Semantic-Aware Data Augmentation for Text-to-Image Synthesis

Zhaorui Tan[1,2], **Xi Yang**[1*], **Kaizhu Huang**[3*]

[1]Department of Intelligent Science, Xi'an Jiaotong-Liverpool University
[2]Department of Computer Science, University of Liverpool
[3] Data Science Research Center, Duke Kunshan University
Zhaorui.Tan21@student.xjtlu.edu.cn, Xi.Yang01@xjtlu.edu.cn, kaizhu.huang@dukekunshan.edu.cn

Abstract

Data augmentation has been recently leveraged as an effective regularizer in various vision-language deep neural networks. However, in text-to-image synthesis (T2Isyn), current augmentation wisdom still suffers from the semantic mismatch between augmented paired data. Even worse, semantic collapse may occur when generated images are less semantically constrained. In this paper, we develop a novel Semantic-aware Data Augmentation (SADA) framework dedicated to T2Isyn. In particular, we propose to augment texts in the semantic space via an Implicit Textual Semantic Preserving Augmentation, in conjunction with a specifically designed Image Semantic Regularization Loss as Generated Image Semantic Conservation, to cope well with semantic mismatch and collapse. As one major contribution, we theoretically show that Implicit Textual Semantic Preserving Augmentation can certify better text-image consistency while Image Semantic Regularization Loss regularizing the semantics of generated images would avoid semantic collapse and enhance image quality. Extensive experiments validate that SADA enhances text-image consistency and improves image quality significantly in T2Isyn models across various backbones. Especially, incorporating SADA during the tuning process of Stable Diffusion models also yields performance improvements.

1 Introduction

Text-to-image synthesis (T2Isyn) is one mainstream task in the visual-language learning community that has yielded tremendous results. Image and text augmentations are two popular methods for regularizing visual-language models (Naveed 2021; Liu et al. 2020). As shown in Figure 2 (a), existing T2Isyn backbones (Xu et al. 2018; Tao et al. 2022; Wang et al. 2022) typically concatenate noises to textual embeddings as the primary text augmentation method (Reed et al. 2016) whilst employing simply basic image augmentations (*e.g.*, Crop, Flip) on images' raw space. Recent studies (Dong et al. 2017; Cheng et al. 2020) suggest text augmentation to be more critical and robust than image augmentation for T2Isyn, given that real texts and their augmentations involve the inference process.

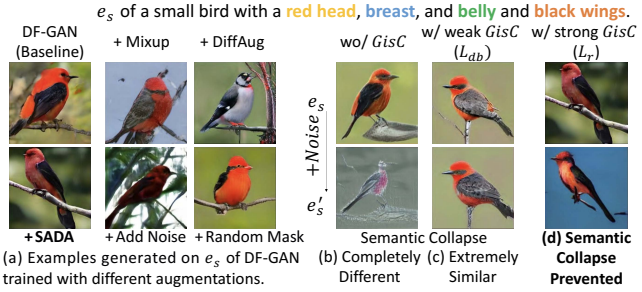

Figure 1: (a) Current augmentations cause semantic mismatch and quality degradation in T2Isyn task. (b)(c) Illustrations of semantic collapse. (d) Our method prevents semantic collapse. See Supplementary Materials D for more.

Albeit their effectiveness, we argue that current popular augmentation methods exhibit two major limitations in the T2Isyn task: 1) *Semantic mismatch* exists between augmented texts/images and generated pairs, it triggers accompanied semantic distribution disruption across both modalities, leading to augmented texts/images lacking corresponding visual/textual representations. As shown in Figure 1 (a), advanced image augmentation, such as Mixup (Zhang et al. 2017a), DiffAug (Zhao et al. 2020), along with text augmentation like Random Mask[1] or Add Noise[2] might weaken both semantic and visual supervision from real images. 2) *Semantic collapse* occurs in the generation process, *i.e.*, when two *slightly* semantic distinct textual embeddings are given, the model may generate either *completely* different or *extremely* similar images. This indicates that the models may be underfitting or over-fitting semantically (see Figure 1 (b)(c)). Both issues will compromise semantic consistency and generation quality. While imposing semantic constraints on generated images can alleviate semantic collapse, the study (Wang et al. 2022) solely focuses on regulating the direction of semantic shift, which may not be entirely adequate.

Motivated by these findings, this paper proposes a novel Semantic-aware Data Augmentation (SADA) framework that offers semantic preservation of texts and images. SADA consists of an Implicit Textual Semantic Preserving Aug-

*Corresponding authors

[1]Randomly masking words in raw texts.
[2]Directly adding random noise to textual semantic embeddings.

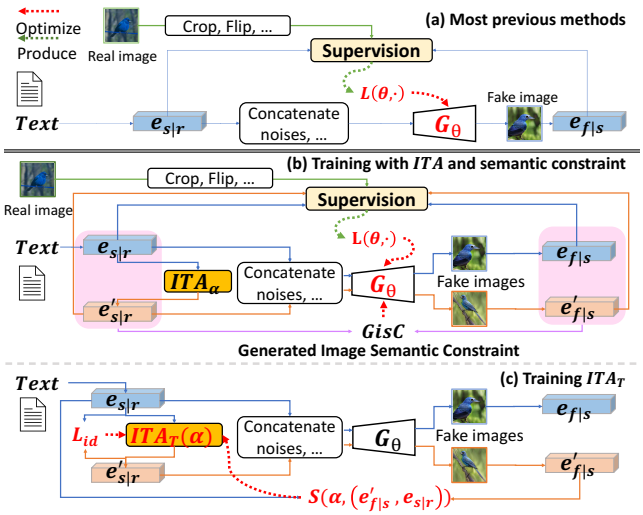

Figure 2: $L(\theta,\cdot)$ is optimization loss for G. $S(\theta,(\cdot,\cdot))$ measures semantic consistency. (a) Simplified training paradigm of previous methods. (b) Training paradigm of SADA. (c) Training of ITA_T where generators are frozen.

mentation (ITA) and a Generated Image Semantic Conservation ($GisC$). ITA efficiently augments textual data and alleviates the semantic mismatch; $GisC$ preserves generated image semantics distribution by adopting constraints on semantic shifts. As one major contribution, we show that SADA can both certify better text-image consistency and avoid semantic collapse with a theoretical guarantee.

