SSD: Towards Better Text-Image Consistency Metric in Text-to-Image Generation

Zhaorui Tan[1], Xi Yang[1,†], Zihan Ye[1], Qiufeng Wang[1], Yuyao Yan[1], Anh Nguyen[2], and Kaizhu Huang[3,†]

Abstract—Generating consistent and high-quality images from given texts is essential for visual-language understanding. Although impressive results have been achieved in generating high-quality images, text-image consistency is still a major concern in existing GAN-based methods. Particularly, the most popular metric R-precision may not accurately reflect the text-image consistency, often resulting in very misleading semantics in the generated images. Albeit its significance, how to design a better text-image consistency metric surprisingly remains under-explored in the community. In this paper, we make a further step forward to develop a novel CLIP-based metric termed as Semantic Similarity Distance (SSD), which is both theoretically founded from a distributional viewpoint and empirically verified on benchmark datasets. Benefiting from the proposed metric, we further design the Parallel Deep Fusion Generative Adversarial Networks (PDF-GAN) that aims at improving text-image consistency by fusing semantic information at different granularities and capturing accurate semantics. Equipped with two novel plug-and-play components: Hard-Negative Sentence Constructor and Semantic Projection, the proposed PDF-GAN can mitigate inconsistent semantics and bridge the text-image semantic gap. A series of experiments show that, as opposed to current state-of-the-art methods, our PDF-GAN can lead to significantly better text-image consistency while maintaining decent image quality on the CUB and COCO datasets.

Index Terms—Text-to-image, Image Generation, Generative Adversarial Networks

I. Introduction

Generating images from text descriptions, usually known as Text-to-Image Generation (T2I), is a challenging task that requires both generating high-quality images and maintaining text-image consistency. Although Generative Adversarial Networks (GANs) based methods [1]–[7] have achieved impressive results in generating high-quality images from text descriptions, they still struggle to keep the text-image consistency within complex semantics. Once the text descriptions become more complex, the semantics of the generated image will likely mismatch its text, although it comes with a high-quality score. Thus, the measurement of text-image consistency remains a critical concern in T2I generation.

Albeit its significance, how to design a better text-image consistency metric surprisingly remains under-explored in the community. To illustrate it clearly, we conduct a simple ex-

Fig. 1: Illustration of synthesized images with CFID, CS, R, and SSD score for Attn-GAN, DM-GAN, DAE-GAN, DF-GAN, and our PDF-GAN. Better metric scores are highlighted with higher saturation. SSD demonstrates much better text-image consistency than R, CS and CFID.

periment for current potential text-image consistency metrics in Fig. 1.[1]

The most widely used T2I synthesis metric, R-precision (R) [8], judges text-image consistency by evaluating if the generated image is more consistent with the given text than the other 99 randomly sampled texts. Such a measure may not accurately reflect the direct consistency between texts and images. The right-most column of Fig. 1 shows that the R score on ground truth (GT) is even seriously worse than synthetic images.[2] It results in very misleading semantics in the generated image. As highlighted by rectangles in Fig. 1, DAE-GAN achieves the highest R, but still produces very text-inconsistent images. Meanwhile, random sampling may also be highly biased by datasets. The more significant variation between the sampled texts and the given descriptions leads to significantly better scores on more diverse datasets (i.e., COCO [9] v.s. CUB [10] in Fig. 1 (R)).

Semantic Object Accuracy (SOA) [11], one recently-proposed metric specifically designed for evaluating multi-object text-image consistency, would still fail to measure the entire semantic consistency without evaluating object attributes and relationships. More seriously, SOA cannot be applied to

---

[1] Zhaorui Tan, Xi Yang, Zihan Ye, Qiufeng Wang, and Yuyao Yan are with Xi'an Jiaotong-Liverpool University.

[2] Anh Nguyen is with University of Liverpool.

[3] Kaizhu Huang is with Duke Kunshan University.

†: Corresponding authors: Xi Yang (Xi.Yang01@xjtlu.edu.cn) and Kaizhu Huang (kaizhu.huang@dukekunshan.edu.cn)

[1] SSD, CS and CFID in this paper are scaled by 100 for better readability. SOA is omitted since it cannot be applied to CUB.

