598 | **0.4671** | 0.4612 | 0.4544 | 0.4656 | - | - | - | - | - |
| | Photo → Ukiyo-e | 0.2437 | 0.2467 | 0.3906 | 0.3286 | **0.2315** | 0.4583 | 0.4652 | 0.4406 | 0.4647 | **0.4670** | - | - | - | - | - |
| Dog | Dog → Cat | 0.1493 | 0.1530 | 0.1644 | 0.1642 | **0.1320** | 0.4439 | 0.4522 | 0.4547 | 0.4436 | **0.4628** | 70.87 | 83.29 | 65.79 | 88.71 | **56.20** |
| | Dog → Hamster | 0.1616 | 0.1457 | 0.1826 | 0.1282 | **0.1306** | 0.4196 | **0.4340** | 0.3918 | 0.3822 | 0.4213 | - | - | - | - | - |
| | Dog → Capybara | 0.1359 | 0.1446 | 0.1861 | 0.1543 | **0.1121** | 0.4312 | 0.4377 | 0.4264 | 0.4217 | **0.4401** | - | - | - | - | - |
| | Dog → Wolf | 0.1480 | 0.1519 | 0.2249 | **0.1199** | 0.1421 | 0.4305 | 0.4261 | 0.4272 | 0.4056 | **0.4349** | - | - | - | - | - |
| | Dog → the Joker | 0.3708 | 0.3827 | 0.4574 | 0.3230 | **0.3131** | 0.4155 | 0.4206 | **0.4327** | 0.4058 | 0.4173 | - | - | - | - | - |
| | Photo → Cartoon | 0.2433 | 0.2419 | 0.2543 | 0.2936 | **0.2258** | 0.4356 | 0.4413 | 0.4400 | 0.3741 | **0.4427** | - | - | - | - | - |
| | Photo → Watercolor | 0.1535 | 0.1711 | 0.1646 | 0.1611 | **0.1507** | 0.4639 | **0.4703** | 0.4622 | 0.4566 | 0.4667 | - | - | - | - | - |
| Church | Church → Skyscraper | 0.3860 | 0.3441 | 0.4209 | 0.3631 | **0.3270** | 0.4823 | 0.481 | 0.4777 | 0.4433 | **0.4839** | - | - | - | - | - |
| | Church → Temple | 0.3197 | 0.3632 | 0.3177 | 0.3229 | **0.3146** | 0.4689 | 0.4623 | 0.4839 | 0.5103 | **0.4862** | - | - | - | - | - |
| | Photo → Sketch | 0.2948 | 0.3070 | 0.3401 | 0.2957 | **0.2805** | 0.5084 | **0.5373** | 0.5356 | 0.5085 | 0.5176 | - | - | - | - | - |
| | Photo → Anime | 0.2067 | 0.2180 | 0.3503 | 0.3504 | **0.1990** | 0.5223 | **0.5392** | 0.5356 | 0.5071 | 0.5245 | - | - | - | - | - |
| | Avg. | 0.2711 | 0.2793 | 0.3293 | 0.2802 | **0.2465** | 0.4519 | 0.4575 | 0.4552 | 0.4359 | **0.4588** | 69.91 | 75.86 | 112.28 | 106.13 | **59.16** |

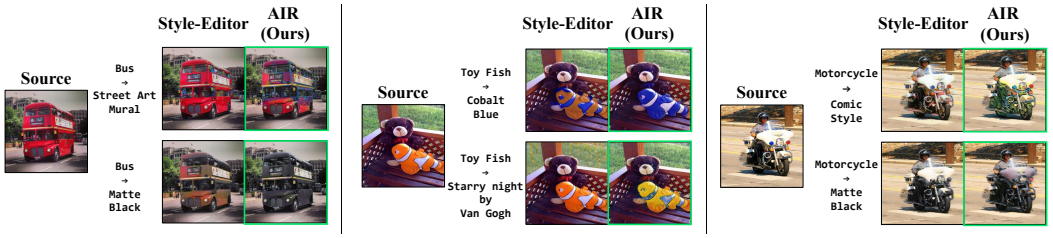



Figure 5: **Qualitative comparison for text-guided object-centric image editing.** Compared with Style-Editor, AIR better follows the target text while preserving image regions unrelated to the edit. These result indicates that AIR's refined directional guidance improves object-centric editing. **(Best viewed with color and zoom in.)**

a range of source objects and object-level transformations, including changes in color, material, and artistic style. See more details in Sec. C.

### 5.2.1 Improved Directional Guidance Leads to Better Editing

We evaluate whether AIR improves object-centric editing quality by providing more accurate directional guidance. As shown in Fig. 5, AIR follows the target text more faithfully than Style-Editor while preserving regions unrelated to the edited object.

Following Style-Editor, we report foreground metrics $\mathrm{Sim}_F$ and $\mathrm{Con}_F$, and background metrics $\mathrm{Con}_B$ and $\mathrm{Sty}_B$ (details in Sec. C.2). Tab. 4 shows that AIR improves foreground editing quality while maintaining comparable background preservation, indicating that refined guidance improves text-guided object-centric image editing.

## 5.3 Validate Learned Anchor Concept Descriptions

Additionally, to validate our concept description learning, we visualize the sampled anchor images and the corresponding learned anchor concept descriptions in CLIP space. As shown in Fig. 6, the anchor text descriptions learned by our proposed concept description learning accurately represent the anchor images (We show 3 of 5 anchors for presentation clarity).

## 5.4 Ablation Study

We conduct an ablation study on zero-shot generative model domain adaptation to evaluate the proposed concept description learning strategy. We compare: i) $\mathcal{I} \rightarrow \mathcal{T}$: We follow IPL to learn

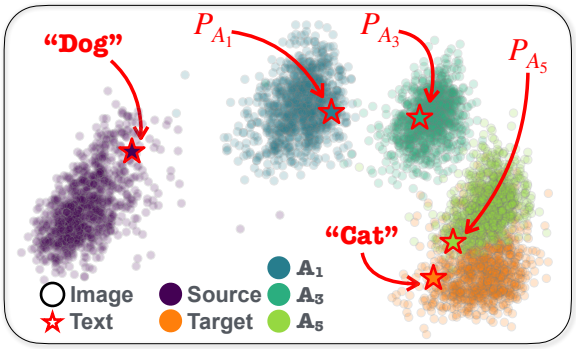



Figure 6: PCA visualization of anchor description learning. We use ZSGM setting of Dog → Cat as an example. For each anchor point $\mathcal{A}_i$, our learned concept description $P_{\mathcal{A}_i}$ lies within the distribution of 1000 sampled anchor images $I_{\mathcal{A}_i}$ for that anchor point. This shows that the learned descriptions produced by our concept description learning can accurately represent the anchor.

a mapper that produces a concept description for each image. ii) $\mathcal{S} \rightarrow \mathcal{A}_i$: We learn the concept description by capturing the semantic difference between $\mathcal{S}$ and $\mathcal{A}_i$ with the directional loss $\mathcal{L}^{\mathcal{S}}_{align} = 1 - \cos(\Delta I_{\mathcal{S}\rightarrow\mathcal{A}_i}, \Delta P_{\mathcal{S}\rightarrow\mathcal{A}_i})$. iii) $\mathcal{A}_{i-1} \rightarrow \mathcal{A}_i$: We use our proposed strategy, which captures the semantic difference between consecutive anchors $\mathcal{A}_{i-1}$ and $\mathcal{A}_i$ using the loss in Eq. 5. The results in Tab. 5 show that our concept description learning strategy produces more accurate textual description and improves zero-shot adaptation, supporting the use of shorter, less misaligned offsets for learning anchor descriptions. Our visual ablation results in Sec. C.5.3 further confirm this observation. See more of our comprehensive ablation studies in Sec. C.5.

Table 3: Quantitative results of zero-shot diffusion domain adaptation. AIR improves adaptation quality while maintaining competitive diversity. FID is reported only for Baby and Cat, which have sufficient target-domain samples. See more results in Sec. C.3.

| Pre-trained Dataset | Adaptation | CLIP Distance (↓) | | Intra-LPIPS (↑) | | FID (↓) | |
|---|---|---|---|---|---|---|---|
| | | NADA | AIR | NADA | AIR | NADA | AIR |
| FFHQ | Human → Baby | 0.2612 | **0.2172** | 0.5690 | **0.5763** | 63.68 | **58.74** |
| | Human → Werewolf | 0.2777 | **0.2323** | **0.5194** | 0.5185 | - | - |
| | Human → Pixar character | 0.4296 | **0.3865** | **0.4565** | 0.4558 | - | - |
| | Photo → Sketch | 0.4390 | **0.3601** | **0.4873** | 0.4815 | - | - |
| | Photo → Wall painting | 0.4781 | **0.4744** | 0.5252 | **0.5270** | - | - |
| | Photo → Watercolor | 0.3264 | **0.3229** | 0.6224 | **0.6399** | - | - |
| | Photo → Ukiyo-e | 0.2307 | **0.2152** | 0.5431 | **0.5524** | - | - |
| Dog | Dog → Cat | 0.1434 | **0.1428** | 0.5411 | **0.5494** | 84.75 | **77.15** |
| | Dog → Wolf | 0.1460 | **0.1337** | 0.4649 | **0.4735** | - | - |
| | Dog → Hamster | 0.1851 | **0.1588** | 0.4854 | **0.4952** | - | - |
| | Dog → Capybara | 0.1576 | **0.1204** | **0.4832** | 0.4604 | - | - |
| | Dog → the Joker | 0.3955 | **0.3208** | 0.4571 | **0.4663** | - | - |
| | Photo → Cartoon | 0.2516 | **0.2481** | 0.5572 | **0.5611** | - | - |
| | Photo → Watercolor | 0.1932 | **0.1853** | 0.5173 | **0.5322** | - | - |
| | Avg. | 0.2797 | **0.2513** | 0.5164 | **0.5207** | 74.22 | **67.95** |

Table 4: Quantitative results of text-guided object-centric image editing.

| Methods | Foreground quality metrics | | Background quality metrics | |
|---|---|---|---|---|
| | $\text{Sim}_F$ ↑ | $\text{Con}_F$ ↓ | $\text{Con}_B$ ↓ | $\text{Sty}_B$ ↓ |
| Style-Editor | 0.27 | 4.58 | **1.10** | **0.10** |
| AIR (ours) | **0.39** | **3.22** | 1.11 | **0.10** |

Table 5: Ablation study on concept description learning strategy.

| Methods | Human → Baby | | Dog → Cat | |
|---|---|---|---|---|
| | FID (↓) | Intra-LPIPS (↑) | FID (↓) | Intra-LPIPS (↑) |
| NADA | 68.35 | 0.4474 | 70.87 | 0.4439 |
| $\mathcal{I} \rightarrow \mathcal{T}$ | 98.35 | 0.4308 | 104.59 | 0.4452 |
| $\mathcal{S} \rightarrow \mathcal{A}_i$ | 64.39 | 0.4503 | 61.75 | **0.4630** |
| $\mathcal{A}_{i-1} \rightarrow \mathcal{A}_i$ | **62.13** | **0.4520** | **56.20** | 0.4628 |

## 6 Conclusion

In this work, we revisit the widely adopted assumption of image-text offset alignment in multimodal contrastive representation spaces. Through comprehensive empirical analysis, we reveal noticeable image-text offset misalignment and show that it has a meaningful positive correlation with semantic concept distance across multiple datasets and contrastive vision-language models. Motivated by this finding, we propose AIR, an iterative refinement framework that reduces the downstream effect of offset misalignment by replacing a single source-to-target text offset with refined anchor-to-target guidance during text-driven visual transformation. Extensive experiments on zero-shot generative model domain adaptation and text-guided image editing demonstrate that AIR improves directional guidance and consistently enhances transformation quality, achieving state-of-the-art performance in both tasks.

## Acknowledgment

This research is supported by Temasek Laboratories, Singapore University of Technology and Design; the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021); the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

## References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021. 1, 5, 19, 44

[2] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 5, 19, 25, 32

[3] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1, 7, 23, 24, 25, 38, 39, 43, 44

[4] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Humphrey Shi, Gao Huang, and Shiji Song. Zero-shot generative model adaptation via image-specific prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11494–11503, 2023. 1, 7, 23, 24, 43, 44

[5] Seogkyu Jeon, Bei Liu, Pilhyeon Lee, Kibeom Hong, Jianlong Fu, and Hyeran Byun. Improving diversity in zero-shot gan adaptation with semantic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7258–7267, 2023. 1, 7, 23, 24, 25, 43

[6] Abdul Basit Anees, Ahmet Canberk Baykal, Duygu Ceylan, Aykut Erdem, Erkut Erdem, and Muhammed Burak Kızıl. Hypergan-clip: A unified framework for domain adaptation, image synthesis and manipulation. In *Proceedings of the ACM (SIGGRAPH Asia)*, 2024. 1, 7, 42

[7] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2065–2074, 2021. 1, 43

[8] 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 (CVPR)*, pages 2426–2435, June 2022. 1, 41, 43

[9] Yeongtak Oh, Saehyung Lee, Uiwon Hwang, and Sungroh Yoon. On mitigating stability-plasticity dilemma in clip-guided image morphing via geodesic distillation loss. *International Journal of Computer Vision*, 133(5):2721–2751, 2025. 1

[10] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18062–18071, 2022. 1, 43

[11] Jihun Park, Jongmin Gim, Kyoungmin Lee, Seunghun Lee, and Sunghoon Im. Style-editor: Text-driven object-centric style editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18281–18291, June 2025. 1, 8, 24, 25, 32, 43

[12] Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4950–4958, 2022. 1, 43

[13] Zhenyu Yang, Shengsheng Qian, Dizhan Xue, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Semantic editing increment benefits zero-shot composed image retrieval. In *ACM Multimedia 2024*, 2024. 1

[14] Osman Tursun, Sinan Kalkan, Simon Denman, and Clinton Fookes. Pdv: Prompt directional vectors for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7740–7749, 2026. 1

[15] Qian Wang, Aleksandar Cvejić, Abdelrahman Eldesokey, and Peter Wonka. Editclip: Representation learning for image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15960–15970, 2025. 1, 43

[16] Yoonjeon Kim, Soohyun Ryu, Yeonsung Jung, Hyunkoo Lee, Joowon Kim, June Yong Yang, Jaeryong Hwang, and Eunho Yang. Preserve or modify? context-aware evaluation for balancing preservation and modification in text-guided image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23474–23483, 2025. 1, 43

[17] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013. 3, 41

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013*, 2013. 3, 41, 42

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3, 41, 42

[20] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, 2014. 3, 41

[21] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics*, 3:211–225, 2015. 3, 42

[22] Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th international conference on computational semantics*, pages 40–45, 2015. 3, 42

[23] Anna Rogers, Aleksandr Drozd, and Bofang Li. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148, 2017. 3, 42

[24] Louis Fournier, Emmanuel Dupoux, and Ewan Dunbar. Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 365–375, 2020. 3, 42

[25] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. 4, 20, 21, 22

[26] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 5, 19

[27] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations*, 2023. 5, 19

[28] Fartash Faghri, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, Alexander T Toshev, Oncel Tuzel, and Hadi Pouransari. Mobileclip2: Improving multi-modal reinforced training. *Transactions on Machine Learning Research*, 2025. 5, 19