Specifically, ITA preserves the semantics of augmented text by adding perturbations to semantic embeddings while constraining its distribution without using extra models. It bypasses the risks of semantic mismatch and enforces the corresponding visual representations of augmented textual embeddings. Crucially, we provide a theoretical basis for ITA enhancing text-image consistency, a premise backed by the group theory for data augmentation (Chen, Dobriban, and Lee 2020). As illustrated in Figure 2 (b), the augmented text embeddings are engaged with the inference process, providing semantic supervision to enhance their regularization role. On the implementation front, two variants for ITA: a closed-form calculation ITA_C (training-free), and its simple learnable equivalent ITA_T. It is further proved that a theoretical equivalence of ITA_C arrives at the same solution to recent methods (Dong et al. 2017; Cheng et al. 2020) that employ auxiliary models for textual augmentation when these auxiliary models are well-trained. This suggests that ITA_C offers an elegant and simplified alternative to prevent semantic mismatch.

Meanwhile, we identify that an effective $GisC$ diminishes semantic collapse and benefits the generated image quality. Inspired by variance-preservation (Bardes, Ponce, and LeCun 2021), we design an Image Semantic Regularization Loss (L_r) to serve as a $GisC$ with ITA_C, which constrains both the semantic shift direction and distance of generated images (see Figure 3 (d)). Through Lipschitz continuity and semantic constraint tightness analysis (as seen in Propositions 4.3 and 4.4), we theoretically justify that L_r prevents the semantic collapse, consequently yielding superior image quality compared to methods that solely bound semantic direction (Gal et al. 2022). Notably, SADA can serve as a theoretical framework for other empirical forms of ITA and $GisC$ in the future.

Our contributions can be summarized as follows:

- This paper proposes a novel Semantic-aware Data Augmentation (SADA) framework that consists of an Implicit Textual Semantic Preserving Augmentation (ITA) and a Generated Image Semantic Conservation ($GisC$).
- Drawing upon the group theory for data augmentation (Chen, Dobriban, and Lee 2020), we prove that ITA certifies a text-image consistency improvement. As evidenced empirically, ITA bypasses semantic mismatch while ensuring visual representation for augmented textual embeddings.
- We make the first attempt to theoretically and empirically show that $GisC$ can additionally affect the raw space to improve image quality. We theoretically justify that using Image Semantic Regularization Loss L_r to achieve $GisC$ prevents semantic collapse through the analysis of Lipschitz continuity and semantic constraint tightness.
- Extensive experimental results show that SADA can be simply applied to typical T2Isyn frameworks, such as diffusion-model-based frameworks, effectively improving text-image consistency and image quality.

The extended version with full Supplementary Materials is available at https://arxiv.org/abs/2312.07951.

2 Related Work

T2Isyn Frameworks and Encoders: Current T2Isyn models have four main typical frameworks: attentional stacked GANs accompanied with a perceptual loss produced by pre-trained encoders (Zhang et al. 2017b, 2018; Xu et al. 2018; Zhu et al. 2019; Ruan et al. 2021), one-way output fusion GANs (Tao et al. 2022), VAE-GANs with transformers (Gu et al. 2022), and diffusion models (DMs) (Dhariwal and Nichol 2021). Two encoders commonly used for T2Isyn are DAMSM (Xu et al. 2018; Tao et al. 2022) and CLIP (Radford et al. 2021). Our proposed SADA is readily applied to these current frameworks with different encoders.

Augmentations for T2Isyn: Most T2Isyn models (Reed et al. 2016; Xu et al. 2018; Tao et al. 2022; Gu et al. 2022) only use basic augmentations such as image corp, flip, and noise concatenation to textual embedding without exploiting further augmentation facilities. To preserve textual semantics, I2T2I (Dong et al. 2017) and RiFeGAN (Cheng et al. 2020) preserve textual semantics using an extra pre-trained captioning model and an attentional caption-matching model respectively, to generate more captions for real images and to refine retrieved texts for T2Isyn. They still suffer from semantic conflicts between input and retrieved texts, and their costly retrieval process leads to infeasibility on large datasets, prompting us to propose a more tractable augmentation method.

Variance Preservation: Stylegan-nada (Gal et al. 2022) presents semantic Direction Bounding (L_{db}) to constrain

semantic shift directions of texts and generated images, which may not guarantee the prevention of semantic collapse. Inspired by variance preservation in contrastive learning (Bardes, Ponce, and LeCun 2021) based on the principle of maximizing the information content (Ermolov et al. 2021; Zbontar et al. 2021; Bardes, Ponce, and LeCun 2021), we constrain the variables of the generated image semantic embeddings to have a particular variance along with its semantic shift direction.

3 Implicit Textual Semantic Preserving Augmentation

Consider observations $\hat{X}_1, ..., \hat{X}_k \in \hat{\mathcal{X}}$ sampled i.i.d. from a probability distribution $\mathbb{P}$ in the sample space $\hat{\mathcal{X}}$, where each $\hat{\mathcal{X}}$ includes real image r and its paired text s. According to $\hat{X} \in \hat{\mathcal{X}}$, we then have $X_1, ..., X_k \in \mathcal{X}$ where each X includes real image embedding e_r and text embedding e_s. We take G with parameter θ as a universal annotation for generators in different frameworks; $L(\theta, \cdot)$ represents total losses for G used in the framework. Following the Group-Theoretic Framework for Data Augmentation (Chen, Dobriban, and Lee 2020), we also assume that:

Assumption 3.1. *If original and augmented data are a group that is exact invariant (i.e., the distribution of the augmented data is equal to that of the original data), semantic distributions of texts/images are exact invariant.*

Consider augmented samples $X' \in \mathcal{X}'$, where X' includes e_r, and augmented textual embedding e'_s. According to Assumption 3.1, we have an equality in distribution:

$$\mathcal{X} =_d \mathcal{X}', \tag{1}$$

which infers that both X and X' are sampled from $\mathcal{X}$. Bringing it down to textual embedding specifically, we further draw an assumption:

Assumption 3.2. *If the semantic embedding e_s of a given text follows a distribution Q_s, then e'_s sampled from Q_s also preserves the main semantics of e_s.*

This assumption can be intuitively understood to mean that for the given text, there is usually a group of synonymous texts. Satisfying exact invariant, e'_s sampled from Q_s preserves the main semantics of e_s. e'_s can be guaranteed to drop within the textual semantic distribution and correspond to a visual representation that shares the same semantic distribution with the generated image on e_s. Thus, e'_s can be used to generate a reasonable image. Under Assumption 3.2, we propose the Implicit Textual Semantic Preserving Augmentation (ITA) that can obtain Q_s. As shown in Figure 3 (a)(b), ITA boosts the generalization of the model by augmenting implicit textual data under Q_s.