[2] We use the 99 random samples that can reproduce the R in AttnGAN [8] to calculate R on GT pairs.
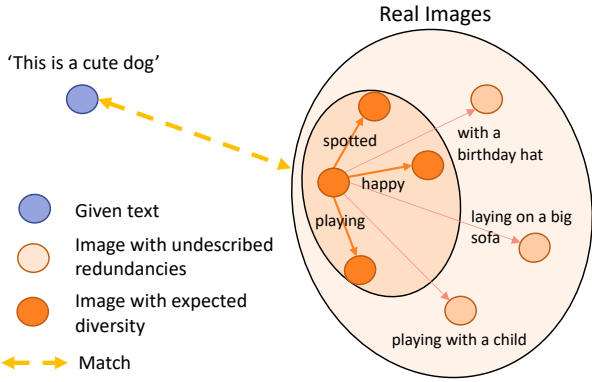
Fig. 2: Diagram of semantic redundancies. Real images often obtains semantics that may not have been described by given texts. This naturally causes semantic gap.
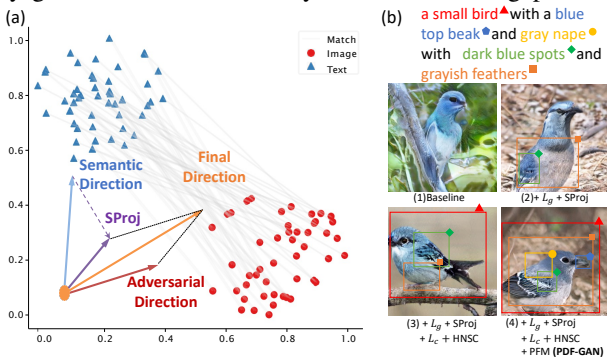

Fig. 3: (a) T-SNE map of CLIP embeddings for randomly sampled GT text-image pairs in the CUB dataset. (b) Examples generated of different components are added.

datasets where only one object usually appears in the generated images, such as CUB.

To alleviate these issues, researchers have to rely on Human Evaluation [7]. However, the process is usually costly, and its settings vary a lot among different methods, making it harder to apply in practical scenarios.

In this work, we make a further step forward and propose a novel CLIP-based text-image consistency metric from a distributional viewpoint, termed Semantic Similarity Distance (SSD). For T2I synthesis tasks, CLIP offers a joint language-vision embedding space where the similarity between semantic distributions of images and text can be directly measured. Our SSD is designed by combining two terms: 1) the first-moment term measures directly the text-image semantic similarity, reflecting the semantic bias between generated images and texts; 2) the second-moment term evaluates the difference of semantic variation between synthesized and real images conditioned on texts, suggesting that the diversity of semantics in the generated images should also be consistent with that of the real images. The second term can bring more credit to precise semantics, balancing the evaluation between overall and detailed consistency. Moreover, due to the large-scale pre-trained CLIP, SSD alleviates the bias in datasets and can be compared across different datasets.

On the theoretical front, we show that SSD's rationale is rooted in using a modified Wasserstein Distance for measuring the divergence of two distributions. We also show that it can be closely linked with two recent metrics, CLIPScore (CS) [12] and Conditional Frechet Inception Distanc (CFID) [13], but exhibits more desirable properties in measuring the semantic consistency. CS directly evaluates the similarity between images and texts' CLIP [14] embeddings, merely accounting for the first term in SSD. As shown in Fig. 1, generated images may even achieve better CS than GT, partially showing its limitation as a metric. Meanwhile, we show that CFID inappropriately measures the distance between text-conditioned real and fake image distributions, which is seriously affected by semantic redundancies (i.e. the semantics not specified by given texts as shown in Fig. 2) in real images. Fig. 1 also illustrates its incapability of measuring text-image consistency.

With the benefits of SSD, there are two findings as follows: 1) Different levels of semantic information can significantly help with text-image consistency. However, the semantic gap [15] will cause optimization conflicts between adversarial loss and semantic perceptual loss [8], as shown in Fig. 3 (a). As such, brutally adding semantic perceptual loss weakens the semantic supervision, leading to a sub-optimal performance in text-image consistency. 2) The mismatched samples for discrimination usually utilize shifted samples in a batch or random samples from other classes, which may lead to degradation of text-image consistency, especially in case of contrastive losses.