[29] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. 5

[30] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding*, 106(1):59–70, 2007. 5

[31] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 5

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 25

[33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 5

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 5

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 7, 35

[36] Christopher TH Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai-man Cheung. Fairqueue: Rethinking prompt learning for fair text-to-image generation. *arXiv preprint arXiv:2410.18615*, 2024. 7

[37] Cheng Zhang, Xuanbai Chen, Siqi Chai, Henry Chen Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In *ICCV*, 2023. 7

[38] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 23, 39

[39] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410, 2019. 23, 27, 28, 31, 32, 33, 38, 39

[40] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8185–8194, 2020. 23, 24, 28, 29, 30, 31, 32, 34, 39, 44

[41] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 23, 32, 33, 34

[42] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11472–11481, 2022. 23

[43] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 24

[44] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 24

[45] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10743–10752, 2021. 24, 25, 42

[46] 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, 2022. 24

[47] 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*, pages 7380–7391, 2023. 24

[48] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 25

[49] 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. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 25, 32

[50] Leonard Kaufman and Peter J Rousseeuw. *Partitioning Around Medoids (Program PAM)*, chapter 2. John Wiley & Sons, 2009. 25

[51] 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*, abs/1506.03365, 2015. 26, 31

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 35

[53] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 40, 41

[54] 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. 40, 41, 44

[55] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, page 160–167, 2008. 41

[56] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394, 2010. 41

[57] Joachim Daiber, Lautaro Quiroz, Roger Wechsler, and Stella Frank. Splitting compounds by semantic analogy. In *Proceedings of the 1st Deep Machine Translation Workshop*, pages 20–28, 2015. 42

[58] Trevor Cohen, Dominic Widdows, and Thomas Rindflesch. Expansion-by-analogy: A vector symbolic approach to semantic search. In *Quantum Interaction*, pages 54–66, 2015. 42

[59] Nguyen Tuan Duc, Danushka Bollegala, and Mitsuru Ishizuka. Cross-language latent relational search between japanese and english languages using a web corpus. In *ACM Transactions on Asian Language Information Processing (TALIP)*, pages 54–66, 2015. 42

[60] Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2015. 42

[61] Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, 2016. 42

[62] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 42

[63] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 42

[64] 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*, pages 9140–9150, 2022. 42

[65] Yang Zhou, Zichong Chen, and Hui Huang. Deformable one-shot face stylization via dino semantic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7787–7796, 2024. 42

[66] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Domainstudio: Fine-tuning diffusion models for domain-driven image generation using limited data. *International Journal of Computer Vision*, pages 1–25, 2025. 42

[67] Jiayi Guo, Junhao Zhao, Chaoqun Du, Yulin Wang, Chunjiang Ge, Zanlin Ni, Shiji Song, Humphrey Shi, and Gao Huang. Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30503–30513, 2025. 42

[68] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18434–18443, 2025. 42

[69] Milad Abdollahzadeh, Guimeng Liu, Touba Malekzadeh, Christopher T.H Teo, Keshigeyan Chandrasegaran, and Ngai-Man Cheung. A survey on generative modeling with limited data, few shots, and zero shot. *Transactions on Machine Learning Research*, 2025. Survey Certification. 42

[70] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 43

[71] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020. 43

[72] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12863–12872, 2021. 43

[73] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3):1–21, 2021. 43

[74] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 43

[75] Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation. *arXiv preprint arXiv:2303.06285*, 2023. 43

[76] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 44

## Supplementary

In this supplementary material, we provide additional experiments, ablation studies, and reproducibility details to support our findings. These sections are not included in the main paper due to space constraints. In particular:

- Sec. A expands our offset misalignment study with additional multimodal contrastive models, for a total of **8 different multimodal contrastive models** considered in the whole paper. Across all multimodal contrastive models, we consistently observe a meaningful positive correlation between image-text offset misalignment and concept distance.
- Sec. C reports extensive experimental results on zero-shot generative model adaptation and text-guided object-centric image editing, bringing the total number of adaptation settings evaluated in this work to **40**. Across all settings, we consistently observe that our proposed AIR achieves state-of-the-art performance. These results support our claim that reducing image-text offset misalignment improves directional guidance and downstream transformation quality.

Codes are available in the supplementary files.

## Contents


**A More Details and Results of Empirical Analysis** **19**

A.1 Details of Empirical Analysis . . . . . 19

A.2 Full Empirical Analysis Results . . . . . 19

**B More Details of AIR** **23**

B.1 Algorithm . . . . . 23

**C More Details and Results of Experiments** **23**

C.1 Details of AIR Implementation . . . . . 23

C.2 Details of AIR Evaluation . . . . . 24

C.3 More Experimental Results of Downstream Tasks . . . . . 25

C.3.1 Zero-shot GAN Adaptation . . . . . 31

C.3.2 Zero-shot Diffusion Model Adaptation . . . . . 32

C.4 Hyperparameters of AIR . . . . . 32

C.4.1 Hyperparameters of AIR on ZSGM . . . . . 32

C.4.2 Hyperparameters of AIR on TOIE . . . . . 32

C.5 Additional Ablation Studies . . . . . 35

C.5.1 Ablation on Hyperparameters Selection . . . . . 35

C.5.2 Ablation on Anchor Label Initialization . . . . . 35

C.5.3 Visual Ablation Studies . . . . . 35

C.5.4 Ablation on Loss Weighting . . . . . 35

C.6 Zero-shot Adaptation with Other Multimodal Contrastive Models . . . . . 37

C.7 Concept Shifts during Transformation . . . . . 38

C.8 User Study . . . . . 38

C.9 Impact of Offset Misalignment on Downstream Tasks . . . . . 38

C.10 Additional Experiments of ZSGM . . . . . 39

C.10.1 Latent Space Interpolation . . . . . 39
C.10.2 Cross-model Interpolation . . . . . 40
C.10.3 Image Manipulation . . . . . 40

**D Related Work** **41**

**E Limitation** **43**

**F Social Impact** **43**

**G Licenses** **44**

# A More Details and Results of Empirical Analysis

## A.1 Details of Empirical Analysis

For datasets with a single class label for each image, such as ImageNet, Caltech-101, and CIFAR-100, we use the original images from the dataset. For datasets with multiple objects in an image, such as OpenImages, MS COCO, and Visual Genome, to better align with the setting in NADA, we extract the objects using bounding boxes and classify them into their labeled classes.

For a certain concept $\alpha$, we use the images of the class as $I_\alpha$. For text description $T_\alpha$, we use the corresponding class label with INt, e.g., "a photo of a [cat]" when $\alpha = \texttt{cat}$.

## A.2 Full Empirical Analysis Results

In this section, we present an additional empirical analysis of offset misalignment for **6 additional contrastive learning-based multimodal representation spaces**. Following the experimental setup in Sec. 3 and in addition to CLIP ViT-B/32 [1] and SigLIP SoViT-400m [2], we use CLIP ConvNext-L, CLIP RN50x64, SigLIP ViT-L/16-256 [2], EVA-CLIP-8B [26], DFN5B-CLIP [27], and MobileCLIP [28] as multimodal contrastive representation spaces. We plot offset misalignment against concept distance for six public datasets in Fig. 7–12. Our results demonstrate consistent positive correlations between offset misalignment and concept distance across different multimodal contrastive representation spaces.

One possible explanation for the consistent and meaningful positive correlation lies in the geometry of multimodal contrastive representation spaces. Contrastive models align individual image-text pairs but do not explicitly enforce relational consistency between concepts. Consequently, semantically close concepts, which often lie within locally smooth regions of the representation manifold, tend to yield more consistent transformation directions across modalities, whereas larger concept shifts span broader regions of the manifold where curvature and modality-specific distortions may lead to greater misalignment. Understanding the factors underlying this phenomenon is an interesting direction for future work.

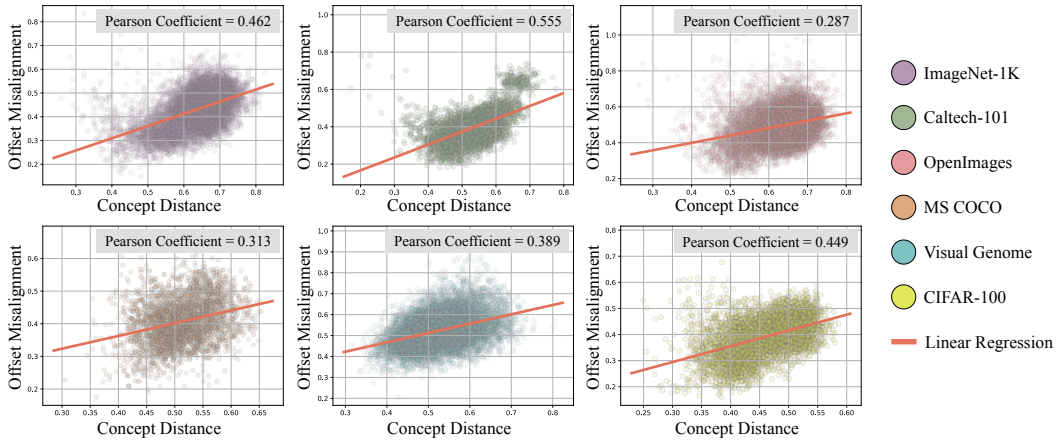



Figure 7: Empirical analysis of offset misalignment in CLIP ConvNext-L space. Experiment setup is the same as Sec. 3.1 except that CLIP ConvNext-L is used as the multimodal contrastive representation space. Our results show that the **meaningful correlation** (measured by Pearson coefficient [25]) between offset misalignment and concept distance consistently exists in various multimodal contrastive representation spaces.

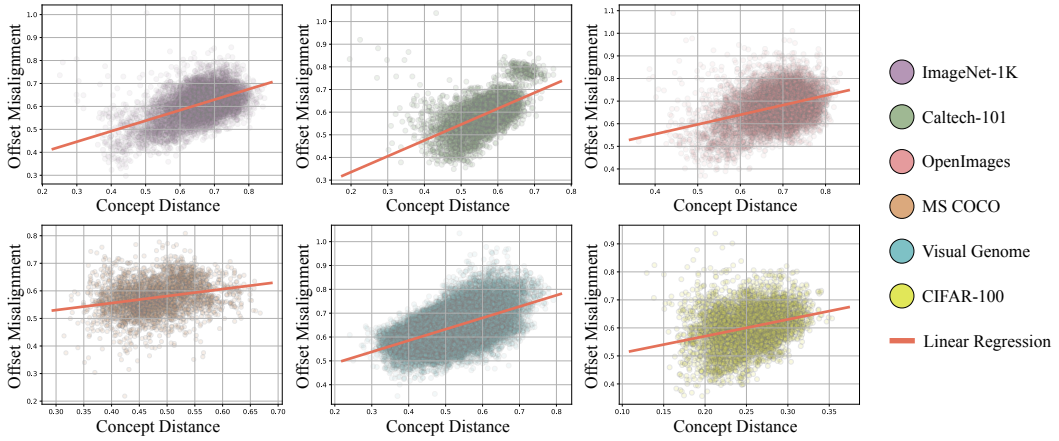



Figure 8: Empirical analysis of offset misalignment in CLIP RN50x64 space. Experiment setup is the same as Sec. 3.1 except that CLIP RN50x64 is used as the multimodal contrastive representation space. Our results show that the **meaningful correlation** (measured by Pearson coefficient [25]) between offset misalignment and concept distance consistently exists in various multimodal contrastive representation spaces.

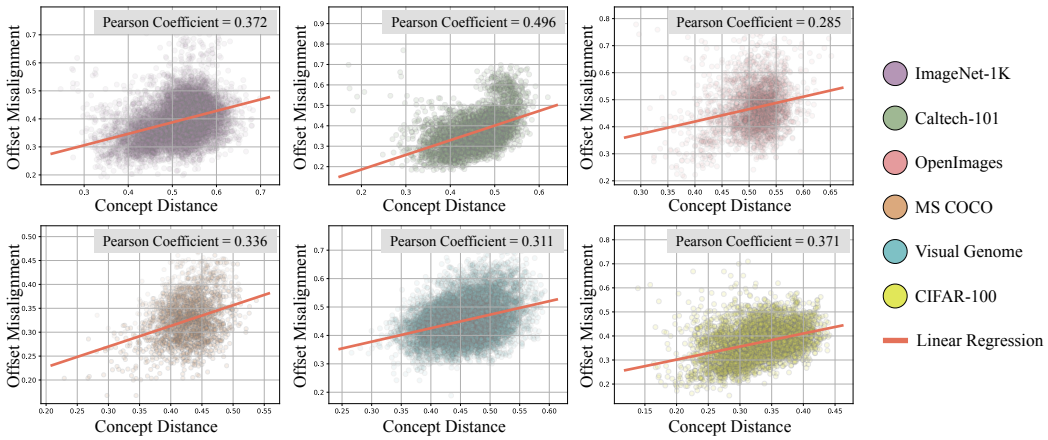



Figure 9: Empirical analysis of offset misalignment in SigLIP ViT-L/16-256 space. Experiment setup is the same as Sec. 3.1 except that SigLIP ViT-L/16-256 is used as the multimodal contrastive representation space. Our results show that the **meaningful correlation** (measured by Pearson coefficient [25]) between offset misalignment and concept distance consistently exists in various multimodal contrastive representation spaces.

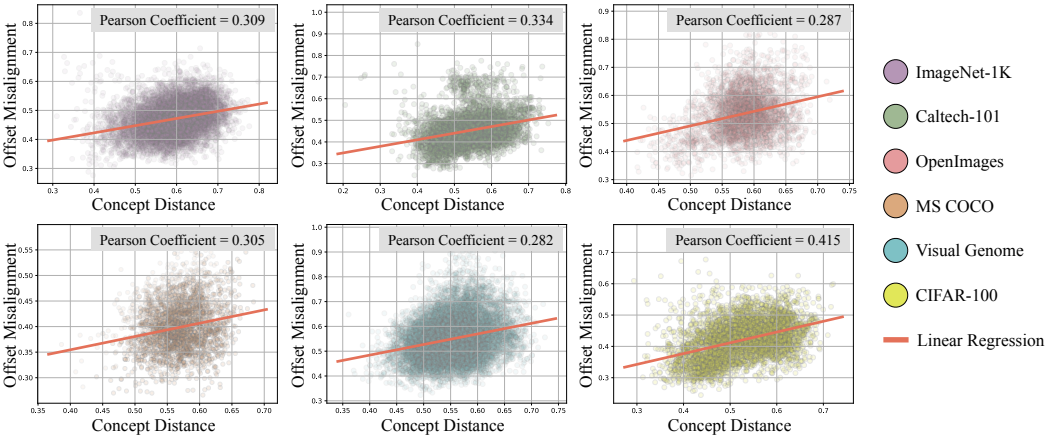



Figure 10: Empirical analysis of offset misalignment in EVA-CLIP-8B space. Experiment setup is the same as Sec. 3.1 except that EVA-CLIP-8B is used as the multimodal contrastive representation space. Our results show that the **meaningful correlation** (measured by Pearson coefficient [25]) between offset misalignment and concept distance consistently exists in various multimodal contrastive representation spaces.