3.1 Training Objectives for G with ITA

The general sample objective with ITA is defined as:

$$\min_{\theta} \hat{R}_k(\theta) := \frac{1}{k} \sum_{i=1}^{k} L(\theta, ITA(X_i)). \tag{2}$$

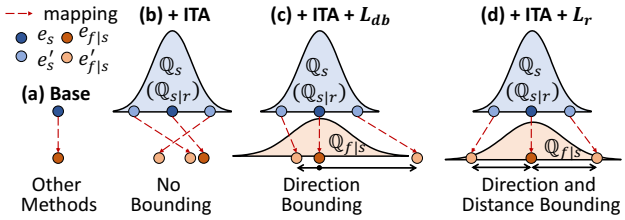

Figure 3: Diagram of augmentation effects of our proposed SADA ($+ITA$, $+ITA + L_{db}$, $ITA + L_r$).

We then define the solution of θ based on Empirical Risk Minimization (ERM) (Naumovich 1998) as:

$$\text{ERM: } \theta^*_{ITA} \in arg\min_{\theta \in \Theta} \frac{1}{k} \sum_{i=1}^{k} L(\theta, ITA(X_i)), \tag{3}$$

where Θ is defined as some parameter space. See detailed derivation based on ERM in Supplementary Materials A.1.

Proposition 3.3 (ITA **increases T2Isyn semantic consistency**). *Assume exact invariance holds. Consider an unaugmented text-image generator $\hat{\theta}(X)$ of G and its augmented version $\hat{\theta}_{ITA}$. For any real-valued convex loss $S(\theta, \cdot)$ that measures the semantic consistency, we have:*

$$\mathbb{E}[S(\theta, \hat{\theta}(X))] \geq \mathbb{E}[S(\theta, \hat{\theta}_{ITA}(X))], \tag{4}$$

which means with ITA, a model can have lower $\mathbb{E}[S(\theta, \hat{\theta}_{ITA}(X)]$ thus a better text-image consistency.

Proof. we obtain a direct consequence that: $Cov[\hat{\theta}_{ITA}(X)] \preceq Cov[\hat{\theta}(X)]$, where $Cov[\cdot]$ means the covariance matrix decreases in the Loewner order. Therefore, G with ITA can obtain better text-image consistency. See proof details in Supplementary Materials A.2. □

For a clear explanation, we specify a form $S(\theta, \cdot) := S(\theta, (\cdot, \cdot))$ where $(\cdot, \cdot)$ take a e_s and e_r for semantic consistency measuring, and θ denotes the set of training parameters. Since we preserve the semantics of e'_s, its generated images should also semantically match e_s. Thus, the total semantic loss of G is defined as:

$$\begin{aligned} L_S =& S(\theta, (e_s, \mathcal{G}(e_s))) + S(\theta, (e'_s, \mathcal{G}(e'_s))) \\ &+ S(\theta, (e_s, \mathcal{G}(e'_s))) + S(\theta, (e'_s, \mathcal{G}(e_s))) \ , \end{aligned} \tag{5}$$

where $\mathcal{G} = h(G(\cdot))$, $(\cdot)$ takes a textual embedding and $h(\cdot)$ maps images into semantic space. Typically, as the first term is included in the basic framework, it is omitted while other terms are added for SADA applications.

3.2 Obtaining Closed-from ITA_C

Theoretical Derivation of ITA_C Assume that exact invariance holds. We treat each textual semantic embedding e_s as a Gaussian-like distribution $\phi = \mathcal{N}(e_s, \sigma)$, where each sample $e'_s \sim \mathcal{N}(e_s, \sigma)$ can maintain the main semantic m_s of e_s. In other words, σ is the variation range of e_s conditioned by m_s, ϕ derives into:

$$\phi = \mathcal{N}(e_s, \sigma | m_s) \ . \tag{6}$$

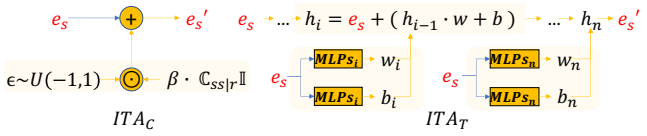

Figure 4: Network structure of ITA_C and ITA_T. Note that e_s and e'_s are equivalent to $e_{s|r}$ and $e'_{s|r}$ respectively.

By sampling e'_s from ϕ, we can efficiently obtain augmented textual embedding for training. We need to draw support from real images to determine the semantics m_s that need to be preserved. Empirically, real texts are created based on real images. e_s is thus naturally depending on e_r, leading to the inference: $e_{s|r} \triangleq e_s, m_{s|r} \triangleq m_s, Q_{s|r} \triangleq Q_s$. Given a bunch of real images, $\sigma|m_s$ is assumed to represent the level of variation inherent in text embeddings, conditioned on the real images. We can redefine ϕ in Eq. (6) for ITA_C augmentation as: $\phi \triangleq \mathcal{N}(e_{s|r}, \sigma|m_{s|r}) = \mathcal{N}(e_{s|r}, \beta \cdot \mathbb{C}_{ss|r}\mathbb{I})$, where $\mathbb{C}_{**}$ denotes covariance matrix of semantic embeddings; r, s stand for real images and real texts; $\mathbb{C}_{ss|r}$ is the self-covariance of e_s conditioned by semantic embedding of real images e_r; $\mathbb{I}$ denotes an identity matrix; β is a positive hyper-parameter for controlling sampling range. As such, we define: $\phi \triangleq Q_{s|r}$. According to (Kay 1993), conditional $\mathbb{C}_{ss|r}$ is equivalent to:

$$\mathbb{C}_{ss|r} = \mathbb{C}_{ss} - \mathbb{C}_{sr}\mathbb{C}_{rr}^{-1}\mathbb{C}_{rs} \ , \tag{7}$$

where all covariances can be directly calculated. Then ϕ is calculated from the dataset using semantic embeddings of texts and images for s and r. In practice, $\mathbb{C}_{ss|r}$ is calculated using real images and their given texts from the training set.

Remarks of ITA_C We explore the connections between ITA_C and previous methods (Dong et al. 2017; Cheng et al. 2020), assuming all models are well-trained.

Proposition 3.4. *ITA_C can be considered a closed-form solution for general textual semantic preserving augmentation methods of T2Isyn.*

Proof details can be seen in Supplementary Materials A.2. Therefore, training with bare ITA_C is equivalent to using other textual semantic preserving augmentation methods.