According to the above findings, we propose a novel one-stage T2I generation framework named PDF-GAN as Fig. 5, consisting of Parallel Fusion Modules (PFMs) with semantic perceptual losses to fuse different-level granularity textual data. To further improve text-image consistency, we design two novel plug-and-play modules: Hard-Negative Sentence Constructor (HNSC) and Semantic Projection (SProj). HNSC constructs stable and controllable hard negative textural samples instead of sampling mismatched textual samples from the dataset to alleviate dataset bias. SProj constrains the optimization direction of semantic loss, projecting it to the direction that does not conflict with the adversarial loss to overcome the semantic gap. As Fig. 3 (b) shows, our PDF-GAN with HNSC and SProj significantly improves text-image consistency while maintaining decent image quality. Our code is available at https://github.com/zhaorui-tan/PDF-GAN/tree/main.

Our contributions are summarized threefold:

- We introduce a novel metric, Semantic Similarity Distance, which evaluates both text-image similarity and semantic variation difference between generated images and real images conditioned by the texts. SSD is theoretically well founded and can be cross-compared on different datasets.
- We propose a novel framework, Parallel Deep Fusion Generative Adversarial Networks (PDF-GAN), with semantic perceptual losses and PFMs to fuse semantic information at different levels.
- We design an HNSC that mines hard negative textual samples and SProj that alleviates the semantic gap and enhances text-image consistency.

II. RELATED WORK

A. GANs for T2I

To improve the image quality and size in the first GAN-based approach [16], most methods adopt a multi-stage architecture for a coarse-to-fine generating process [8], [17]–[20]. The attention mechanism [8], [21]–[24] and extra networks [20], [25]–[29] are frequently applied to emphasize the semantics. DF-GAN [7] proposes a one-stage backbone with Deep Fusion Blocks using Affine and a one-way discriminator output with Matching-Aware Gradient Penalty. It avoids entanglements between generators without using the computationally expensive cross attention.

B. Contrastive Language-Image Pre-training

CLIP [14] is a large-scale multi-modal pre-training model that maps images and language to a joint latent space, aligning them by maximizing their Cosine similarity. CLIP has been widely used as pre-trained encoders for T2I GAN-based models [30], [31], transformer-based generators [32], and diffusion models [33].

C. Text-Image Consistency Metrics

Built upon the Cosine similarity between image and text embeddings, the widely-used metric R [8] evaluates if the generated image is more similar to the GT texts than random samples from the dataset. R does not measure directly the semantic consistency, which may be highly biased by the dataset. To this end, SOA [11] uses a pre-trained object detection model to evaluate whether an object mentioned by text exists in the generated image. Failing to measure the entire semantic consistency, SOA cannot be applied to datasets where only one object appears in the generated images (e.g. CUB). Owing to CLIP's popularity, CS [12] is designed for image captioning, yet Cosine similarity of CLIP embeddings may not explicitly bind attributes to objects and neglects semantic variations [33]. With the conditional distribution, CFID [13] evaluates the distance between text-conditioned fake and real image distributions. However, directly aligning fake and real distributions may mismatch the redundant parts in real images, i.e. those contents not specified by texts. This severely affects CFID's effect in measuring text-image consistency.

III. SEMANTIC SIMILARITY DISTANCE

In this section, we set out our novel metric for quantitatively evaluating T2I models. For better measuring the text-image consistency, our metric evaluates not only direct text-image semantic similarity but also the semantic variation difference between synthesized and real images conditioned on texts. From a distributional perspective, we assume that normalized embeddings of generated image $\tilde{e}_f$, real image $\tilde{e}_r$, and text $\tilde{e}_s$ distributions are all Gaussian-like distributions Φ in a joint language-vision embedding space (CLIP space):

$$\mathbb{Q}_f = \Phi(m_f, \mathbb{C}_{ff}) \, , \; \mathbb{Q}_r = \Phi(m_r, \mathbb{C}_{rr}) \, , \; \mathbb{Q}_s = \Phi(m_s, \mathbb{C}_{ss}) \, ,$$

where m and $\mathbb{C}$ denote the mean and covariance; f, r and s mean the generated images, real images, and texts,

respectively. Conditioned on the same text s, the generated and real images's conditional distribution, $\mathbb{Q}_{f|s}, \mathbb{Q}_{r|s}$, are given as:

$$\mathbb{Q}_{f|s} = \Phi(m_{f|s}, \mathbb{C}_{ff|s}) \, , \mathbb{Q}_{r|s} = \Phi(m_{r|s}, \mathbb{C}_{rr|s}) \, ,$$

where $\mathbb{C}_{ff|s}$ and $\mathbb{C}_{rr|s}$ represent conditional covariances of $\tilde{e}_f$ and $\tilde{e}_r$ that are constant and independent of condition $\tilde{e}_s$. We are now ready to define our Semantic Similarity Distance.