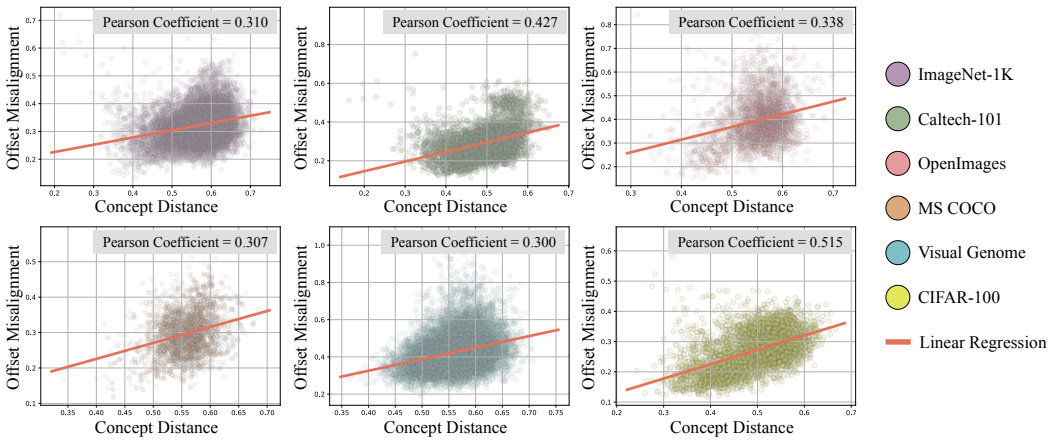



Figure 11: Empirical analysis of offset misalignment in DFN5B-CLIP space. Experiment setup is the same as Sec. 3.1 except that DFN5B-CLIP is used as the multimodal contrastive representation space. Our results show that the **meaningful correlation** (measured by Pearson coefficient [25]) between offset misalignment and concept distance consistently exists in various multimodal contrastive representation spaces.

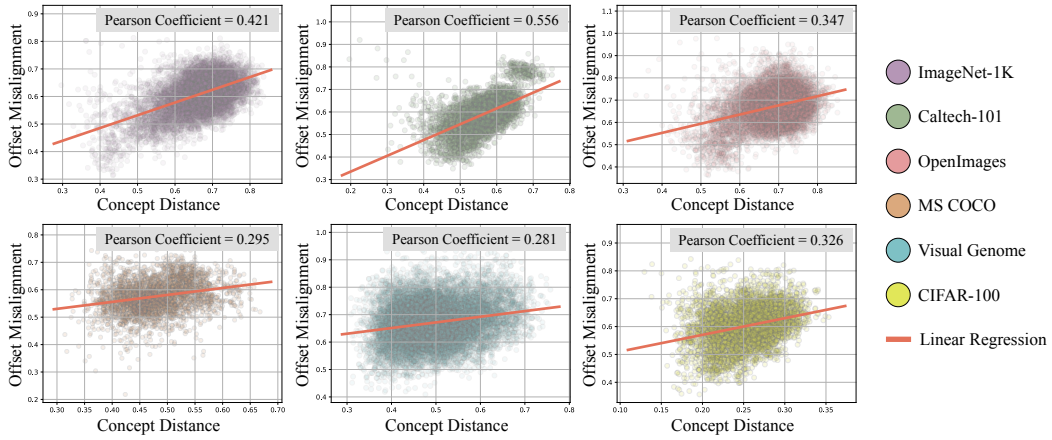



Figure 12: Empirical analysis of offset misalignment in MobileCLIP space. Experiment setup is the same as Sec. 3.1 except that MobileCLIP is used as the multimodal contrastive representation space. Our results show that the **meaningful correlation** (measured by Pearson coefficient [25]) between offset misalignment and concept distance consistently exists in various multimodal contrastive representation spaces.

# B More Details of AIR

## B.1 Algorithm

We provide a figure illustration corresponding pseudo-code for applying AIR to text-driven visual transformation in Fig. 3(a) and Alg. 1, and for the proposed concept description learning module in Fig. 3(b) and Alg. 2. We use $I_{\mathcal{S}}(z)$ and $I_t(z;\theta_t)$ to denote the source visual state and the transformed visual state at iteration $t$, respectively, where $\theta_t$ denotes the task-specific trainable parameters. For zero-shot generative model domain adaptation, the visual states correspond to $G_{\mathcal{S}}(w)$ and $G_t(w)$ generated using the same latent code $w$. In the GAN setting, $\theta_0$ is initialized from the pretrained source generator. In the diffusion setting, the pretrained denoiser remains frozen and $\theta_t$ denotes the trainable LoRA adapter parameters. For text-guided object-centric image editing, the visual states correspond to the source and current edited images, $\theta_t$ denotes the StyleNet parameters optimized for the given source image following Style-Editor, and $z$ can be omitted.

We initialize the anchor sequence with the source concept, *i.e.*, $I_{\mathcal{A}_0}(z) = I_{\mathcal{S}}(z)$ and $P_{\mathcal{A}_0} = T_{\mathcal{S}}$. Thus, the first intermediate anchor is learned relative to the original source, while each subsequent anchor is learned relative to its preceding anchor.

**Algorithm 1:** AIR for Text-Driven Visual Transformation

**Require:** Source visual state $I_{\mathcal{S}}(z)$, task-specific transformation producing $I_t(z;\theta_t)$, textual descriptions $T_{\mathcal{S}}$ and $T_{\mathcal{T}}$, $t_{adapt}$, $t_{thresh}$, $t_{int}$, learning rate $\eta$, and multimodal contrastive image and text encoders $E_I$ and $E_T$

**Output:** Optimized transformed visual state $I_{t_{adapt}}(z;\theta_t)$

Initialize $\theta_0$ following the task-specific baseline and keep $I_{\mathcal{S}}$ fixed, $i = 0$, $\mathcal{L}_{adaptive} = 0$
$\Delta T_{\mathcal{S}\to\mathcal{T}} = E_T(T_{\mathcal{T}}) - E_T(T_{\mathcal{S}})$
**for** $t = 0$; $t{+}{+}$; $t < t_{adapt}$ **do**
  $\Delta I_{\mathcal{S}\to t} = E_I(I_t(z;\theta_t)) - E_I(I_{\mathcal{S}}(z))$
  $\mathcal{L}_{direction} = 1 - \cos(\Delta I_{\mathcal{S}\to t}, \Delta T_{\mathcal{S}\to\mathcal{T}})$
  **if** $t \% t_{int} = 0$ **then**
    $i{+}{+}$
    $I_{\mathcal{A}_i}(z) = I_t(z;\theta_t)$
    $P_{\mathcal{A}_i}$ = Concept-Description-Learning $(I_{\mathcal{A}_i}, I_{\mathcal{A}_{i-1}}, P_{\mathcal{A}_{i-1}})$ /* refer to Alg. 2 for details */
  **end**
  **if** $t > t_{thresh}$ **then**
    $\Delta I_{\mathcal{A}_i\to t} = E_I(I_t(z;\theta_t)) - E_I(I_{\mathcal{A}_i}(z))$ /* if $I_t = I_{\mathcal{A}_i}$, apply the task-specific perturbation to $I_t$ */
    $\Delta T_{\mathcal{A}_i\to\mathcal{T}} = E_T(T_{\mathcal{T}}) - E_T(P_{\mathcal{A}_i})$
    $\mathcal{L}_{adaptive} = 1 - \cos(\Delta I_{\mathcal{A}_i\to t}, \Delta T_{\mathcal{A}_i\to\mathcal{T}})$
  **end**
  $\mathcal{L} = \mathcal{L}_{baseline} + \mathcal{L}_{adaptive}$
  Update $\theta_{t+1} \leftarrow \theta_t - \eta\nabla_{\theta_t}\mathcal{L}$
**end**

# C More Details and Results of Experiments

## C.1 Details of AIR Implementation

In this work, we implement zero-shot generative model adaptation for both GANs and diffusion models. The implementation details for each type of model is as follows:

- **Zero-Shot Adaptation of GANs.** We follow previous ZSGM works [3, 4, 5] setups to adapt StyleGAN2-ADA [38] pre-trained on FFHQ [39] and AFHQ-Dog [40] to various target domains.
- **Zero-Shot Adaptation of Diffusion Models.** We use Guided Diffusion [41] pre-trained on FFHQ and AFHQ-Dog from P2-Weighting [42] as source generator. To speedup training, we

**Algorithm 2:** Concept Description Learning for Visual Anchors

**Require:** Current and previous anchor visual states $I_{\mathcal{A}_i}$ and $I_{\mathcal{A}_{i-1}}$, learned concept description for the previous anchor $P_{\mathcal{A}_{i-1}}$, learning rate $\mu$, and multimodal contrastive image and text encoders $E_I$ and $E_T$

**Output:** Textual description $P_{\mathcal{A}_i}$ to represent the current anchor concept.

$\Delta I_{\mathcal{A}_{i-1}\to\mathcal{A}_i} = E_I(I_{\mathcal{A}_i}(z)) - E_I(I_{\mathcal{A}_{i-1}}(z))$

**for** $k = 0$; $k{+}{+}$; $k < k_{iter}$ **do**

  $\Delta P_{\mathcal{A}_{i-1}\to\mathcal{A}_i} = E_T(P_{\mathcal{A}_i}) - E_T(P_{\mathcal{A}_{i-1}})$.

  $\mathcal{L}_{align} = 1 - \cos(\Delta I_{\mathcal{A}_{i-1}\to\mathcal{A}_i}, \Delta P_{\mathcal{A}_{i-1}\to\mathcal{A}_i})$

  Update $P_{\mathcal{A}_i} \leftarrow P_{\mathcal{A}_i} - \mu\nabla_{P_{\mathcal{A}_i}}\mathcal{L}_{align}$

**end**

use DPM-Solver [43] for 10-step image generation. To prevent overfitting, we keep all pretrained denoiser weights frozen and optimize only the LoRA adapter parameters [44].

During the adaptation of both generators, we utilize the pre-trained ViT-B/32 as vision encoder for CLIP. Hyperparameter details can be found in Sec. C.4. *Notably, the only varying hyperparameter for all adaptation setups is the total number of adaptation iterations (strictly follow NADA).*

Note that none of the four baselines NADA/IPL/SVL/HyperGAN-CLIP of ZSGM provides an official implementation for diffusion models. We therefore implemented NADA for the diffusion setting on generative domain adaptation, while the other three methods cannot be directly transferred for the diffusion setting without substantial methodological redesign. This is because:

- IPL is not generator-agnostic, as it learns a mapper from StyleGAN style codes and therefore relies on the generator architecture and latent-space structure of StyleGAN.
- HyperGAN-CLIP trains a hypernetwork to modulate StyleGAN style layers, making its formulation specific to StyleGAN-based generators.
- SVL also relies on a parameter-importance estimation which is designed for a one-pass GAN generator. In diffusion models, however, the same denoising network is repeatedly applied across different noise levels, and the importance of a parameter may vary substantially across timesteps. This behavior cannot be captured by directly applying SVL's output-image-based importance estimation.

Consequently, extending IPL, SVL, or HyperGAN-CLIP to diffusion models would require nontrivial new designs beyond a straightforward reimplementation.

For implementation of text-guided object-centric image editing, we follow the SoTA Style-Editor [11]. AIR is applied by refining Style-Editor's directional guidance with anchor-based text offsets while keeping the remaining editing framework unchanged

### C.2 Details of AIR Evaluation

**Zero-Shot Generative Model Adaptation** A well-trained image generator is defined by its ability to produce high-quality and diverse images from target distribution. We follow existing zero-shot works in evaluation setup when applicable, and further improve on them. Specifically, following previous works [3, 4, 5], we have conducted comparisons on both public datasets and images collected from the internet. Our evaluations include both visual inspections for qualitative evaluations and quantitative evaluations using the following metrics:

- **FID.** For target domains with large and publicly available datasets, we follow previous work [5] to use FFHQ-Baby [45] (for target domain `Baby`), and AFHQ-Cat [40] (for target domain `Cat`) as target distribution. Then, we generate 5000 samples for each target domain [46, 47], and use FID to evaluate the quality and diversity of domain adaptation.
- **CLIP Distance.** The public data is scarce for other target domains, *e.g.,* `Pixar`. For these domains, we follow IPL's idea [4] to collect internet images as reference. However, since IPL did not make the collected samples publicly available, we had to repeat the same practice and

collect the images. Then, we use the CLIP Distance [48] which is defined as the cosine distance between the CLIP embeddings of the collected samples and the generated samples to measure the similarity of the generated samples to the target domain. Note that these collected target images are never exposed during adaptation, so the evaluation does not introduce target-image leakage into the optimization. Since no text is involved in this evaluation, the metric cannot be improved through trivial alignment with the target text.

- **SigLIP/DINO Distance.** We further evaluate the quality of the domain adaptation by introducing two additional metrics SigLIP Distance [2] and DINOv2 Distance [49]. Similar to the computation of CLIP Distance, SigLIP and DINOv2 Distance are defined as the cosine distance between the SigLIP/DINOv2 embeddings of collected and generated samples. SigLIP provides evaluation in a different multimodal representation space, while DINOv2 Distance is a fully image-based metric that does not involve text embeddings.
- **Intra-LPIPS.** To measure the diversity of the domain adaptation, we use Intra-LPIPS metric [45] which first assigns generated samples to one of $K$ clusters, then averages pair-wise distance within the cluster members and reports the average value over $K$ clusters. In zero-shot setup, since there are no training images, we follow [3, 5] to cluster around generated samples using $K$-Medoids [50], with $K = 10$.
- **User Study.** We also conducted a user study to compare the quality and the diversity of the domain adaptation with different schemes based on human feedback. See more details in Sec. C.8.

We remark that similarly NADA reports Intra-LPIPS on AFHQ-Cat, and SVL reports both FID and Intra-LPIPS on AFHQ-Cat.

**Text-Guided Object-Centric Image Editing** Following Style-Editor [11], we evaluate the edited foreground and the preserved background separately using the ground-truth object masks from MS COCO [32]. Let $I^{\text{src}}$ and $I^{\text{out}}$ denote the source and edited images, respectively, $T^{\text{tgt}}$ the target text, $M_F$ the foreground mask, and $M_B = 1 - M_F$ the corresponding background mask. We define $I_F^x = \text{crop}(I^x \odot M_F)$ and $I_B^x = I^x \odot M_B$, where $x \in \{\text{src}, \text{out}\}$. The masks are used only for evaluation and are not provided to either editing method.