ITA_C Structure Based on Eq. (7), we obtain $e'_{s|r}$ from calculated ITA_C:

$$e'_{s|r} = e'_{s|r} \sim \phi = e_{s|r} + z \triangleq e_{s|r} + \epsilon \odot \beta \cdot \mathbb{C}_{ss|r}\mathbb{I}, \tag{8}$$

where $z \sim \mathcal{N}(0, \beta \cdot \mathbb{C}_{ss|r}\mathbb{I})$, ϵ is sampled from a uniform distribution $U(-1, 1)$, as shown in Figure 4. ITA_C requires no training and can be used to train or tune a T2Isyn model.

3.3 Obtaining Learnable ITA_T

We also design a learnable ITA_T as a clever substitute. Proposition 3.4 certifies that well-trained ITA_T is equivalent to ITA_C. To obtain ITA_T through training, we need to achieve the following objectives:

$$\max_{\alpha} L_d(\alpha, (e'_{s|r}, e_{s|r})), \ \min_{\alpha} S(\alpha, (e_{s|r}, \mathcal{G}(e'_{s|r}))) \ ,$$

where $L_d(\alpha, \cdot, \cdot)$ denotes a distance measurement, enforcing that the augmented $e'_{s|r}$ should be far from $e_{s|r}$ as much as possible; α is training parameters of ITA_T. $S(\alpha, (\cdot, \cdot))$ bounds the consistency between $e_{s|r}$ and generated images on $e'_{s|r}$, preserving the semantics of $e'_{s|r}$. The first objective can be easily reformed as minimizing the inverse distance:

$$\min_{\alpha} L_{id}(\alpha, (e'_{s|r}, e_{s|r})) := \min_{\alpha} -L_d(\alpha, (e'_{s|r}, e_{s|r})).$$

The final loss for training ITA_T is a weighted combination of L_d and $S(\alpha, (\cdot, \cdot))$:

$$\begin{aligned} L_{ITA_T} =& r \cdot L_{id}(\alpha, (e'_{s|r}, e_{s|r})) \\ &+ (1 - r) \cdot S(\alpha, (e_{s|r}, \mathcal{G}(e'_{s|r}))), \end{aligned} \tag{9}$$

where r is a hyper-parameter controlling the augmentation strength. Note that L_{ITA_T} is only used for optimizing α of ITA_T and parameters of G are frozen here (as Figure 2 (c)).

ITA_T Structure Since the augmented $e'_{s|r}$ should maintain the semantics in $e_{s|r}$, ϵ in Eq. (8) is maximized but does not disrupt the semantics in $e_{s|r}$. As such, ϵ is not a pure noise but a $e_{s|r}$-conditioned variable. Hence, Eq. (8) can be reformed as $e'_{s|r} = e_{s|r} + f(e_{s|r})$ to achieve ITA_T, where $f(e_{s|r})$ means a series of transformations of $e_{s|r}$. The final ITA_T process can be formulated as $e'_{s|r} = ITA_T(e_{s|r}) = e_{s|r} + f(e_{s|r})$. We deploy a recurrent-like structure as shown in Figure 4 to learn the augmentation. ITA_T takes $e_{s|r}$ as an input. For i^{th} step in overall n steps, there is a group of Multilayer Perceptrons to learn the weights w_i and bias b_i conditioned by $e_{s|r}$ for the previous module's output h_{i-1}. Then $h_i = e_{s|r} + (h_{i-1} \cdot w_i + b_i)$ will be output to the following processes. We empirically set $n = 2$ for all our experiments. ITA_T can be trained simultaneously with generative frameworks from scratch or used as a tuning trick.

4 Generated Image Semantic Conservation

Enabled by ITA's providing $e_{s|r}, e'_{s|r}$, we show that using Generated Image Semantic Conservation ($GisC$) will affect generated images' raw space. Consider a frozen pretrained image encoder (E_I) that maps images into the same semantic space. Consider a feasible and trainable generator G that learns how to generate text-consistent images: $G(X) \rightarrow \mathcal{F}$, $E_I(\mathcal{F}) \rightarrow \mathcal{E}$, where $\mathcal{F}$ and $\mathcal{E}$ are the sets for generated images f and their semantic embeddings e_f. Since images are generated on texts, we have $e_{f|s} \triangleq e_f$. We show that semantically constraining generated images can additionally affect their raw space.

Proposition 4.1. *Assume that E_I is linear and well-trained. Constraining the distribution $Q_{\mathcal{E}}$ of $e_{f|s}$ can additionally constrain the distribution $\mathcal{F}$ of f.*

Proof. There are two scenarios: 1) If E_I is inevitable, Proposition 4.1 is obvious. 2) If E_I is not inevitable, it is impossible that $\mathcal{F}$ all locates in the $Null(E_I)$ (nullspace of E_I) for well trained E_I, thus constraining $\mathcal{F}$ can affect $\mathcal{E}$. See more proof details in Supplementary Materials A.2. □

We further assume that the positive effeteness of feasible $GisC$ can pass to the raw generated image space. The non-linear case is non-trivial to proof. Our results of using non-linear encoders (DAMSM (Xu et al. 2018) and CLIP (Radford et al. 2021)) with different feasible $GisC$ methods suggest that Proposition 4.1 holds for non-linear E_I and positively affect image quality.

4.1 Image Semantic Regularization Loss

We design an Image Semantic Regularization Loss L_r to attain $GisC$ for preventing semantic collapse and providing tighter semantic constraints than direction bounding $\mathcal{L}_{db}$ (Gal et al. 2022).

Theoretical Derivation of L_r To tackle semantic collapse empirically, we constrain the semantic distribution of generated images, which draws inspiration from the principle of maximizing the information content of the embeddings through variance preservation (Bardes, Ponce, and LeCun 2021). Since semantic redundancies undescribed by texts in real images are not compulsory to appear in generated images, the generated images are not required to be the same as real images. Therefore, conditioned by the texts, generated images should obtain semantic variation in real images. For example, when text changes from 'orange' to 'banana', 'orange' in real images should likewise shift to 'banana' despite the redundancies, and fake images should obtain this variance (Tan et al. 2023). If exact invariance holds and the model is well-trained, the text-conditioned semantic distribution of its generated images $Q_{f|s} = \mathcal{N}(m_{f|s}, \mathbb{C}_{ff|s}\mathbb{I})$ should have the semantic variance as close as that of the real images $Q_{rr|s} = \mathcal{N}(m_{r|s}, \mathbb{C}_{rr|s}\mathbb{I})$:

$$\min_{e_f} ||\mathbb{C}_{ff|s}\mathbb{I} - \mathbb{C}_{rr|s}\mathbb{I}||^2, \mathbb{C}_{rr|s} = \mathbb{C}_{rr} - \mathbb{C}_{rs}\mathbb{C}_{ss}^{-1}\mathbb{C}_{sr} \ , \quad (10)$$

where $\mathbb{C}_{rr|s}$ is the self-covariance of e_r conditioned by real text embeddings.