Definition 1. *As the ultimate goal is to measure the semantic distance among $\tilde{e}_f$ and $\tilde{e}_s$, we consider the distance between Q_f, and Q_s, and the distance between $Q_{ff|s}$ and $Q_{rr|s}$. Our SSD is then defined as follows:*

$$SSD(\mathbb{Q}_f, \mathbb{Q}_s, \mathbb{Q}_{f|s}, \mathbb{Q}_{r|s}) = \underbrace{[1 - cos(m_f, m_s)]}_{SS} + \quad (1)$$

$$\underbrace{||d(\mathbb{C}_{ff|s}) - d(\mathbb{C}_{rr|s})||^2}_{dSV} \, , \quad (2)$$

where $d(\cdot)$ represents matrix's the diagonal part. SS and dSV represents the first and second term of SSD respectively.

Since a pre-trained CLIP model is used to map the image and text to a joint language-vision embedding space, it is intuitive to measure their embeddings' Cosine distance, as done in the first-moment term of Eq. (1). Due to the semantic gap between Q_f and Q_s, solely measuring the Cosine distance cannot fully reflect the distribution divergence. We then take $Q_{f|s}$ and $Q_{r|s}$ into consideration to bridge the semantic gap. We argue that if a model can fully capture semantics, its generated images should share the same semantic variation as the real images. Semantic variation can also help bind objects and attributes, leading to more precise semantic alignment. Note that we do not align $Q_{f|s}$ and $Q_{r|s}$ directly because it over-concerns the redundancies that are not described by the text. Therefore, we design a second-moment term in Eq. (1) to evaluate the semantic variation by calculating the diagonal differences between text conditioned covariance of fake and real image distributions.

Subsequently, we will support the plausibility of SSD by proving some lemmas.

Lemma 1. *If $\mathbb{C}$ is a non-negative diagonal matrix, the second-moment term can be rewritten as:*

$$||d(\mathbb{C}_{ff|s}) - d(\mathbb{C}_{rr|s})||^2 \propto Tr[(\mathbb{C}_{ff|s}^{\frac{1}{2}} - \mathbb{C}_{ff|s}^{\frac{1}{2}})^2] \quad (3)$$

$$= Tr[\mathbb{C}_{ff|s} + \mathbb{C}_{rr|s} - 2(\mathbb{C}_{ff|s}^{\frac{1}{2}} \mathbb{C}_{rr|s} \mathbb{C}_{ff|s}^{\frac{1}{2}})^{\frac{1}{2}}] \, . \quad (4)$$

Proof. According to [34], conditional covariances can be equivalently written as follows:

$$\mathbb{C}_{ff|s} = \mathbb{C}_{ff} - \mathbb{C}_{fs}\mathbb{C}_{ss}^{-1}\mathbb{C}_{sf} \, , \; \mathbb{C}_{rr|s} = \mathbb{C}_{rr} - \mathbb{C}_{rs}\mathbb{C}_{ss}^{-1}\mathbb{C}_{sr} \, .$$

$\mathbb{C}_{**}$ is defined as a covariance matrix, which is positive semi-definite. Meanwhile, in CLIP space we only focus the diagonal part of $\mathbb{C}$ because CLIP tries to maximize the Cosine similarity between embeddings via training. Thus $\mathbb{C}$ can be simplified as a non-negative diagonal matrix. Therefore for Eq. (4), we have:

$$Tr[\mathbb{C}_{ff|s} + \mathbb{C}_{rr|s} - 2(\mathbb{C}_{ff|s}^{\frac{1}{2}} \mathbb{C}_{rr|s} \mathbb{C}_{ff|s}^{\frac{1}{2}})^{\frac{1}{2}}]$$

$$= Tr[d(\mathbb{C}_{ff|s}) + d(\mathbb{C}_{rr|s}) - 2(d(\mathbb{C}_{ff|s})^{\frac{1}{2}} d(\mathbb{C}_{rr|s}) d(\mathbb{C}_{ff|s})^{\frac{1}{2}})^{\frac{1}{2}}] \, .$$