- **Foreground text similarity** ($\text{Sim}_F$)**.** This metric measures how well the edited foreground matches the target text using the cosine similarity between their CLIP embeddings: $\text{Sim}_F = \cos(E_T(T^{\text{tgt}}), E_I(I_F^{\text{out}}))$ . A higher value indicates stronger alignment between the target description and the edited object.
- **Foreground content distance** ($\text{Con}_F$)**.** We compute the VGG content loss between the source and edited foregrounds, $\text{Con}_F = \mathcal{L}_{\text{content}}^{\text{VGG}}(I_F^{\text{src}}, I_F^{\text{out}})$. A lower value indicates better preservation of the object's semantic content and structure during stylization.
- **Background content distance** ($\text{Con}_B$)**.** We apply the same VGG content loss to the source and edited backgrounds, $\text{Con}_B = \mathcal{L}_{\text{content}}^{\text{VGG}}(I_B^{\text{src}}, I_B^{\text{out}})$. A lower value indicates less structural or semantic change outside the edited object.
- **Background style distance** ($\text{Sty}_B$)**.** This metric is the mean–variance VGG style loss between the source and edited backgrounds, $\text{Sty}_B = \mathcal{L}_{\text{style}}^{\text{VGG}}(I_B^{\text{src}}, I_B^{\text{out}})$. A lower value indicates that the original background appearance is better preserved and that less target style leaks into irrelevant regions.

Other perceptual metrics such as LPIPS are not directly suitable for our tasks, because they generally assume that the two images depict the same underlying content and are approximately spatially aligned. They are therefore most meaningful for measuring perceptual distortion or reconstruction quality between paired images. In our settings, however, no paired ground-truth target image is available for either zero-shot generative model adaptation or text-guided object-centric image editing. In addition, we believe the included visual results can help in transparency and reflecting the superior performance of our proposed method in terms of adaptation quality.

### C.3 More Experimental Results of Downstream Tasks

We include a total of **32** different configurations of zero-shot adaptation in this paper. The experimental setting and evaluation metric follow Sec. 5 in the main paper.

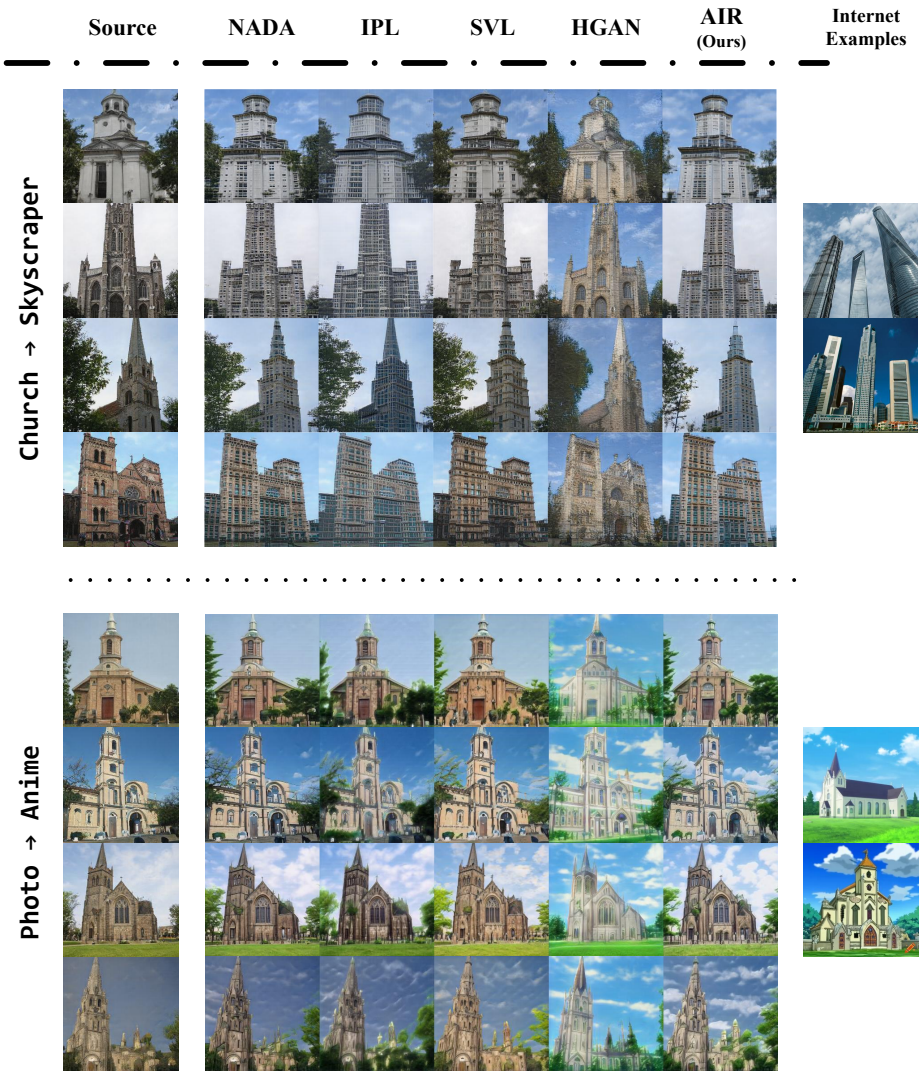



Figure 13: **Additional zero-shot adaptation results from source domain** `Church`**.** Here we use a StyleGAN2 generator pre-trained on the LSUN-Church [51] dataset as $G_{\mathcal{S}}$ and shift this to various target domains using different zero-shot approaches. We report the qualitative results for two setups: `Church`→`Skyscraper` and `Photo`→`Anime`. We also compute CLIP Distance on 5K generated samples as quantitative results, as shown in Tab. 2, our proposed AIR approach results in less CLIP distance meaning that the adapted generator is closer to the target domain. Additionally, qualitative results indicate that **our method generally adapts better to the target domain and produces higher-quality outputs than NADA, IPL, and SVL, while HGAN suffers severely from mode collapse to one of the target examples.** For example, in `Church`→`Skyscraper`, NADA and IPL samples contain artifacts around windows, and SVL still has some structures related to the church like the arch in the middle of the skyscraper. In `Photo`→`Anime`, NADA, IPL, and SVL failed to fully adapt to target style, and HGAN generates buildings that are partially transparent and exhibit undesirable green artifacts.

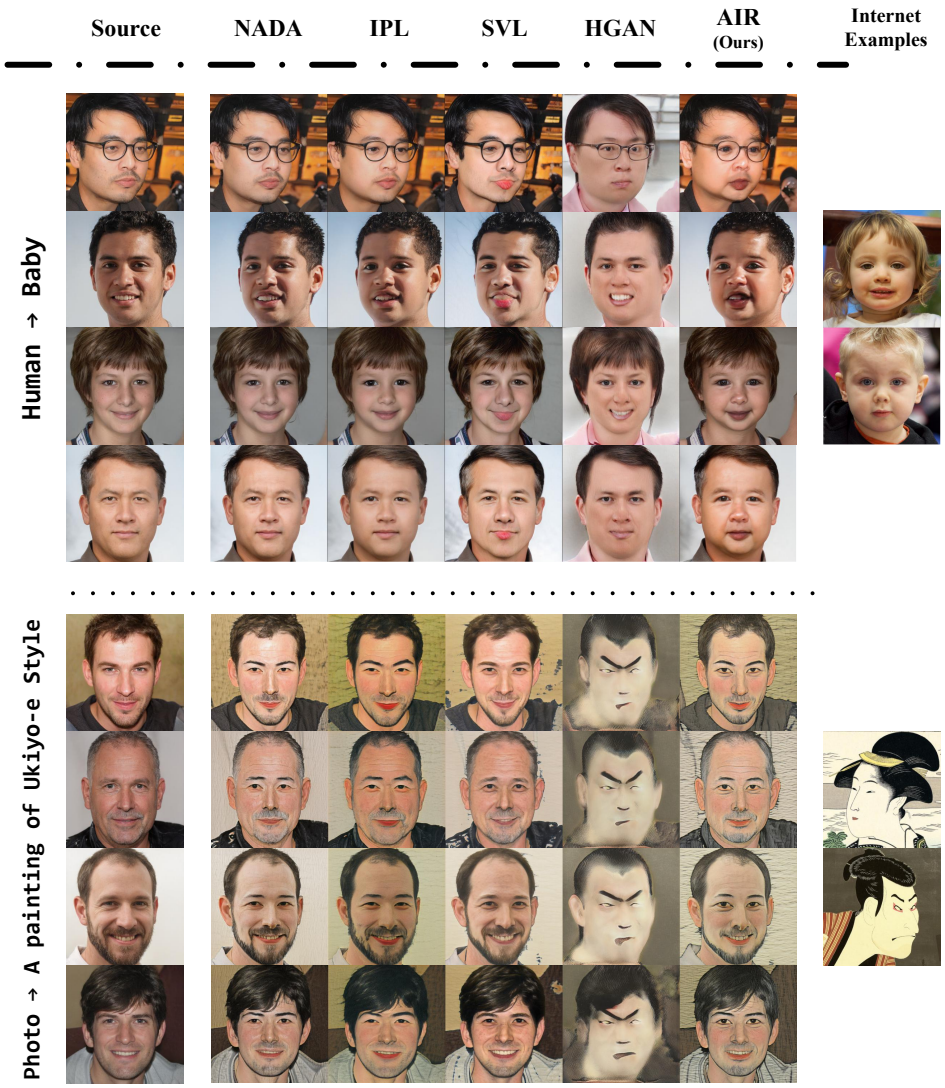



Figure 14: **Additional zero-shot adaptation results from source domain** `FFHQ`**.** Here we use a StyleGAN2 generator pre-trained on the FFHQ [39] (human faces) dataset as $G_S$ and shift this to various target domains using different zero-shot approaches. We report the qualitative results for two setups: `Photo`→`Baby` and `Photo`→`A Painting of Ukiyo-e Style`. We also compute CLIP Distance on 5K generated samples as quantitative results, as shown in Tab. 2, our proposed AIR approach results in less CLIP Distance meaning that the adapted generator are closer to the target domain. Additionally, qualitative results indicate that **our method generally adapts better to the target domain and produces higher-quality outputs than NADA, IPL, and SVL, while HGAN suffers severely from mode collapse to one of the target examples.**

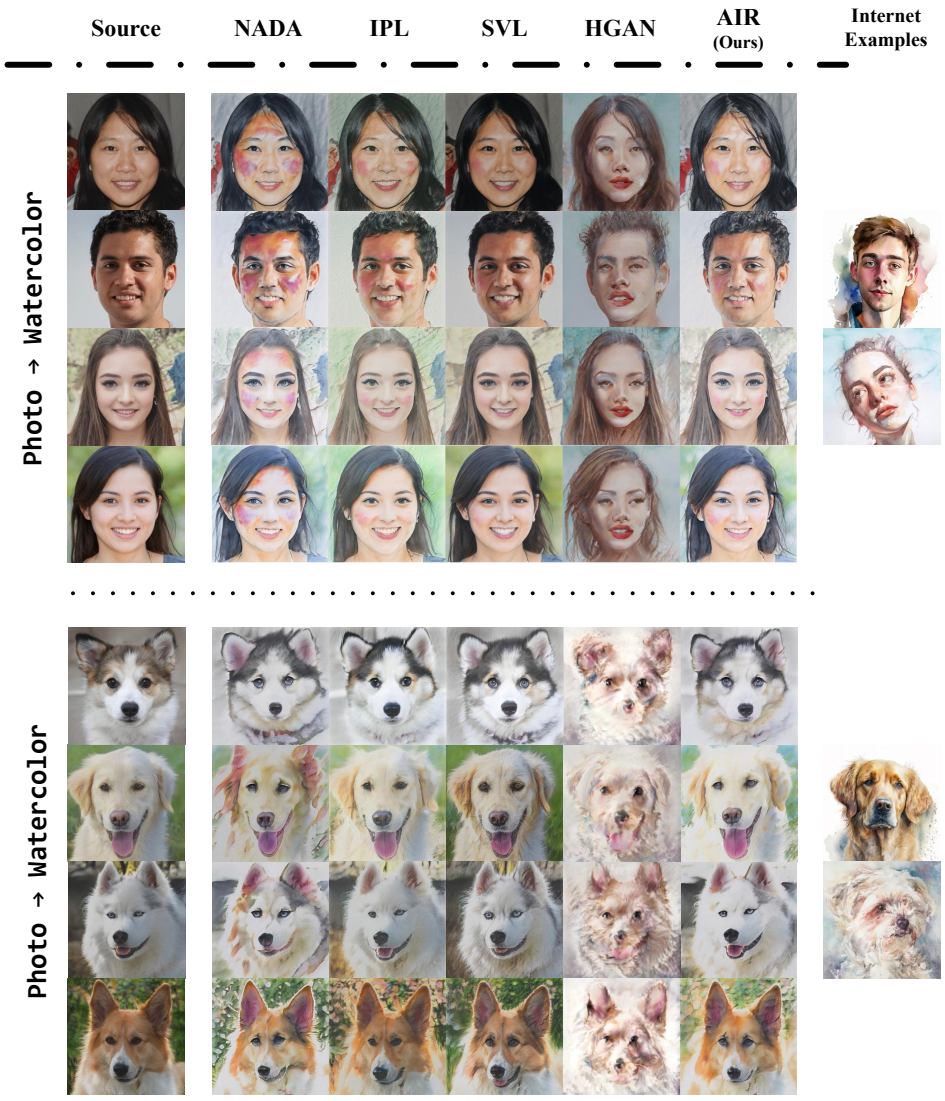



Figure 15: **Additional zero-shot adaptation results from source domain** `FFHQ`**.** Here we use a StyleGAN2 generator pre-trained on the FFHQ [39] (human faces) and AFHQ-Dog [40] (dog face) dataset as $G_{\mathcal{S}}$ and shift this to various target domains using different zero-shot approaches. We report the qualitative results for the setups: `Photo`→`Watercolor`. We also compute CLIP Distance on 5K generated samples as quantitative results, as shown in Tab. 2, our proposed AIR approach results in less CLIP Distance meaning that the adapted generator are closer to the target domain. Additionally, qualitative results indicate that **our method generally adapts better to the target domain and produces higher-quality outputs than NADA, IPL, and SVL, while HGAN suffers severely from mode collapse to one of the target examples.**