Aim to maintain latent space alignment, an existing $GisC$ method, direction bonding (Gal et al. 2022) is defined as:

$$L_{db} = 1 - \frac{(e'_{s|r} - e_{s|r}) \cdot (e'_{f|s} - e_{f|s})}{||(e'_{s|r} - e_{s|r})||^2 \cdot ||(e'_{f|s} - e_{f|s})||^2} \ . \quad (11)$$

L_{db} follows that semantic features are usually linearized (Bengio et al. 2013; Upchurch et al. 2017; Wang et al. 2021).

Given a pair of encoders that maps texts and images into the same semantic space, inspired by L_{db}, we assume that:

Assumption 4.2. *If the paired encoders are well-trained, aligned, and their semantic features are linearized. The semantic shifts images are proportional to texts:*

$$(e'_{f|s} - e_{f|s}) \propto (e'_{s|r} - e_{s|r}). \quad (12)$$

Assumption 4.2 holds for T2Isyn intuitively because when given textual semantics changes, its generated image's semantics also change, whose shifting direction and distance are based on textual semantics changes. Otherwise, semantic mismatch and collapse would happen. If Assumption 4.2 holds, based on ITA_C that preserves $e'_{s|r} - e_{s|r}$, we have:

$$\begin{aligned} e'_{f|s} - e_{f|s} &\leq \epsilon \odot \beta \cdot d(\mathbb{C}_{ff|s}) \\ \text{s.t. } e'_{s|r} - e_{s|r} &\leq \epsilon \odot \beta \cdot d(\mathbb{C}_{ss|r}) \ . \end{aligned} \quad (13)$$

If we force that each dimension of ${\epsilon^*}_{i=1}^{d} \sim \{-1, 1\}$ where $d = \{1, ..., n\}$ and n is the dimension of the semantic embedding, we have:

$$\begin{aligned} e''_{f|s} - e_{f|s} &= \epsilon^* \odot \beta \cdot d(\mathbb{C}_{ff|s}) \\ \text{s.t. } e''_{s|r} - e_{s|r} &= \epsilon^* \odot \beta \cdot d(\mathbb{C}_{ss|r}) \ . \end{aligned} \quad (14)$$

Derived form Eqs. (10) and (14), we define our Image Semantic Regularization Loss L_r as:

$$L_r = \varphi \cdot || \ (e''_{f|s} - e_{f|s}) - \epsilon^* \odot \beta \cdot d(\mathbb{C}_{rr|s})||^2 \ , \quad (15)$$

where $\beta \cdot d(\mathbb{C}_{ff|s})$ can be considered a data-based regularized term. ϵ constrains the shifting direction, as shown in Figure 3 (d). φ is a hyper-parameter for balancing L_r with other loss. Note that for ITA_T, the range of $e'_{s|r} - e_{s|r}$ is not closed-form. Thus, we cannot apply L_r with ITA_T.

Remarks of L_r We show the effect of L_r on the semantic space of generated images:

Proposition 4.3 (L_r **prevent semantic collapse: completely different**). *L_r leads to $|e'_{f|s} - e_{f|s}|$ is less than or equal to a sequence Λ of positive constants, further constrains the semantic manifold of generated embeddings to meet the Lipschitz condition.*

Proof. From Eq. (15), we have the constraint $|e'_{f|s} - e_{f|s}| \leq \Lambda$. Therefore, we have: $\frac{|e'_{f|s} - e_{f|s}|}{|e'_{s|r} - e_{s|r}|} \leq K$, s.t. $e'_{s|r} \neq e_{s|r}$, where K is a Lipschitz constant. See more proof details in Supplementary Materials A.2. □

Proposition 4.3 justifies why image quality can be improved with L_r. According to Proposition 4.1, we believe that the Lipschitz continuity can be passed to visual feature distribution, leading to better continuity in visual space as well. Our experiments verify that with L_r methods, T2Isyn models achieve the best image quality.

Proposition 4.4 (L_r **prevent semantic collapse: extremely similar**). *L_r prevents $|e''_{f|s} - e_{f|s}| = 0$ and provides tighter image semantic constraints than direction bounding L_{db}.*

Proof. For Eq. (11), assume $L_{db} = 0$ and use $e''_{s|r}$ to substitute $e_{s|r}$, combining with Eq. (8), we have: $|e''_{f|s} - e_{f|s}| \geq 0$. Preservation of semantic collapse is not guaranteed due to the distance between $e''_{f|s}$ ($e'_{f|s}$) and $e_{f|s}$ is not strictly contained. Assume $L_r = 0$, we have: $|e''_{f|s} - e_{f|s}| > 0$, where provides tighter constraints than L_{db}. See visual explanation in Figure 3 (c)(d) and proof details in Supplementary Materials A.2. □

Propositions 4.3-4.4 show that L_r prevents semantic collapse. See SADA' algorithms in Supplementary Materials B.

	Image Retrieval		Text Retrieval	
	Top1	Top5	Top1	Top5
CLIP	30.40	54.73	49.88	74.96
Tuned	44.43	72.38	61.20	85.16
+ITA	**44.88**(+0.45)	**72.42**(+0.04)	**62.76**(+1.56)	**85.38**(+0.22)

Table 1: Text-Image Retrieval results of CLIP tune w/ and wo/ SADA ITA. Please refer to Supplementary Material D.1 for tuning CLIP with different numbers of samples.

5 Experiments

Our experiments include three parts: 1) To demonstrate how ITA improves text-image consistency, we apply ITA of SADA to Text-Image Retrieval tasks. 2) To exhibit the feasibility of our SADA, we conduct extensive experiments by using different T2Isyn frameworks with GANs, Transformers, and Diffusion Models (DM) as backbones on different datasets. 3) Detailed ablation studies are performed; we compare our SADA with other typical augmentation methods to show that SADA certifies an improvement in text-image consistency and image quality in T2Isyn tasks. Particularly noteworthy is the observation that $GisC$ can alleviate semantic collapse. Due to page limitations, key findings are presented in the main paper. For detailed application and training information, as well as more comprehensive results and visualizations, please refer to Supplementary Materials C and D. Codes are available at https://github.com/zhaorui-tan/SADA.

5.1 SADA on Text-Image Retrieval

Experimental setup We compare tuning CLIP (Wang et al. 2022)(ViT-B/16) performance w/ ITA and wo/ ITA on the COCO (Lin et al. 2014) dataset. Evaluation is based on Top1 and Top5 retrieval accuracy under identical hyperparameter settings.