For the diagonal part, it has:

$$\sigma_{ff|s} + \sigma_{rr|s} - 2(\sigma_{ff|s}^{\frac{1}{2}} \ \sigma_{rr|s} \ \sigma_{ff|s}^{\frac{1}{2}})^{\frac{1}{2}}$$
$$=\sigma_{ff|s} + \sigma_{rr|s} - 2(\sigma_{ff|s} \ \sigma_{rr|s})^{\frac{1}{2}}$$
$$=[\sigma_{ff|s}^{\frac{1}{2}} - \sigma_{rr|s}^{\frac{1}{2}}]^2 \propto [\sigma_{ff|s} - \sigma_{rr|s}]^2 \ ,$$

where σ_{**} is the diagonal part of C_{**}. Therefore, for Eq. (4) we have:

$$Tr[\mathbb{C}_{ff|s} + \mathbb{C}_{rr|s} - 2(\mathbb{C}_{ff|s}^{\frac{1}{2}}\mathbb{C}_{rr|s}\mathbb{C}_{ff|s}^{\frac{1}{2}})^{\frac{1}{2}}]$$
$$= Tr[(d(\mathbb{C}_{ff|s})^{\frac{1}{2}} - d(\mathbb{C}_{rr|s})^{\frac{1}{2}})^2] \propto ||d(\mathbb{C}_{ff|s}) - d(\mathbb{C}_{rr|s})||^2 \ .$$

Thus, Eq. (3) is proportional to Eq. (4). □

Lemma 2. *When m_f, m_s are normalized, the first-moment term can be rewritten as:*

$$1 - cos(m_f, m_s) \triangleq ||m_f - m_s||^2 \ . \tag{5}$$

Proof. Cosine distance is equivalent to Euclidean distance of normalized vectors. In CLIP space, m_f, m_s are normalized embeddings of generated image $\tilde{e}_f$ and text $\tilde{e}_s$. □

We now show how our SSD can be theoretically linked with the other metrics including CS and CFID in T2I synthesis, but exhibits more desirable characteristics.

Proposition 1. *If we only use the first term in Eq. (1), our SSD is converted to CLIP-Score (CS).*

Proof. The Cosine similarity term $cos(m_f, m_s)$ in Eq. (5) of SS is equivalent to CS:

$$cos(m_f, m_s) = \mathbb{E}\,[\,cos(\tilde{e}_f, \tilde{e}_s)\,] \propto \omega$$
$$* \mathbb{E}\,[\,\max(\,cos(\,\tilde{e}_f, \tilde{e}_s\,), 0)\,] = \text{CS}(\,\tilde{e}_f, \tilde{e}_s\,),$$

where ω is a constant scale-coefficient used in CS. □

Proposition 2. *If we measure the distance between $m_{f|s}$ and $m_{r|s}$ for the first term, SSD will be proportional to Conditional Frechet Inception Distance (CFID).*

Proof. By Lemma 1 and 2, SSD can be rewritten as:

$$SSD \propto ||m_f - m_s||^2$$
$$+ Tr[\mathbb{C}_{ff|s} + \mathbb{C}_{rr|s} - 2(\mathbb{C}_{ff|s}^{\frac{1}{2}}\mathbb{C}_{rr|s}\mathbb{C}_{ff|s}^{\frac{1}{2}})^{\frac{1}{2}}] \ . \tag{6}$$

If we use $m_{f|s}$ and $m_{r|s}$ for the first term, we have:

$$||m_{f|s} - m_{r|s}||^2 +$$
$$\underbrace{Tr[\mathbb{C}_{ff|s} + \mathbb{C}_{rr|s} - 2(\mathbb{C}_{ff|s}^{\frac{1}{2}}\mathbb{C}_{rr|s}\mathbb{C}_{ff|s}^{\frac{1}{2}})^{\frac{1}{2}}]}_{TrSV} \tag{7}$$
$$= CFID(Q_{ff|s}, Q_{rr|s}) \ . \tag{8}$$

Then Eq. (1) can be changed to CFID. □

A. Text-based Semantic Similarity Distance

Since SSD measures the semantic distance between texts and generated images as well as the text-conditioned variation difference between synthesized images and real image semantic distributions, it can also be applied to image captioning tasks subject to slight modification.