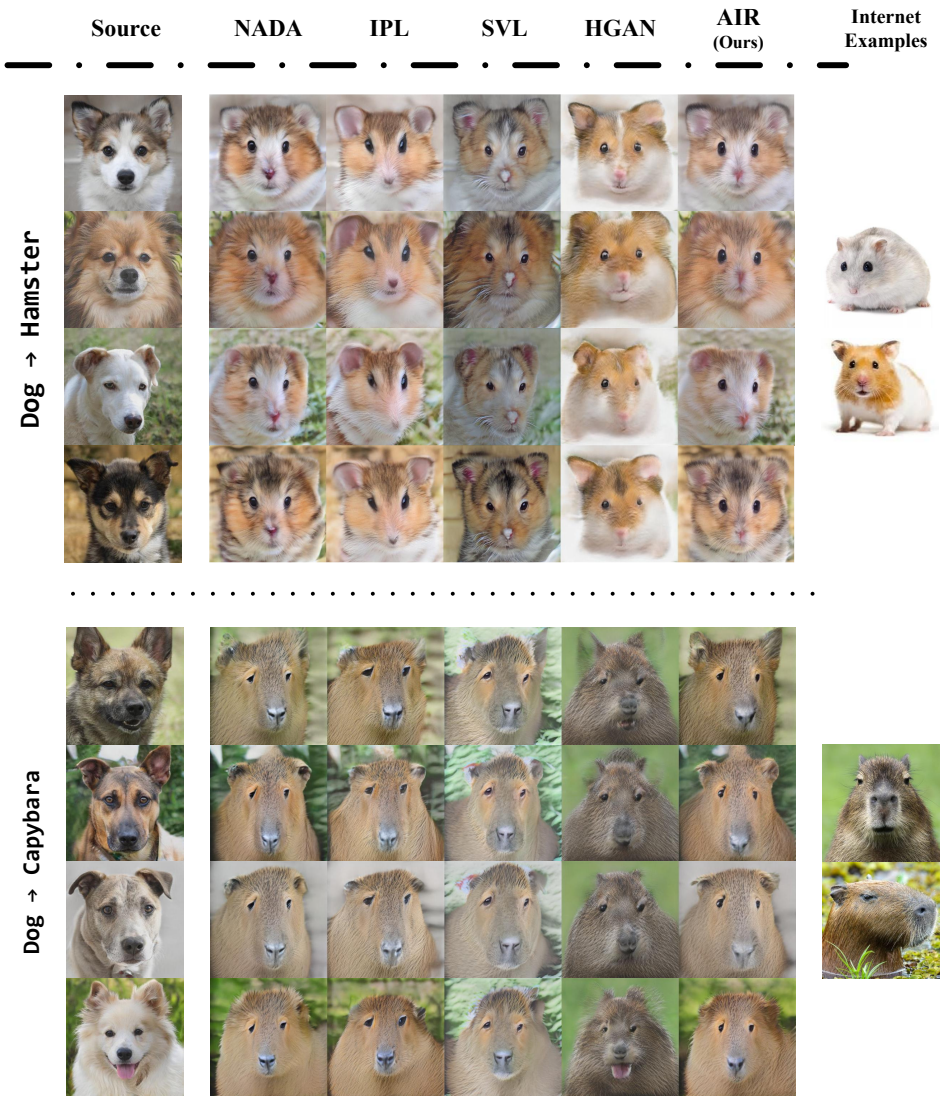



Figure 16: **Additional zero-shot adaptation results from source domain** Dog**.** Here we use a StyleGAN2 generator pre-trained on the AFHQ-Dog [40] dataset as $G_{\mathcal{S}}$ and shift this to various target domains using different zero-shot approaches. We report the qualitative results for two setups: Dog→Hmaster and Dog→Capybara. We also compute CLIP Distance on 5K generated samples as quantitative results, as shown in Tab. 2, for both setups, our proposed AIR approach results in less CLIP Distance meaning that the adapted generator are closer to the target domain. Additionally, qualitative results indicate that **our method generally adapts better to the target domain and produces higher-quality outputs than NADA, IPL, and SVL, while HGAN suffers severely from mode collapse to one of the target examples.** For example, for Dog→Capybara setup, generated samples with other approaches have degradations like unsymmetrical faces or eyes.

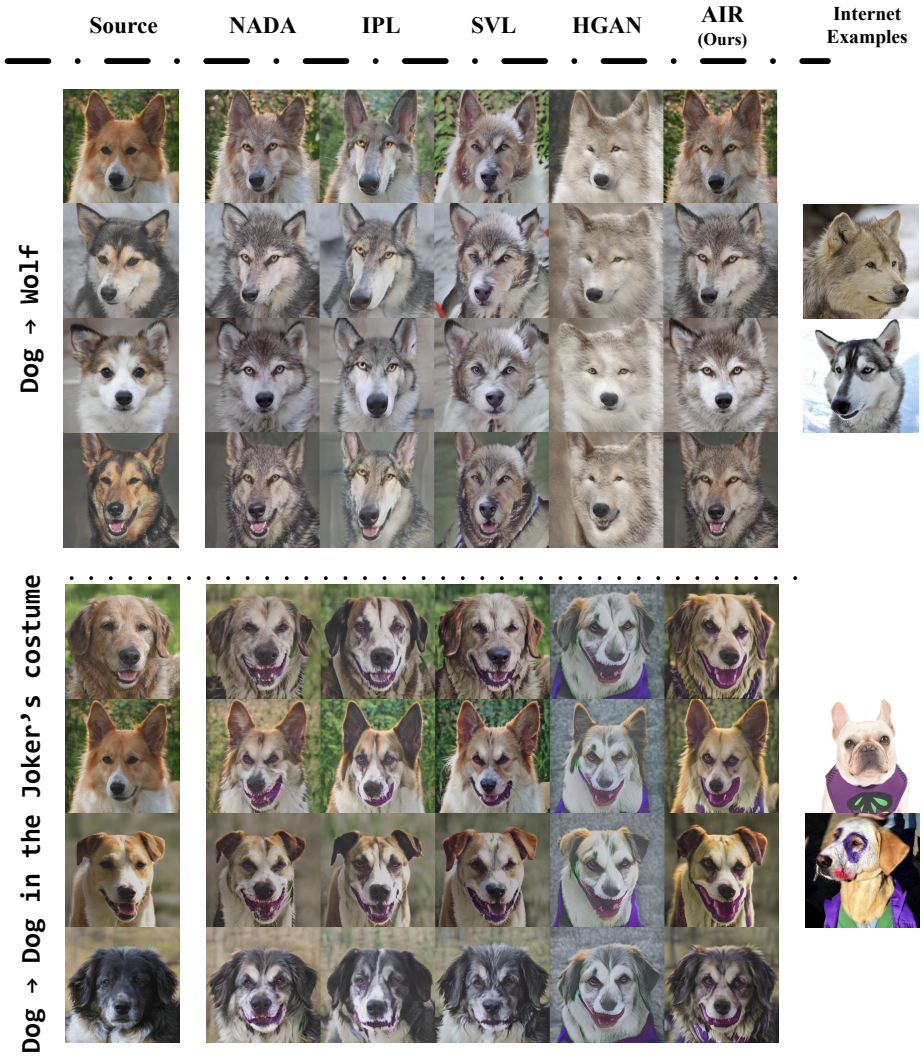



Figure 17: **Additional zero-shot adaptation results from source domain** Dog**.** Here we use a StyleGAN2 generator pre-trained on the AFHQ-Dog [40] dataset as $G_S$ and shift this to various target domains using different zero-shot approaches. We report the qualitative results for two setups: Dog→Wolf and Dog→The Joker. We also compute CLIP Distance on 5K generated samples as quantitative results, as shown in Tab. 2, for both setups, our proposed AIR approach results in less CLIP Distance meaning that the adapted generator are closer to the target domain. Additionally, qualitative results indicate that **our method generally adapts better to the target domain and produces higher-quality outputs than NADA, IPL, and SVL, while HGAN suffers severely from mode collapse to one of the target examples.** For example, for Dog→Wolf setup, IPL generates an unnaturally big snout and SVL has some artifacts in the generated sample. For Dog→The Joker setup, our approach attains the mouth feature with a proper style and quality.

Table 6: Full per-setting guidance-misalignment results for zero-shot diffusion model adaptation. We compare the directional guidance used by each method with the corresponding reference visual offset. Results are reported as mean ± standard deviation.

| Pre-trained Dataset | Source → Target Setting | Guidance Misalignment ↓ | |
|---|---|---|---|
| | | NADA | AIR |
| FFHQ | Human → Baby | 0.68± 0.02 | **0.52**± 0.08 |
| | Human → Werewolf | 0.71± 0.01 | **0.34**± 0.05 |
| | Human → Pixar character | 0.58± 0.01 | **0.38**± 0.09 |
| | Photo → Sketch | 0.42± 0.02 | **0.41**± 0.07 |
| | Photo → Wall painting | 0.71± 0.01 | **0.52**± 0.10 |
| | Photo → Watercolor | 0.56± 0.02 | **0.32**± 0.06 |
| | Photo → Ukiyo-e | 0.73± 0.01 | **0.52**± 0.08 |
| Dog | Dog → Cat | 0.54± 0.02 | **0.24**± 0.09 |
| | Dog → Wolf | 0.33± 0.02 | **0.24**± 0.07 |
| | Dog → Hamster | 0.36± 0.01 | **0.26**± 0.11 |
| | Dog → Capybara | 0.29± 0.01 | **0.24**± 0.05 |
| | Dog → the Joker | 0.64± 0.02 | **0.42**± 0.10 |
| | Photo → Cartoon | 0.59± 0.01 | **0.38**± 0.08 |
| | Photo → Watercolor | 0.49± 0.02 | **0.41**± 0.06 |

Table 7: Full quantitative results of zero-shot diffusion model adaptation. Results are reported as mean ± standard deviation computed from 3 runs with different random seeds. All methods use the same CLIP guidance for fair comparison. FID is available only for Human → Baby and Dog → Cat, which have sufficient data for reliable evaluation

| Pre-trained Dataset | Adaptation | CLIP Distance (↓) | | Intra-LPIPS (↑) | | FID (↓) | |
|---|---|---|---|---|---|---|---|
| | | NADA | AIR | NADA | AIR | NADA | AIR |
| FFHQ | Human → Baby | 0.2612± 0.0018 | **0.2172**± 0.0022 | 0.5690± 0.0026 | **0.5763**± 0.0037 | 63.68± 0.75 | **58.74**± 1.56 |
| | Human → Werewolf | 0.2777± 0.0013 | **0.2323**± 0.0026 | **0.5194**± 0.0025 | 0.5185± 0.0048 | - | - |
| | Human → Pixar character | 0.4296± 0.0011 | **0.3865**± 0.0025 | **0.4565**± 0.0024 | 0.4558± 0.0039 | - | - |
| | Photo → Sketch | 0.4390± 0.0014 | **0.3601**± 0.0037 | **0.4873**± 0.0028 | 0.4815± 0.0035 | - | - |
| | Photo → Wall painting | 0.4781± 0.0011 | **0.4744**± 0.0022 | 0.5252± 0.0021 | **0.5270**± 0.0037 | - | - |
| | Photo → Watercolor | 0.3264± 0.0012 | **0.3229**± 0.0033 | 0.6224± 0.0019 | **0.6399**± 0.0037 | - | - |
| | Photo → Ukiyo-e | 0.2307± 0.0014 | **0.2152**± 0.0036 | 0.5431± 0.0018 | **0.5524**± 0.0035 | - | - |
| Dog | Dog → Cat | 0.1434± 0.0017 | **0.1428**± 0.0023 | 0.5411± 0.0018 | **0.5494**± 0.0034 | 84.75± 1.07 | **77.15**± 1.49 |
| | Dog → Wolf | 0.1460± 0.0015 | **0.1337**± 0.0036 | 0.4649± 0.0022 | **0.4735**± 0.0035 | - | - |
| | Dog → Hamster | 0.1851± 0.0017 | **0.1588**± 0.0021 | 0.4854± 0.0018 | **0.4952**± 0.0039 | - | - |
| | Dog → Capybara | 0.1576± 0.0011 | **0.1204**± 0.0026 | **0.4832**± 0.0022 | 0.4604± 0.0052 | - | - |
| | Dog → the Joker | 0.3955± 0.0012 | **0.3208**± 0.0022 | 0.4571± 0.0018 | **0.4663**± 0.0039 | - | - |
| | Photo → Cartoon | 0.2516± 0.0012 | **0.2481**± 0.0025 | 0.5572± 0.0027 | **0.5611**± 0.0040 | - | - |
| | Photo → Watercolor | 0.1932± 0.0016 | **0.1853**± 0.0027 | 0.5173± 0.0022 | **0.5322**± 0.0038 | - | - |
| | Avg. | 0.2797± 0.0014 | **0.2513**± 0.0028 | 0.5164± 0.0022 | **0.5207**± 0.0039 | 74.22± 0.91 | **67.95**± 1.53 |

### C.3.1 Zero-shot GAN Adaptation

In this section, we provide additional experimental results including quantitative and qualitative results for different adaptation setups using GAN as the generator and introduce more evaluation metrics.

**Qualitatives Results.** In Fig. 13 , we perform zero-shot adaptation of a StyleGAN2 pre-trained on LSUN-Church [51] to four different target domains including Skyscraper, and Anime. In Fig. 14 and Fig. 15 show the qualitative results of zero-shot adaptation of a StyleGAN2 pre-trained on FFHQ [39] dataset to two different target domains including Baby, Watercolor, and A Painting of Ukiyo-e Style. Finally, Fig. 15, Fig. 16 and Fig. 17 we report the zero-shot adaptation of a StyleGAN2 pre-trained on AFHQ-Dog [40] to four different target domains including Watercolor, Hamster, Capybara, Wolf and The Joker. The results show that our approach in general adapts better to the style of the target domain and has better sample quality (please check the caption of each image for more detailed discussion).

**Quantitative Results.** Quantitative results are reported by computing the CLIP Distance between the embeddings of 5K generated samples with each approach and the embedding of the reference images of the target domain in CLIP space. As the results show, generated samples by proposed

AIR has smaller CLIP distance meaning that the adapted generators is closer to the target domain compared to the adapted generator by other zero-shot approaches.

**Additional Metrics.** We further evaluate the quality of the domain adaptation by introducing two additional metrics SigLIP Distance [2] and DINOv2 Distance [49]. Similar to the computation of CLIP Distance, SigLIP and DINOv2 Distance are defined as the cosine distance between the SigLIP/DINOv2 embeddings of collected and generated samples. As shown in Tab. 8, the results align with those in the main paper, further support the superiority of our proposed AIR.

Table 8: Additional quantitative evaluation of zero-shot GAN adaptation using additional metrics of SigLIP/DINOv2, with the same setting of Tab. 2 in main paper.

| Pre-trained Dataset | Target Domain | SigLIP Distance (↓) | | | | | DINOv2 Distance (↓) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NADA | IPL | SVL | HGAN | AIR | NADA | IPL | SVL | HGAN | AIR |
| FFHQ | Baby | 0.1925 | 0.1884 | 0.3474 | 0.3080 | **0.1833** | 0.5943 | 0.5993 | 0.8026 | 0.7703 | **0.5887** |
| | Werewolf | 0.3192 | 0.3930 | 0.4831 | 0.2425 | **0.2274** | 0.8500 | 0.8923 | 0.9365 | 0.6345 | **0.6097** |
| | Pixar | 0.2803 | 0.2762 | 0.4582 | 0.2857 | **0.2630** | 0.6935 | **0.6690** | 0.7848 | 0.7362 | 0.6785 |
| | Sketch | 0.2897 | 0.3173 | 0.3598 | 0.2939 | **0.2837** | 0.4682 | 0.5918 | 0.6291 | 0.4703 | **0.4420** |
| | Wall painting | 0.4205 | 0.4277 | 0.4489 | **0.4052** | 0.4103 | 0.7256 | 0.7267 | 0.7836 | **0.6761** | 0.7004 |
| AFHQ-Dog | Cat | 0.1395 | 0.2287 | 0.1819 | 0.1762 | **0.1297** | 0.8553 | 0.8737 | 0.8842 | 0.8623 | **0.8338** |
| | Cartoon | 0.2580 | 0.2618 | 0.3140 | 0.3101 | **0.2518** | 0.7899 | 0.8174 | 0.9078 | 0.8513 | **0.7644** |
| | Watercolor | 0.1934 | 0.1980 | 0.2569 | 0.1916 | **0.1819** | 0.8059 | 0.8328 | 0.8330 | **0.7721** | 0.7943 |

### C.3.2 Zero-shot Diffusion Model Adaptation

In this section, we provide more qualitative and quantitative results of zero-shot diffusion model adaptation.