Results As exhibited in Table 1, using ITA results in a boost in image-text retrieval accuracy in both the Top1 and Top5 rankings, reflecting its proficiency in enhancing the consistency between text and images. The increase of 0.45% and 1.56% in Top1 retrieval accuracy explicitly suggests a precise semantic consistency achieved with SADA, providing empirical validation to our Proposition 3.3.

5.2 SADA on Various T2Isyn Frameworks

Experimental setup We test SADA on GAN-based AttnGAN (Xu et al. 2018) and DF-GAN (Tao et al. 2022), transformer-based VQ-GAN+CLIP (Wang et al. 2022), vanilla DM-based conditional DDPM (Ho, Jain, and Abbeel 2020) and Stable Diffusion (SD) (Rombach et al. 2021) with different pretrianed text-image encoders (CLIP and DAMSM (Xu et al. 2018)). Parameter settings follow the original models of each framework for all experiments unless specified. Datasets CUB (Wah et al. 2011), COCO (Lin et al. 2014), MNIST, and Pokémon BLIP (Deng 2012) are employed for training and tuning (see the 2^{nd} column in Table 2 for settings). Supplementary Material D.2 offers additional SD-tuned results. For qualitative evaluation, we use CLIPScore (CS) (Hessel et al. 2021) to assess text-image consistency (scaled by 100) and Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate image quality (computed over 30K generated images).

Backbone	Encoder, Method Settings, Dataset		CS↑	FID↓
Transformer	CLIP	VQ-GAN+CLIP	62.78	16.16
+SADA	Tune	COCO	**62.81**	**15.56**
DM	CLIP	SD	72.72	55.98
+SADA	Tune	Pokémon BLIP	**73.80**	**46.07**
DM	CLIP	DDPM	70.77	8.61
+SADA	Train	MNIST	**70.91**	**7.78**
GANs	DAMSM	AttnGAN	68.00	23.98
+SADA	Train	CUB	**68.20**	**13.17**
GANs	DAMSM	AttnGAN	62.59	29.60
+SADA	Tune	COCO	**64.59**	**22.70**
GANs	DAMSM	DF-GAN	58.10	12.10
+SADA	Train	CUB	**58.24**	**10.45**
GANs	DAMSM	DF-GAN	50.71	15.22
+SADA	Train	COCO	**51.02**	**12.49**

Table 2: Performance evaluation of SADA with different backbones with different datasets. Results better than the baseline are in bold.

Results As shown in Table 2 and corresponding Figure 6, the effectiveness of our SADA can be well supported by improvements across all different backbones, datasets, and text-image encoders, which experimentally validate the efficacy of SADA in enhancing text-image consistency and image quality. Notably, facilitated by $ITA_C + L_r$, AttnGAN achieves 13.17 from 23.98 on CUB. For tuning VQ-GAN+CLIP and SD that have been pre-trained on large-scale data, SADA still guarantees improvements. These results support Propositions 3.3, 4.1 and 4.3. It's worth noting that the tuning results of models with DM backbones (SD) are influenced by the limited size of the Pokémon BLIP dataset, resulting in a relatively high FID score. Under these constraints, tuning with SADA performed better than the baseline, improving the CS from 72.72 to 73.80 and lowering the FID from 55.98 to 46.07.

5.3 Ablation Studies

Experimental setup Based on AttnGAN and DF-GAN, we compare Mixup (Zhang et al. 2017a), DiffAug (Zhao et al. 2020), Random Mask (RandMask), Add Noise, with SADA components in terms of CS and FID. Refer to Supplementary Materials C, D.3 for more detailed settings and the impact of r in ITA_T.

Quantitative results Quantitative results are reported in Table 3.[3] We discuss the results from different aspects.

1). Effect of other competitors: Mixup and DiffAug weaken visual supervision, resulting in worse FID than baselines. They also waken text-image consistency under most situations. Moreover, Random Mask and Add Noise are sen-

[3]Note for task 2, we use the best results among current augmentations as the baseline since no released checkpoint is available.

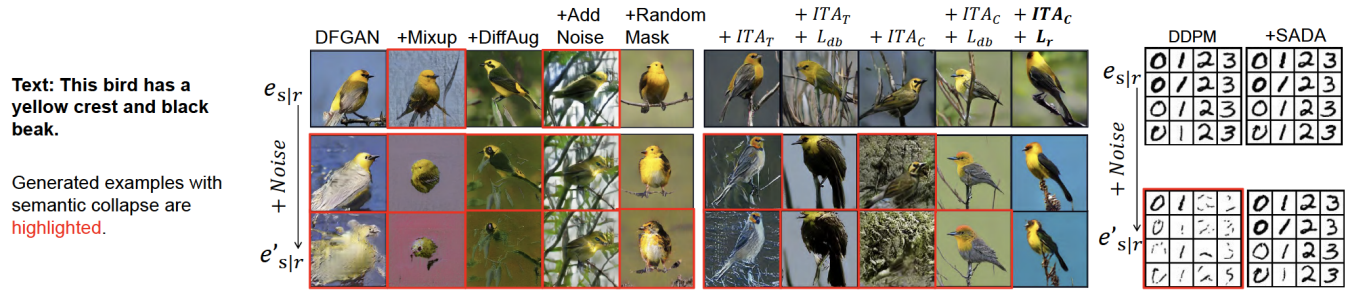

Figure 5: Generated examples of DF-GAN and DDPM trained with different augmentations on $e_{s|r}$ as ascending $Noise \sim \mathcal{N}(0, \beta \cdot \mathbb{C}_{ss|r}\mathbb{I})$ is given. Input noise is fixed for each column. See full examples in Supplementary Materials Figures 18, 19 & 20.

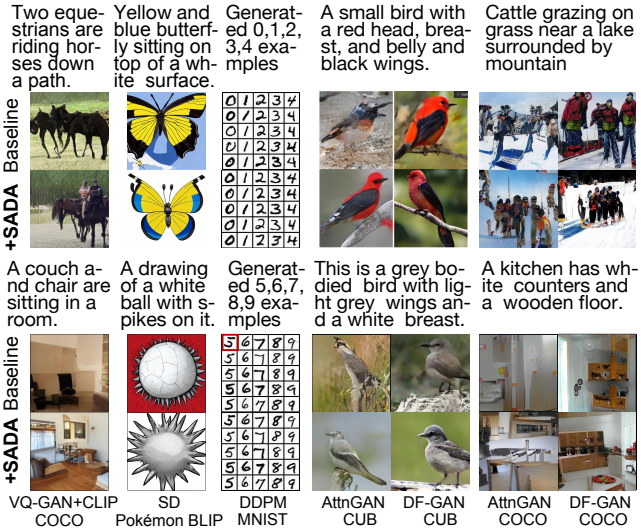

Figure 6: Generated examples of different backbones with different datasets wo/ SADA and w/ SADA. See more examples of different frameworks in Supplementary Materials D.