Definition 2. *SSD in Eq.(1) based on images can be easily altered as SSD_T which is based on texts and can be applied for the image captioning (image-to-text) task:*

$$SSD_T = SSD(\mathbb{Q}_{fs}, \mathbb{Q}_r, \mathbb{Q}_{fs|r}, \mathbb{Q}_{s|r}) \tag{9}$$
$$= [1 - cos(m_{fs}, m_r)] + ||d(\mathbb{C}_{fsfs|r}) - d(\mathbb{C}_{ss|r})||^2 \ , \tag{10}$$

where fs represents fake captions.

We have conducted a series of experiments as shown in Table IV on CUB dataset to show that our SSD_T also reflects image-text consistency for image captioning.

B. SSD's Significance

As shown in Fig. 4, CS only evaluate the first-moment distance between m_s and m_f, which is not enough because the generated image should maintain the shape of $Q_{f|s}$ which is a partial $Q_{r|s}$. CFID only evaluates the first-moment and second-moment distance between $Q_{f|s}$ and $Q_{r|s}$ whose $m_{r|s}$ cannot get rid of undesribed redundancies. Our SSD measures the the first-moment distance between m_s and m_f, and the distortion of $Q_{f|s}$ by comparing its second-moment distance with $Q_{r|s}$.

Our new proposed SSD can be comprehended as evaluating direct consistency between text and images as a first-moment bias term and semantic variation difference between fake images and real images conditioned by text as a second-moment variation term. The first term of SSD is an inversed scaled CS and the second term is a minor correction of the first term that reflects the conditioned semantic distribution distortion of synthesized images, which is similar to the second term of CFID.

In contrast, CS ignores the second-moment variation term, thus causing weakness in estimating semantic variations. It neglects that even if two distributions are very close, the semantics distribution in each modality can be distorted. CFID inappropriately considers the first-moment bias term as the difference between fake and real image distributions, which may mismatch the redundancy contents in the images. Eq. (1) focuses on measuring primary semantic changes, bringing more consistent attention to the major semantics variation than Eq. (7). Semantic gap hinders CFID measuring between real and generated image semantic distributions. For example, the text 'a cute dog' with a real image presents 'a cut dog with a birthday hat', and the generated image presents 'a running cute dog'. The generated image does not match the real image but matches the text. Our experiments also empirically verify the drawbacks of CS, CFID in Table II, and Table IV where generated and modified results exceed GT on CS and CFID fails on text-image consistency.

Fig. 4: Diagrams of CS, CFID and SSD.

Fig. 5: Overall Framework of PDF-GAN.

Furthermore, since we use CLIP as encoders rather than random sampling, our metric mitigates the bias in GT data, enabling a convenient comparison across different datasets. As shown in Fig. 1, our SSD indeed reflects better text-image consistency than the other metrics.

IV. PARALLEL DEEP FUSION GAN

In order to appraise our SSD's feasibility, we further propose PDF-GAN that aims to improving text-image consistency under the guidance of SSD. In this section, we present the Parallel Deep Fusion Generative Adversarial Networks (PDF-GAN) equipped with Hard-negative Sentence Constructor (HNSC) and Semantic Projection (SProj). PDF-GAN fuses semantic information at different levels by using Parallel Fusion Modules (PFM). For semantic supervision, global and local discriminators, semantic perceptual losses, and a contrastive loss are adopted. To capture semantic information in texts more precisely and robustly, HNSC creates stable and controllable hard negative samples, and SProj can overcome the semantic gap by constraining the semantic optimization direction.

A. Parallel Fusion

In our model, CLIP is used as an encoder to map images and texts into a joint semantic space. Guided by SSD, we confirm that using textual data at different levels improves text-image consistency. Global-level feature $\{e_g\}^1$ is the caption embedding. Local-level features $\{e_l\}^n$ are Aspects embeddings which are n key phrases extracted from the caption [22]. For example, 'this black bird has yellow eyes and a long neck', its aspects $\{a\}^{n=3}$ are ['this black bird', 'yellow eyes', 'long neck']. Conditioned by $\{e_g\}^1$ and $\{e_l\}^n$, the generation process is $x_f = G(z \mid \{e_g\}^1, \{e_l\}^n)$ ($G(z)$ for short), where $z \sim N(0,1)$ is a given noise and x_f is the generated image.