**Qualitative Results.** Here, we report the qualitative results of zero-shot diffusion model adaptation for the same configuration used in Tab. 3 (main paper). More specifically, we use the pre-trained Guided Diffusion model [41] on two different source domains FFHQ [39] (Fig. 18) and AFHQ-Dog [40] (Fig. 19) and shift these pre-trained models to different target domains using only text descriptions for both NADA and our proposed AIR approaches. As illustrated in Fig. 18 and Fig. 19, the domain adaptation with NADA suffer from degradation in the form of artifacts compared to our proposed AIR approach.

## C.4 Hyperparameters of AIR

### C.4.1 Hyperparameters of AIR on ZSGM

In Alg. 1, for both GAN and diffusion model adaptation the batch size is set to 2. Adaptation iteration $t_{adapt}$ is set to 200 for in-domain changes like `Human`→`Baby`, 300 for texture-based changes such as `Photo`→`Sketch`, and 2,000 for animal changes like `Dog`→`Cat`. We set $t_{thresh} = 50\%t_{adapt}$ to ensure there are some target domain concept encoded in $G_t$, and $t_{int} = 10\%t_{adapt}$ to facilitate a stable and efficient training.

In Alg. 2, we generate 1,000 pairs of source and anchor images with the same batch of $w$ for each update. The number of learnable concept-description tokens $m$ is set to 4, and is initialized by "`A photo of a`". Each concept description learning session requires $k_{iter} = 200$ iterations.

For all experiments, we use an ADAM optimizer with a learning rate of 0.002 for both Algs. 1 and 2. We conduct all the experiments on a single NVIDIA RTX 6000 Ada GPU. The training time is comparable to NADA as concept description learning in Alg. 2 only requires ∼20 seconds.

### C.4.2 Hyperparameters of AIR on TOIE

We follow Style-Editor [11] to set batch size as 1, and optimize for an iteration $t_{adapt}$ of 200 for all settings. Similar to zero-shot generative model domain adaptation we set $t_{thresh} = 50\%t_{adapt}$ to ensure there are some target domain concept encoded in $G_t$, and $t_{int} = 10\%t_{adapt}$ to facilitate a stable and efficient training.

In Alg. 2, we take the transformed visual directly as the anchor visual. The number of learnable concept-description tokens $m$ is set to 4, and is initialized by "`A photo of a`". Each concept description learning session requires $k_{iter} = 200$ iterations.

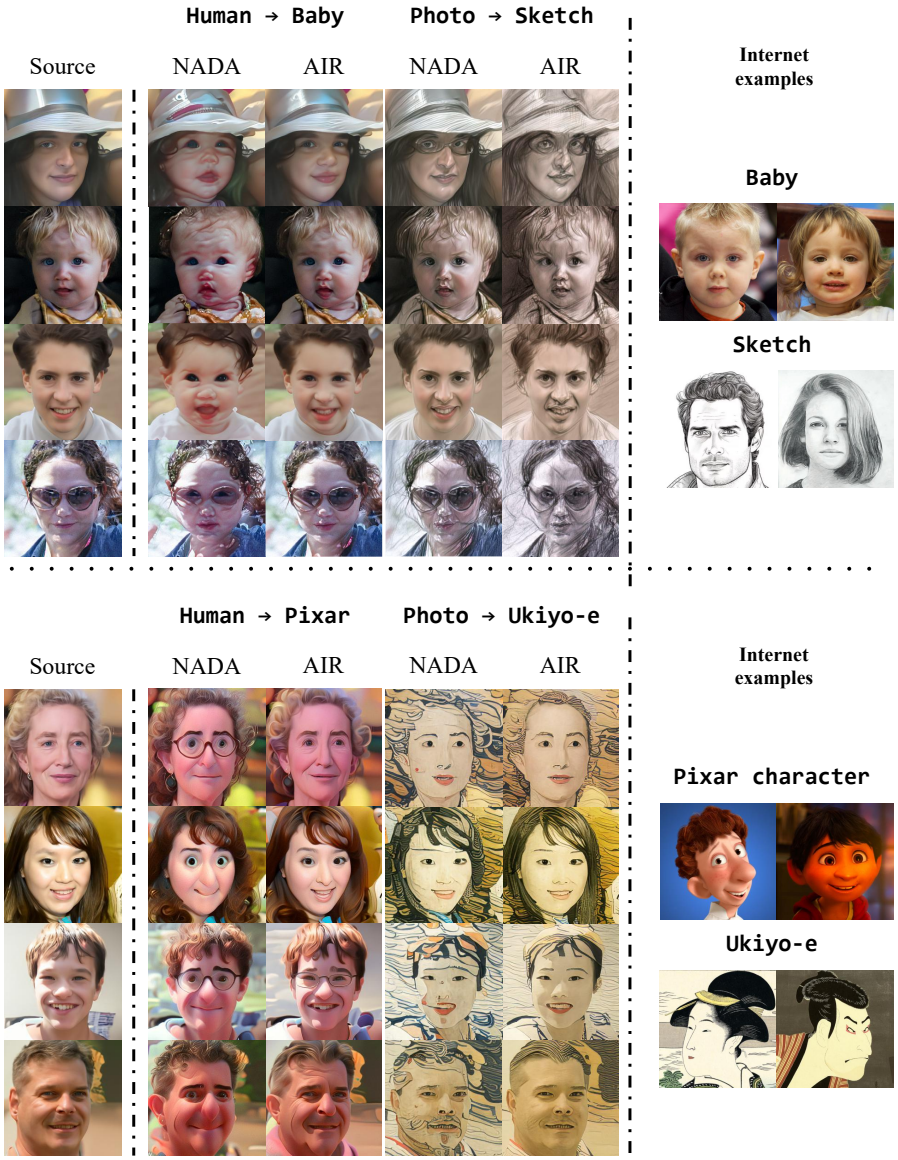



Figure 18: **Additional zero-shot adaptation results.** We use a pre-trained Guided Diffusion model [41] on FFHQ dataset [39] as pre-trained generator $G_S$ and perform zero-shot adaptation in four different setups: Human→ Baby, Photo→ Sketch, Human→ Pixar Character, and Photo→ A painting in Ukiyo-e style using both NADA and our proposed AIR approach. Quantitative results measured by CLIP distance in Tab. 3 shows that the domains adapted by our approach are closer to the target domain. In addition, qualitative results show that NADA suffers from degradation.

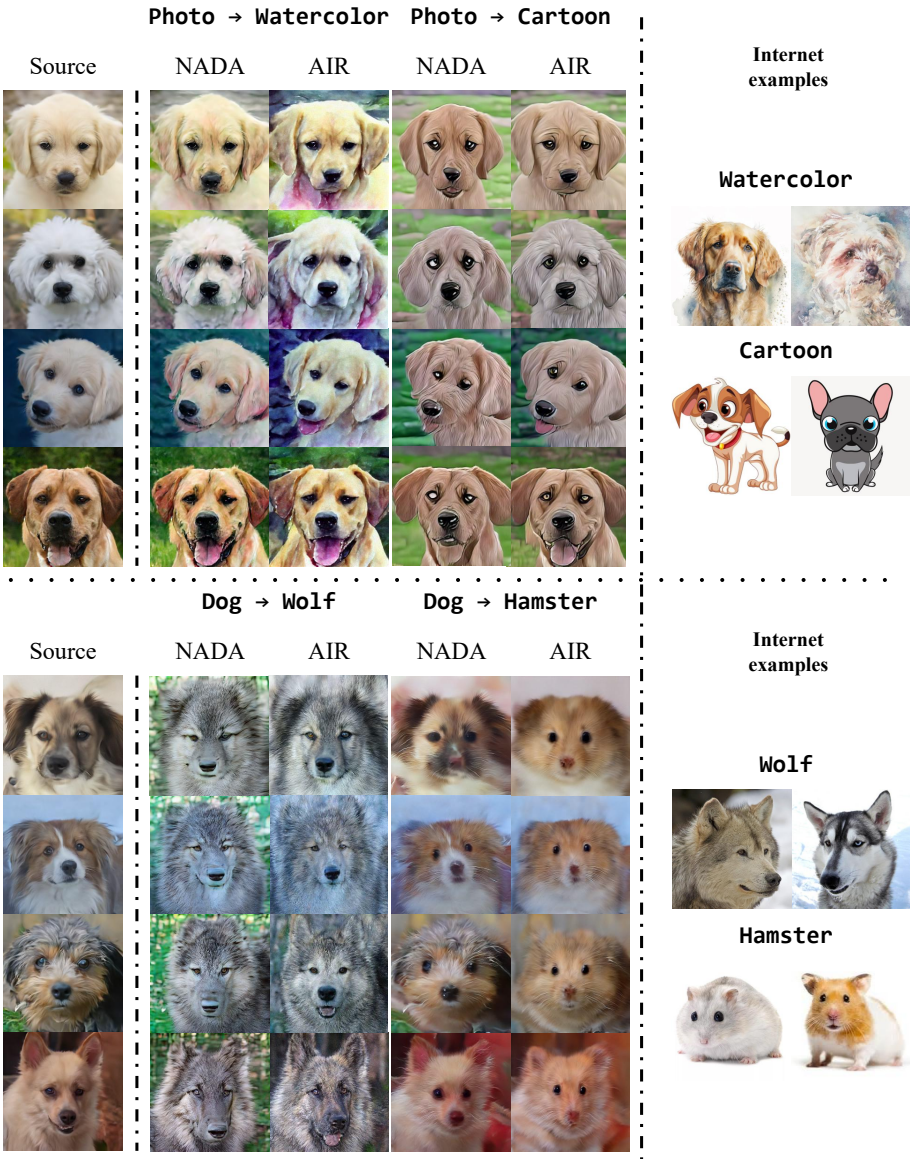



Figure 19: **Additional zero-shot adaptation results.** We use a pre-trained Guided Diffusion model [41] on AFHQ-Dog dataset [40] as pre-trained generator $G_{\mathcal{S}}$ and perform zero-shot adaptation in four different setups: Photo→ Watercolor, Photo→ Cartoon, Dog→ Wolf, and Dog→ Hamster using both NADA and our proposed AIR approach. Quantitative results measured by CLIP distance in Tab. 3 shows that the adapted domain by our approach are closer to the target domain. Similarly, qualitative results show that our proposed AIR approach has better performance compared to NADA.

### C.5 Additional Ablation Studies

#### C.5.1 Ablation on Hyperparameters Selection

We conduct an ablation study to determine the optimal hyperparameters. Specifically, Tab. 9 shows the ablation results for the adaptation interval $t_{int}$ to update anchor. Tab. 10 shows the ablation results for the starting iteration $t_{thresh}$ of applying AIR. A large $t_{int}$ which results in fewer updates of the anchor point, generally leads to a degradation in performance due to less precise adaptation. Conversely, a small $t_{int}$, while more computationally expensive, does not yield significant improvement. Thus, we set $t_{int} = 10\%$ to balance the computation cost and adaptation precision. Similarly, neither excessively large nor small values of $t_{thresh}$ provide optimal adaptation performance. As shown in Fig. 20 (b) and (c), the visual ablation results align with this conclusion. Hence, we empirically select $t_{thresh} = 50\%$.

Table 9: Ablation study on adaptation interval $t_{int}$ to update anchor.

| | $t_{int}$ | | | |
|---|---|---|---|---|
| % of $t_{adapt}$ | Human → Baby | | Dog → Cat | |
| | FID (↓) | Intra-LPIPS (↑) | FID (↓) | Intra-LPIPS (↑) |
| 5% | 59.45 | 0.4512 | 59.97 | 0.4560 |
| 10% | 62.13 | **0.4520** | **56.20** | **0.4628** |
| 15% | 58.87 | 0.4515 | 61.92 | 0.4635 |
| 20% | 64.54 | 0.4496 | 65.49 | 0.4537 |
| 25% | **56.69** | 0.4511 | 67.49 | 0.4374 |
| 30% | 76.39 | 0.4506 | 77.23 | 0.4513 |

Table 10: Ablation study on starting iteration $t_{thresh}$ of applying AIR.

| | $t_{thresh}$ | | | |
|---|---|---|---|---|
| % of $t_{adapt}$ | Human → Baby | | Dog → Cat | |
| | FID (↓) | Intra-LPIPS (↑) | FID (↓) | Intra-LPIPS (↑) |
| 0% | 63.56 | 0.4324 | 83.33 | **0.4815** |
| 12.5% | 75.92 | 0.4259 | 78.21 | 0.4642 |
| 25.0% | 72.17 | 0.4386 | 73.65 | 0.4516 |
| 37.5% | 64.14 | **0.4558** | 59.96 | 0.4496 |
| 50.0% | **62.13** | 0.4520 | **56.20** | 0.4628 |
| 67.5% | 68.25 | 0.4542 | 56.46 | 0.4344 |

#### C.5.2 Ablation on Anchor Label Initialization

Our anchor concept description initialization in Sec. 4.2 is inspired by the standard concept description learning in VLM [35, 52], which initializes label token $Y$ with class label of the image to serve as prior. However, in our setting, the anchor domain encodes both source and target concepts, so it cannot be described with natural language. Therefore, we leverage the continuous and semantically rich representation space of the text encoder to initialize anchor label $Y_{\mathcal{A}_i}$ via **interpolation** between tokenized source and target descriptions. We conduct an ablation study on the initialization of $Y_{\mathcal{A}_i}$. Results shown in Tab. 11 indicate the effectiveness of our idea by obtaining the best FID and Intra-LPIPS.

#### C.5.3 Visual Ablation Studies

We perform visual ablation studies on concept description learning design and hyperparameters selection with the same experiments setting of Sec. 5.4, C.5.1, and Tab. 12. The results in Fig. 20 align consistently with the quantitative findings.