	AttnGAN		DF-GAN			
Settings	Task 1: Train		Task 2: Train		Task 3: Train	
CUB	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓
Paper	68.00*	23.98*	-	14.81*	-	-
RM	68.00	23.98	-	14.81	58.10*	12.10*
+Mixup	65.82	41.47	57.29	28.73	57.36	25.77
+DiffAug	66.94	**22.53**	58.22	17.27	58.05	12.35
+RandMask	67.80	**15.59**	57.96*	15.42	58.07	15.17
+Add Noise	67.79	**17.29**	57.46	48.23	57.58	42.07
$+ITA_T$	**68.53**†	**14.14**	**58.09**	**14.03**	**58.80**†	**12.17**
$+ITA_T+L_{db}$	**68.10**	**14.55**	**58.07**	<u>11.74</u>	<u>58.67</u>	**11.58**
$+ITA_C$	<u>68.42</u>	**13.68**	**58.25**	**12.70**	**58.23**	**11.81**
$+ITA_C+L_{db}$	**68.18**	<u>13.74</u>	**58.30**†	**12.93**	**58.23**	<u>10.77</u>
$+ITA_C+L_r$	**68.20**	**13.17**†	<u>58.27</u>	**11.70**†	**58.24**	**10.45**†
Settings	Task 4: Tune		Task 5: Tune		Task 6: Tune	
COCO	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓
Paper	50.48	35.49	-	19.23	-	-
RM	50.48	35.49	50.94	15.41	50.94	15.41
+ Tuned	62.59*	29.60*	50.63*	15.67*	50.71*	15.22*
+Mixup	62.30	33.41	50.38	23.80	**50.83**	22.86
+DiffAug	**65.44**	33.86	49.45	21.31	**50.94**	18.97
+RandMask	**63.76**	<u>23.82</u>	50.54	15.74	50.64	15.33
+Add Noise	**64.77**†	35.47	**50.94**†	34.90	**50.80**	33.84
$+ITA_T+L_{db}$	**63.31**	**26.65**	50.60	**15.05**	**50.77**	**13.67**
$+ITA_C+L_{db}$	**63.97**	**25.82**	<u>50.92</u>	<u>14.71</u>	<u>50.98</u>	<u>13.28</u>
$+ITA_C+L_r$	<u>64.59</u>	**22.70**†	**50.81**	**13.71**†	**51.02**†	**12.49**†

Table 3: CS↑ and FID↓ for AttnGAN, and DF-GAN with Mixup, Random Mask, Add Noise, and the proposed SADA components on CUB and COCO. *: Baseline results; Bold: Results better than the baseline; †: Best results; Underlines: Second best results; 'RM': Released Model; 'e': epochs.

sitive to frameworks and datasets, thus they cannot guarantee consistent improvements.

2). ITA improves text-image consistency: Regarding text-image consistency, using ITA wo/, or w/ $GisC$ all lead to improvement in semantics, supporting Proposition 3.3. However, ITA_T consumes more time to converge due to its training, weakening its semantic enhancement at the early stage (as in Task 5). As it converged with longer training time, ITA_T improves text-image consistency as in Task 6.

3). $GisC$ promotes image quality: For image quality, it can be observed that using bare ITA wo/ $GisC$, FID is improved in most situations; but using constraints such as L_{db} and L_r with ITA_T and ITA_C can further improve image quality except $ITA_T + L_{db}$ in Task 1. These support our Proposition 4.1 and Proposition 4.3.

4). L_r provides a tighter generated images semantic constraint than L_{db}: Specifically, compared with L_{db}, using our proposed L_r with ITA_C provides the best FID and is usually accompanied by a good text-image consistency, thus validating our Proposition 4.4.

Qualitative Results As depicted in Figure 5 and further examples in Supplementary Materials D, we derived several key insights.

1). Semantic collapse happens in the absence of a sufficient $GisC$**:** As seen in Figure 5, neither non-augmented nor other augmented methods fail to prevent semantic collapse in different backbones. The application of $GisC$ through SADA serves to alleviate this issue effectively.

2). ITA **preserves textual semantics:** It shows that generated images of models wo/ ITA on $e'_{s|r}$ still maintain the main semantics of $e_{s|r}$ though they have low quality, indi-

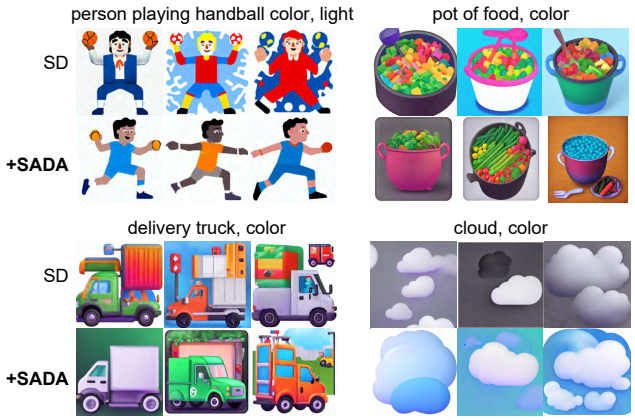

Figure 7: Generated examples of SD tuned on the Emoji dataset wo/ and w/ SADA. A significant improvement in diversity with $+ITA_C + L_r$ can be observed, especially in terms of skin color and view perspective.

cating the textual semantic preservation of ITA.

3). SADA enhances generated image diversity: SADA appears to improve image diversity when input noise is not fixed significantly and $e_{s|r}$ of testing text is used. The greatest improvement in image diversity was achieved by ITA_C+L_r, as the detailed semantics of birds, are more varied than the other semantics. Textual unmentioned details such as skin colors as shown in Figure 7 is more various when using SADA. More textual unmentioned details can be observed in Supplementary Materials Figure 11 (highlighting wing bars, color, and background).