#### C.5.4 Ablation on Loss Weighting

We evaluate the sensitivity of our training objective $\mathcal{L} = \mathcal{L}_{directional} + W\mathcal{L}_{adaptive}$ by varying the weighting factor $W$. As shown in Tab. 12, the effectiveness of our proposed $\mathcal{L}_{adaptive}$ is robust across different weight values, consistently improving over the baseline NADA setting ($W = 0$). It

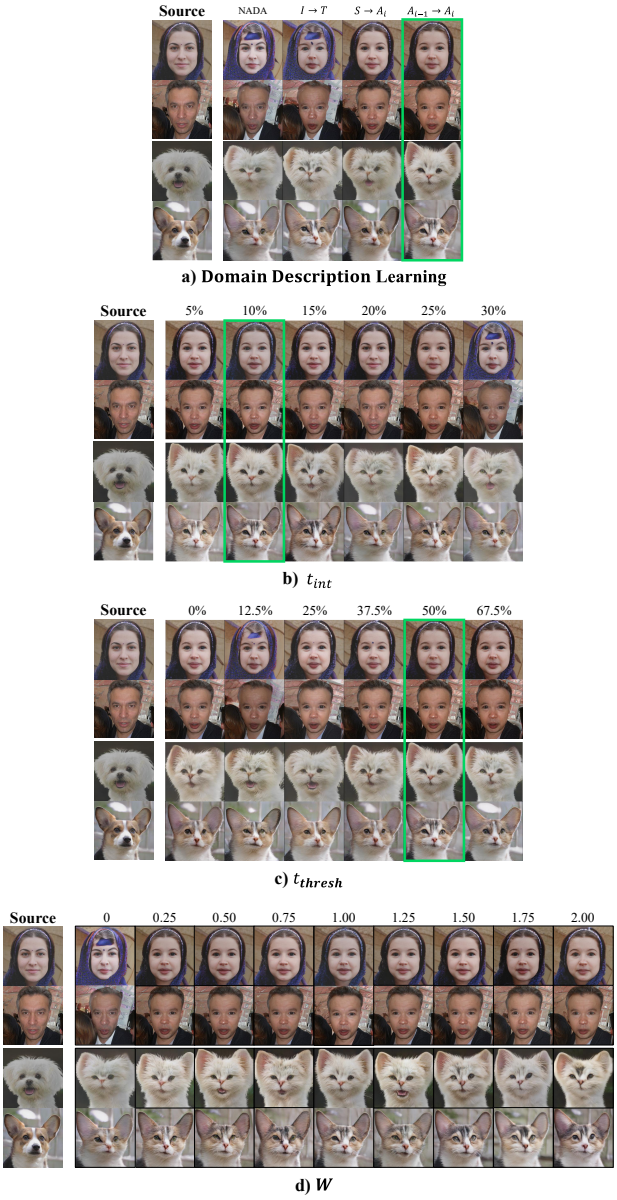



Figure 20: Visual ablation study of: a) Design choice of concept description learning; b) Adaptation interval $t_{int}$ to update anchor; c) Starting iteration $t_{thresh}$ of applying AIR. d) Loss weight $W$ for the proposed $\mathcal{L}_{adaptive}$

Table 11: Ablation study on initialization of $Y_{\mathcal{A}_i}$, using: a) Target domain label; b) Source domain label for the first half of adaptation, then target domain label; c) Interpolation as in our AIR.

| Init. | Human → Baby | | Dog → Cat | |
|---|---|---|---|---|
| | FID (↓) | Intra-LPIPS (↑) | FID (↓) | Intra-LPIPS (↑) |
| a) | 67.53 | 0.4513 | 65.83 | 0.4373 |
| b) | 63.34 | 0.4512 | 56.44 | 0.4466 |
| c) | **62.13** | **0.4520** | **56.20** | **0.4628** |

is important to note that, we set $W = 1$ in the main paper for simplicity and to avoid introducing an additional hyperparameter.

Table 12: Ablation study on loss weight for our proposed $\mathcal{L}_{adaptive}$. Results shows that the effectiveness of **our proposed $\mathcal{L}_{adaptive}$ is robust across different weight values, consistently improving over the baseline NADA setting ($W = 0$).**

| $W$ | Human → Baby | | Dog → Cat | |
|---|---|---|---|---|
| | FID (↓) | Intra-LPIPS (↑) | FID (↓) | Intra-LPIPS (↑) |
| 0.00 | 68.35 | 0.4474 | 70.87 | 0.4439 |
| 0.25 | 66.59 | 0.4501 | 66.90 | 0.4659 |
| 0.50 | 62.54 | 0.4647 | 69.43 | 0.4580 |
| 0.75 | 58.87 | 0.4494 | 66.62 | 0.4515 |
| 1.00 | 62.13 | 0.4520 | 56.20 | 0.4628 |
| 1.25 | 59.35 | 0.4510 | 57.51 | 0.4458 |
| 1.50 | 63.02 | 0.4525 | 52.62 | 0.4459 |
| 1.75 | 66.38 | 0.4468 | 63.55 | 0.4513 |
| 2.00 | 65.17 | 0.4506 | 55.35 | 0.4596 |

### C.6 Zero-shot Adaptation with Other Multimodal Contrastive Models

We present additional adaptation results with guidance from more multimodal constrastive representation spaces in Tab. 13.

Table 13: Quantitative results of zero-shot generative model adaptation with guidance from more multimodal contrastive models. Results demonstrate that better-aligned multimodal contrastive representation spaces generally lead to better adaptation quality and diversity. AIR achieves SOTA across these multimodal contrastive models as guidance.

| Adaptation | Multimodal Contrastive Model | Avg. Offset Misalignment | CLIP Distance (↓) | | Intra-LPIPS (↑) | | FID (↓) | |
|---|---|---|---|---|---|---|---|---|
| | | | NADA | AIR | NADA | AIR | NADA | AIR |
| Human → Baby | CLIP ViT-B/32 | 0.537 | 0.3327 | **0.3325** | 0.4474 | **0.4520** | 68.35 | **62.13** |
| | CLIP ConvNext-L | 0.408 | 0.3204 | **0.3126** | 0.4551 | **0.4566** | 63.02 | **57.61** |
| | CLIP RN50x64 | 0.613 | 0.3418 | **0.3406** | 0.4422 | **0.4471** | 72.96 | **66.04** |
| | SigLIP ViT-L/16-256 | 0.387 | 0.3171 | **0.3014** | 0.4569 | **0.4588** | 61.72 | **56.08** |
| | SigLIP SoViT-400m | 0.352 | 0.3142 | **0.2937** | 0.4599 | **0.4608** | 59.35 | **54.22** |
| | EVA-CLIP 8B | 0.485 | 0.3269 | **0.3261** | 0.4507 | **0.4535** | 66.17 | **60.12** |
| | DFN5B-CLIP ViT-H/14-378 | 0.331 | 0.3113 | **0.3106** | **0.4627** | 0.4620 | 58.84 | **53.67** |
| | MobileCLIP | 0.652 | 0.3497 | **0.3446** | **0.4409** | 0.4411 | 74.82 | **68.31** |
| Dog → Cat | CLIP ViT-B/32 | 0.537 | 0.1493 | **0.1320** | 0.4439 | **0.4628** | 70.87 | **56.20** |
| | CLIP ConvNext-L | 0.408 | 0.1418 | **0.1297** | 0.4564 | **0.4638** | 62.28 | **49.31** |
| | CLIP RN50x64 | 0.613 | 0.1549 | **0.1462** | 0.4387 | **0.4502** | 76.64 | **60.47** |
| | SigLIP ViT-L/16-256 | 0.387 | 0.1409 | **0.1284** | 0.4582 | **0.4645** | 61.36 | **48.86** |
| | SigLIP SoViT-400m | 0.352 | 0.1382 | **0.1196** | 0.4607 | **0.4653** | 58.56 | **46.57** |
| | EVA-CLIP 8B | 0.485 | 0.1454 | **0.1402** | 0.4491 | **0.4632** | 67.44 | **53.29** |
| | DFN5B-CLIP ViT-H/14-378 | 0.331 | 0.1368 | **0.1253** | **0.4669** | 0.4659 | 57.41 | **45.63** |
| | MobileCLIP | 0.652 | 0.1592 | **0.1367** | 0.4332 | **0.4409** | 89.12 | **64.01** |

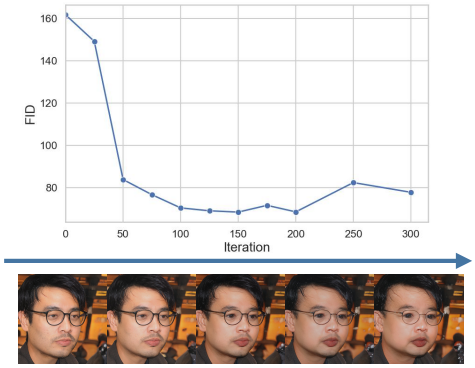



Figure 21: Concept shifts during adaptation. Our result shows that after limited iterations of adaptation using directional loss, **the encoded concept in the adapted generator is already closer to the target domain than the encoded concept in source generator.**

### C.7 Concept Shifts during Transformation

Text-driven visual transformation is an iterative process: even a limited number of updates can move the transformed visual state closer to the target concept than the source. In this section, we design an experiment to demonstrate that the transformed visual already encodes some knowledge related to the target domain. Specifically, following zero-shot generative model domain adaptation setup [3], we perform adaptation on `Human`→`Baby` with StyleGAN2-ADA pretrained on FFHQ [39]. We report FID score throughout the adaptation process to measure the knowledge related to target domain encoded in the transformed visual. As shown in Fig. 21, FID decreases during the initial iterations, indicating that the generated distribution moves closer to the target domain. Additionally, we present qualitative results using the same latent code to further support our findings.

This observation motivates the initial transformation stage of AIR described in Sec. 4. Therefore, we consider NADA generally provides a useful early anchor, and we proposed AIR to subsequently refines the guidance to further improve adaptation quality, without requiring manual identification of a suitable intermediate generator before refinement. The consistency of this behavior is further supported by AIR's improvements across all 40 adaptation settings under the same general configuration.

### C.8 User Study

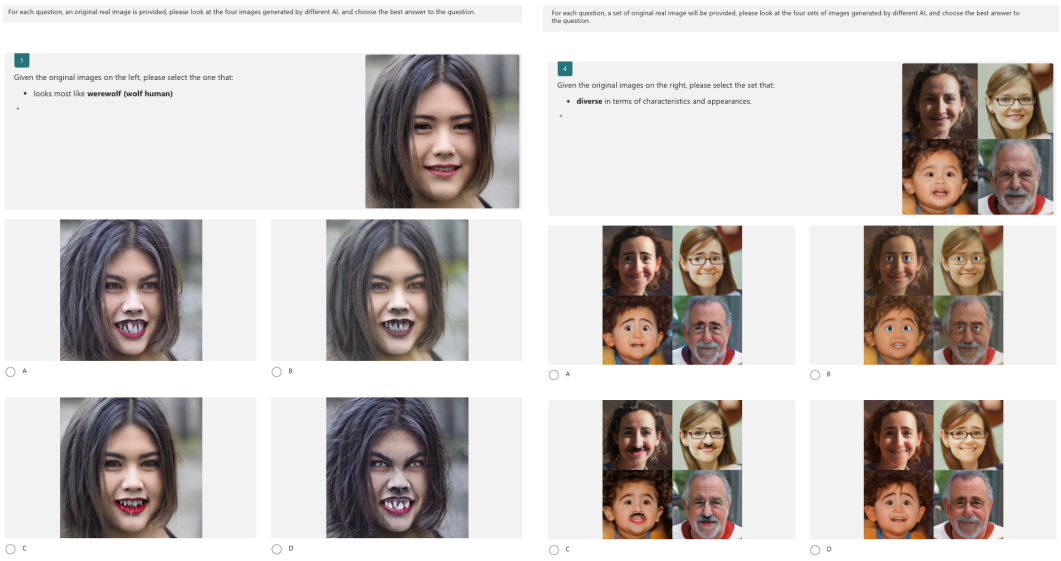
We conduct a user study to compare the quality and the diversity of domain adaptation with different schemes based on human feedback. The questionnaire is performed using the generated samples by different schemes including NADA, IPL, SVL, HGAN and our proposed AIR. It includes 12 questions for quality evaluation and 4 questions for diversity assessment. We include examples for quality and diversity evaluation of our questionnaire in Fig. 22. Finally, we report the percentage of the user preference from 220 responses for each method and for both quality and diversity metrics in Tab. 14. Our user study does not collect identifying information, and does not involve sensitive topics or interventions.

Table 14: Results of our user study (%) experiments.

| Evaluation | NADA | IPL | SVL | HGAN | AIR |
|---|---|---|---|---|---|
| Quality | 24.8 | 4.2 | 3.6 | 11.7 | **55.7** |
| Diversity | 22.4 | **32.8** | 10.8 | 3.4 | 30.6 |

### C.9 Impact of Offset Misalignment on Downstream Tasks

In the previous section, we performed an empirical study that revealed image-text offset misalignment. In this section, we take a step further and investigate the effect of this misalignment on the




(a) Qualitative evaluation (b) Diversity assessment

Figure 22: Examples of user study on (a) Quality assessment and (b) Diversity assessment. We omit one option in the screenshot for better presentation.

generative model adaptation from a source domain (concept) $\mathcal{S}$ to a target domain (concept) $\mathcal{T}$. Specifically, following ZSGM setup [3], for source domain $\mathcal{S}$, we assume a pre-trained generator $G_{\mathcal{S}}$ and a textual description $T_{\mathcal{S}}$ is available. However, for the target domain, only textual description $T_{\mathcal{T}}$ is available. To simulate different degrees of misalignment between source and target, we augment target text to get a set of textual descriptions $\{T^i_{\mathcal{T}}\}$. Then, we perform zero-shot adaptation using the directional loss (Eq. 1) from the source domain $\mathcal{S}$ to each of these target text $T^i_{\mathcal{T}}$ and measure the generation performance of adapted generator.

**Experimental Setup.** For this experiment, we perform adaptation on `Human` $\rightarrow$ `Baby` and `Dog` $\rightarrow$ `Cat`. Our contribution targets the adaptation guidance mechanism and is not tied to a specific generator architecture. To enable rapid validation of the guidance mechanism, we use StyleGAN2-ADA [38] pre-trained on FFHQ [39] and AFHQ-Dog [40] as the pre-trained model for a computationally efficient adaptation testbed. Following [38], we use CLIP. We fix the source text $T_{\mathcal{S}}$ and augment the target text $T_{\mathcal{T}}$ by sampling handcrafted textual descriptions from CLIP ImageNet template (INt)[2] in order to simulate different degrees of misalignment. Then, we follow exactly the same hyperparameters as NADA to adapt the source generator to different target text $T^i_{\mathcal{T}}$. We use FID to measure the performance of the adapted generator against offset misalignment.

Our results in Fig. 23 demonstrate that, in general, **performance of text-driven visual transformation with directional loss degrades as offset misalignment increases**. This downstream validation supports our representation space finding and motivates AIR to iteratively refine text-offset guidance.

### C.10 Additional Experiments of ZSGM

#### C.10.1 Latent Space Interpolation

Building on prior research, we demonstrate that the target domain generators refined through our method retain a smooth latent space property. As illustrated in Fig. 24, each row features a series of images from the same target domain. The left-most and right-most images in each row, labeled as $G_t(w_1)$ and $G_t(w_2)$ respectively, are generated using distinct latent codes $w_1$ and $w_2$. Latent space interpolation between these codes produces an image $G_t((1-\gamma)w_1 + \gamma w_2)$, where $\alpha$ varies from 0 to 1. The visual results show that our method has good robustness and generalization ability. The various target domain spaces obtained by our method are consistently smooth.

[2] https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

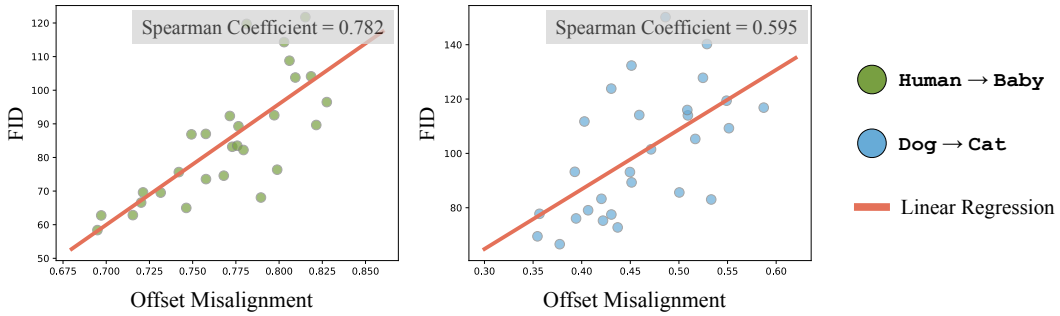



Figure 23: **Impact of offset misalignment on zero-shot generative model adaptation with directional loss:** For each of the two setups, we fix the source domain and augment the textual description of the target domain to simulate various degrees of misalignment between image offset and text offset. Then, we perform the text-driven visual transformation using directional loss in Eq. 1 for each setup. Results show that transformation performance degrades as offset misalignment increases.

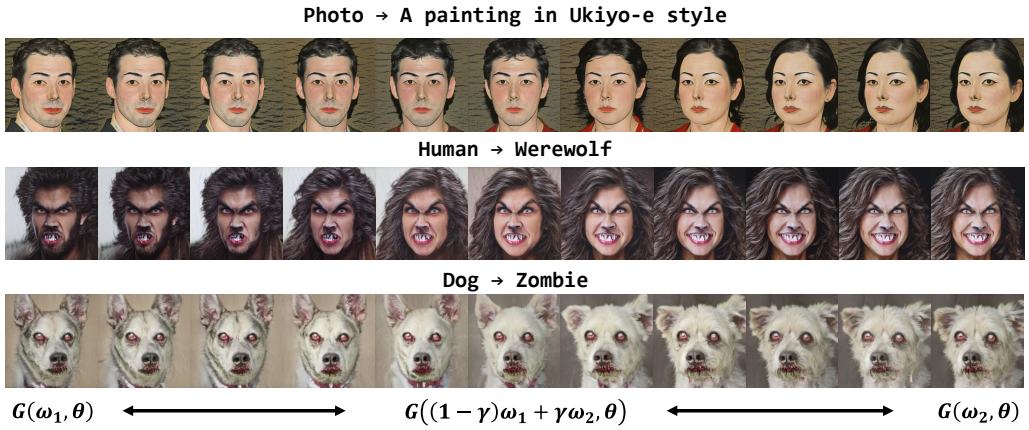



Figure 24: Latent space interpolation. For each row, the left-most column and right-most column are respectively two images synthesized with two different latent codes. The remaining columns refer to images synthesized with interpolated latent codes. Our results demonstrate that **the target domain generators refined through our method retain a smooth latent space property.**

### C.10.2 Cross-model Interpolation

In addition to demonstrating latent space interpolation, we also explore the model's weight smoothness across various domains. Specifically, we perform linear interpolation in the weight space between $G(\cdot, \theta_s)$ and $G(\cdot, \theta_{t_1})$, or between $G(\cdot, \theta_{t_1})$ and $G(\cdot, \theta_{t_2})$. Here, $G(\cdot, \theta_s)$ represents the source domain generator, while $G(\cdot, \theta_{t_1})$ and $G(\cdot, \theta_{t_2})$ are generators adapted to two different target domains. Given a latent code $w$, we produce images via an interpolated model, $G(w, (1-\gamma)\theta_1 + \gamma\theta_2)$, where $\gamma$ ranges from 0 to 1. As illustarted in Fig. 25, our approach effectively supports smooth cross-model interpolation, whether transitioning from a source to a target domain or between different target domains.

### C.10.3 Image Manipulation

To further demonstrate the effectiveness of our proposed method, we also conduct experiments on text-to-image manipulation. It first inverts a image to the latent code by a pre-trained inversion model and then feeds it to the trained target domain generator to get the translated target domain image.

We experiment on both GAN and diffusion model. We use Restyle [53] with e4e encoder [54] to invert a real image into the latent space $w$ for StyleGANs. For the diffusion model, we follow the

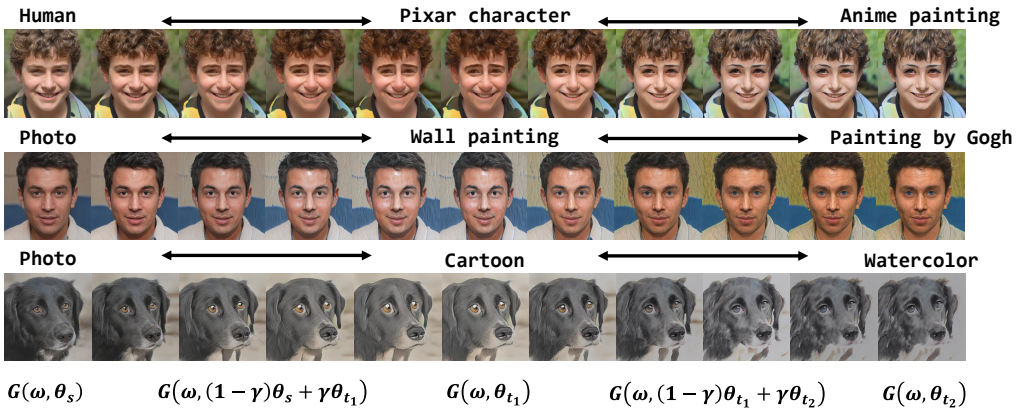



Figure 25: Cross-model interpolation. In each row, the left-most image is generated by the source generator. The middle and the right-most images are synthesized by two different target domain generators. The other images represent cross-model interpolations between two different domains. Our results shows that **the model's weight smoothness across various domains.**

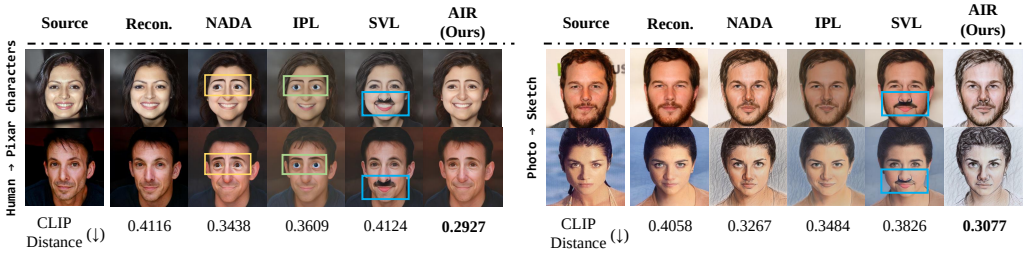



Figure 26: Image manipulation with GAN. The reference image are the same as in Fig. 1

setting of DiffusionCLIP [8] to diffuse a real image and fintune the model to generate an image with target domain features using the diffused image.

**GAN-based Image Manipulation** For GAN-based generators, we perform the experiment by utilizing the inversion model Restyle [53] with e4e encoder [54]. As illustrated in Fig. 26, our method qualitatively exhibits a higher fidelity of target domain features compared to previous methods. Quantitatively, our approach more closely aligns with the reference target images, indicating a greater semantic similarity.

**Diffusion-based Image Manipulation**

We implement based on DiffusionCLIP [8] which seamlessly integrates with the existing zero-shot adaptation methods.

As illustrated in Fig Fig. 27, our method qualitatively exhibits a higher fidelity of target domain feature compared to previous methods. Quantitatively, our approach more closely aligns with the reference target images, indicating a greater semantic similarity.

Fig. 28 illustrates real-world image manipulation results for diffusion AIR.

## D Related Work

**Analogical Reasoning** Research in NLP has shown that word representations of language models are surprisingly good at capturing semantic regularities in language [55, 56]. Specifically, analogical reasoning [17, 18, 19, 20], utilizing the semantic regularities of word representations, aims to solve analogy tasks by using one pair of word vectors to identify the unknown member of a different pair

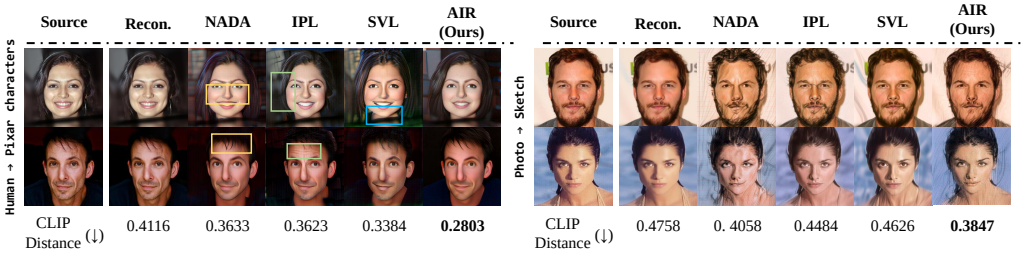



Figure 27: Diffusion model image manipulation.




Figure 28: Additional results of image manipulation with diffusion model.

of words, commonly via alignment of offsets, This is commonly modeled as using the vector offset between two words $a' - a$, and applying it to a new word $b$ to predict the missing word $b'$ that pair with $b$, as illustrated by the famous example of using *v("Man") - v("Woman")* and *v("King")* to identify *v("Queen")*, where $v(\cdot)$ denotes word representation. This approach attracted a lot of attention for the vital role that analogical reasoning plays in human cognition for discovering new knowledge and understanding new concepts. It is already used in many downstream NLP tasks, such as splitting compounds [57], semantic search [58], cross-language relational search [59], etc.

Importantly, previous works [21, 22, 60] demonstrate that the effectiveness of analogical reasoning varies across different categories and semantic relations. More recent studies [23, 24], present a series of experiments performed with BATS dataset [61] on various pre-trained vector space, e.g., GloVe [62], Word2Vec [19], and Skip-gram [18], indicate that it is more effective to use $a' - a$ and $b$ to determine $b'$ when $b$ and $b'$ are close in vector space; and less so when $b$ and $b'$ are more apart.

Inspired by these studies, in this work, we perform an empirical study of offset misalignment in multimodal contrastive representation spaces and observe that for distant concepts in multimodal contrastive representation space, image and text offsets suffer from more misalignment, while closely related concepts suffer less. Based on this analysis, AIR iteratively refines text offsets to reduce image-text offset misalignment and improve directional guidance in text-driven visual transformation tasks.

**Generative model adaptation** studies how a pre-trained generator can be adapted to new domains under limited supervision while preserving sample quality and diversity [63, 45, 64, 65, 6, 66, 67, 68, 69]. A particularly challenging setting within generative model adaptation is *Zero-Shot Generative Model Adaptation (ZSGM)*, where no target domain (concept) images are available and adaptation must rely solely on semantic descriptions of the target domain (concept). Recent methods address this problem by leveraging multimodal contrastive models such as CLIP or SigLIP to derive directional guidance from textual descriptions. In this framework, textual descriptions of the target domain guide generator updates by aligning semantic directions in the representation space, en-

abling the generator to shift from the source domain toward the target domain [3, 4, 5]. In particular, approaches such as StyleGAN-NADA [3] use text offsets in the joint image-text representation space to guide the visual domain transformation during adaptation.

**Image editing** studies how an input image can be modified according to a desired semantic change while preserving content irrelevant to the edit [70, 71, 72, 73, 74]. A particularly flexible setting is *text-guided image editing (TGIE)*, where the desired manipulation is specified by natural-language descriptions rather than predefined attribute labels, reference images, or manually discovered editing directions. Similarly, recent TGIE methods leverage multimodal contrastive models such as CLIP to derive directional guidance from text [7, 8, 10, 11]. In this framework, the edit is modeled as a visual transformation in representation space: the source-to-edited image offset is encouraged to follow the corresponding text offset that describes the intended semantic change. For example, StyleEditor uses text offsets between source object (concept) descriptions and target object (concept) descriptions to guide object-level visual transformations for text-driven image editing [11].

However, the adaptation direction in these two branch of works only focuses on the source and target concepts and is computed once before the visual transformation.

**More application of image-text offset alignment.** Several works address related aspects of text-driven visual transformation. Embedding Arithmetic of Multimodal Queries [12] evaluates offset-vector arithmetic for transformation-based image retrieval, but does not independently quantify the alignment between corresponding visual and textual offsets. DeltaEdit [75] assumes image-text offset alignment and replaces image offsets with text offsets to drive manipulation. EditCLIP [15] learns a task-specific visual encoder to better represent edits and unintended structural changes, rather than analyzing the alignment between native image and text offsets. AugCLIP [16] constructs an ideal edited-image representation from the source image and target text for evaluation, thereby retaining the alignment assumption rather than testing it.

However, none of these works systematically investigates the degree of image-text offset alignment, and assume the image and text offsets in multimodal contrastive representation space are well aligned. In contrast, our work is the first to directly measures the degree of image-text offset alignment and systematically studies how it varies across concept pairs. In this paper, we draw inspiration from a similar problem called analogical reasoning in NLP, and empirically discover that the alignment of image and text offsets in multimodal contrastive representation space is correlated with concept proximity. Based on this finding, AIR iteratively updates the transformation direction, effectively reducing the downstream effect of offset misalignment and achieving SoTA performance on both text-driven visual transformation tasks.

## E Limitation

Our proposed iterative refinement method seeks to improve the quality of zero-shot adaptation. As noted by [4], achieving adaptation across large domain gaps, such as `Human` to `Cat`, is particularly challenging. Similar to previous work, our approach necessitates that the trained generator somewhat closely approximates the target domain before initiating iterative refinement. Additionally, while our experiments on 40 different setups are comprehensive compared to previous work, more setups can be experimented to understand the limitations. We also note that our image-text offset alignment analysis focuses on CLIP/SigLIP-like multimodal contrastive representation spaces and downstream tasks that explicitly use image-text offsets or directional guidance.

## F Social Impact

The AIR methodology holds potential for enhancing artistic image synthesis in social media contexts and could serve as a beneficial data augmentation tool in other computer vision tasks such as recognition and detection. However, its capability to generate realistic images from real-world data raises ethical considerations. It is crucial to address these issues thoughtfully to prevent misuse and ensure responsible application of this technology.

| **Models** | **License** |
|---|---|
| StyleGAN2 [76] | Nvidia Source Code License |
| CLIP [1] | MIT License |
| StyleGAN2-pytorch [76] | MIT License |
| e4e [54] | MIT License |
| StyleGAN-NADA [3] | MIT License |
| IPL [4] | MIT License |
| **Datasets** | **License** |
| FFHQ [76] | CC BY-NC-SA 4.0 |
| AFHQ [40] | CC BY NC 4.0 |

Table 15: Sources and licenses of the utilized models and datasets

## G Licenses

In Table Tab. 15, we specify the source and licenses of the models and datasets used in our work. Note that the FFHQ dataset consists of facial images collected from Flickr, which are under permissive licenses for non-commercial purposes.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Sec. E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Does not include theoretical results

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Sec. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See data and code in supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Sec. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Sec. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Yes

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Sec. F

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: In Sec. F

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Sec. G

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: No.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See Sec. C.8

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We use LLM only for writing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.