4). ITA with $GisC$ improves the model generalization by preventing semantic collapse: Using $ITA_T + L_{db}$ and ITA_C+L_{db}/L_r lead to obvious image quality improvement when more $Noise$ is given, corresponding to our Proposition 4.1 and Proposition 4.3. However, with $ITA_C + L_{db}$, though the model can produce high-quality images, generated images on $e_{s|r}$ and $e'_{s|r}$ are quite similar while $ITA_C + L_r$ varies a lot, especially in the background, implying a not guaranteed semantic preservation of L_{db} and a tighter constraint of L_r as proved in Proposition 4.4. Furthermore, $ITA_C + L_r$ provides the best image quality across all experiments.

5.4 SADA on Complex Sentences and Simple Sentences

We also notice that semantic collapse is more severe when a complex description is given. Applying SADA alleviates the semantic collapse across all descriptions. We explore the effect of SADA on complex sentences and simple sentences. We use textual embeddings of sentences in Table 4 and illustrate interpolation examples at the inference stage between $e_{s|r}$ and $e'_{s|r}$ as shown in Figure 8 right side, where $Noise \sim \mathcal{N}(0, \beta \cdot \mathbb{C}_{ss|r}\mathbb{I})$. It can be observed that models trained with SADA can alleviate the semantic collapse that occurs in models without SADA, and its semantics can resist even larger $Noise$ given. Using $e'_{s|r}$ at the inference stage

sent1	this is a yellow bird with a tail.
sent2	this is a small yellow bird with a tail and gray wings with white stripes.
sent3	this is a small yellow bird with a grey long tail and gray wings with white stripes.

Table 4: Rough, detailed, and in-between description used for generation.

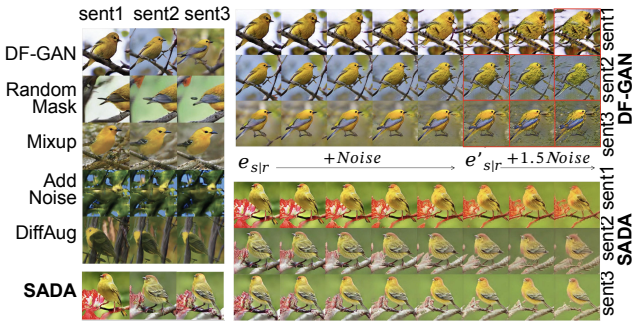

Figure 8: Left: Generated results of DF-GAN with different methods on rough to detailed sentences. Right: Interpolation examples at the inference stage between $e_{r|s}$ and $e'_{r|s}$ of DF-GAN and it with SADA on rough to detailed sentences. $e'_{r|s}$, input noise for generator G, and textual conditions are the same across all rows. Examples of significant collapse are highlighted in red.

can cause image quality degradation, which reveals the robustness of the models.

As shown in Figure 8, on the left side, DF-GAN with SADA generates more text-consistent images with better quality from rough to precise descriptions compared to other augmentations. The Right side indicates that DF-GAN without augmentations experiences semantic collapse when larger $Noise$ is given. The semantic collapse is more severe when a complex description is given. Applying SADA alleviates the semantic collapse across all descriptions. The model with SADA can generate reasonably good and text-consistent images when the $1.5Noise$ with complex description is given. These visualizations further verified the effectiveness of our proposed SADA.

6 Conclusion

In this paper, we propose a Semantic-aware Data Augmentation framework (SADA) that consists of ITA (including ITA_T and ITA_C) and L_r. We theoretically prove that using ITA with T2Isyn models leads to text-image consistency improvement. We also show that using $GisC$ can improve generated image quality, and our proposed $ITA_C + L_r$ promotes image quality the most. ITA relies on estimating the covariance of semantic embeddings, which may, however, be unreliable in the case of unbalanced datasets. We will explore this topic in the future.

Acknowledgments

The work was partially supported by the following: National Natural Science Foundation of China under No. 92370119, No. 62376113, and No. 62206225; Jiangsu Science and Technology Program (Natural Science Foundation of Jiangsu Province) under No. BE2020006-4; Natural Science Foundation of the Jiangsu Higher Education Institutions of China under No. 22KJB520039.

References

Bardes, A.; Ponce, J.; and LeCun, Y. 2021. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Bengio, Y.; Mesnil, G.; Dauphin, Y.; and Rifai, S. 2013. Better mixing via deep representations. In *International Conference on Machine Learning*, 552–560. PMLR.

Chen, S.; Dobriban, E.; and Lee, J. H. 2020. A group-theoretic framework for data augmentation. *The Journal of Machine Learning Research*, 21(1): 9885–9955.

Cheng, J.; Wu, F.; Tian, Y.; Wang, L.; and Tao, D. 2020. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10911–10920.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6): 141–142.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34.

Dong, H.; Zhang, J.; McIlwraith, D.; and Guo, Y. 2017. I2t2i: Learning text to image synthesis with textual data augmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*, 2015–2019. IEEE.

Ermolov, A.; Siarohin, A.; Sangineto, E.; and Sebe, N. 2021. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, 3015–3024. PMLR.

Gal, R.; Patashnik, O.; Maron, H.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10696–10706.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Kay, S. M. 1993. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer.

Liu, P.; Wang, X.; Xiang, C.; and Meng, W. 2020. A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, 191–195. IEEE.

Naumovich, V. 1998. *Statistical learning theory*. Johm Wiley.

Naveed, H. 2021. Survey: Image mixing and deleting for data augmentation. *arXiv preprint arXiv:2106.07085*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, 1060–1069. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ruan, S.; Zhang, Y.; Zhang, K.; Fan, Y.; Tang, F.; Liu, Q.; and Chen, E. 2021. DAE-GAN: Dynamic aspect-aware GAN for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13960–13969.

Tan, Z.; Yang, X.; Ye, Z.; Wang, Q.; Yan, Y.; Nguyen, A.; and Huang, K. 2023. Semantic Similarity Distance: Towards better text-image consistency metric in text-to-image generation. *Pattern Recognition*, 144: 109883.

Tao, M.; Tang, H.; Wu, F.; Jing, X.-Y.; Bao, B.-K.; and Xu, C. 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16515–16525.

Upchurch, P.; Gardner, J.; Pleiss, G.; Pless, R.; Snavely, N.; Bala, K.; and Weinberger, K. 2017. Deep feature interpolation for image content changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7064–7073.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, Y.; Huang, G.; Song, S.; Pan, X.; Xia, Y.; and Wu, C. 2021. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Z.; Liu, W.; He, Q.; Wu, X.; and Yi, Z. 2022. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP. *arXiv preprint arXiv:2203.00386*.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1316–1324.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017a. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017b. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 5907–5915.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1947–1962.

Zhao, S.; Liu, Z.; Lin, J.; Zhu, J.-Y.; and Han, S. 2020. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33: 7559–7570.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5802–5810.