In Generator G, we propose PFM for efficient fusion between global and local features. PFM takes the output from previous steps as inputs and $\{e_g\}^1$ and $\{e_l\}^n$ as conditions (in Fig. 6 (c)). The input h_{t-1} is first upsampled to h'_{t-1}, then deep fusion (DF) is conducted on two groups of conditions. After DF, fused features from two branches are concatenated by channels and then go through a convolution layer and outputted as h_t. Especially for the Local DF branch in Fig. 6 (b), we modify the Fusion Block from DF-GAN [7] to take local features. In Fig. 6 (a), two groups of MLPs learn the scale and the bias conditioned by local semantics, respectively. Input h_{t-1} is first expanded to the proper shape, then scaled and biased. The conditioned features are averaged and passed to later processors. Using PFM to fuse multiple levels of textual information efficiently, our PDF-GAN capture precise semantics while maintaining decent image quality. We use two PFMs in our experiments due to memory limitations.

B. Hard-negative Sentence Constructor

HNSC constructs hard negative sentence samples by randomly replacing tokens in the given description according to
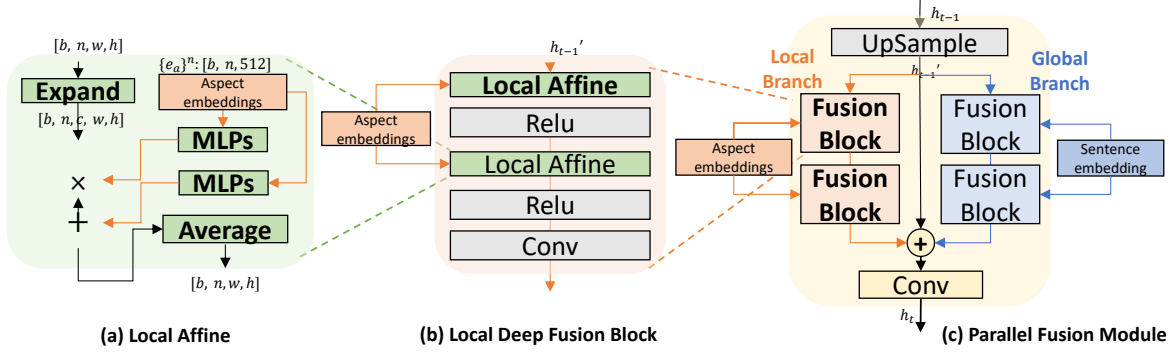
Fig. 6: Parallel Fusion Module

the Part of Speech (POS). Nouns, verbs, and adjectives are replaced by other nouns, verbs, and adjectives. For example, for the text 'this bird is blue on its tail and has a long pointy beak,' our HNSC will randomly replace a certain percentage of words by its POS, (change 'blue' to 'red,' 'tail' to 'head,' etc.). Candidates for replacement are gathered from the dataset. HNSC produces stable and controllable hard negative textual samples which force discriminators to learn precise semantics.

C. Training Objectives

We adopt the hinge loss [35] for Discriminator D and Generator G. D usually takes four kinds of pairs: fake image with real text $(G(z), e)$, real image with real text (x, e), fake image with mismatched text $(G(z), e_m)$ and real image with mismatched text (x, e_m). To capture semantic information at the global and local levels, we use two discriminators D_g and D_l. For each D given by matched k conditions $\{e\}^k$, and mismatched k conditions $\{e_m\}^k$, the loss with modified Matching Aware Gradient Penalty [7] is defined as:

$$
\begin{aligned}
\mathcal{L}_D = & \mathbb{E}_{x_r \sim \mathbb{P}_r, e \in \{e\}^k}[1 - D_g(x, e)] \\
& + \frac{1}{2}[\mathbb{E}_{x_r \sim \mathbb{P}_r, e_m \in \{e_m\}^k}[1 + D_g(x, e_m)] \\
& + \mathbb{E}_{G(z) \sim \mathbb{P}_g, e \in \{e\}^k}[1 + D_g(G(z), e)]] \\
& + q\mathbb{E}_{x \sim \mathbb{P}_r}[(\|\bigtriangledown_x D_g(x, \bar{e})\| + \|\bigtriangledown_e D(x, \bar{e})\|)^p] .
\end{aligned}
$$
(11)

The penalized term uses averaged $\{e\}^k$ as the condition. Loss functions for D_g and D_l both follow Eq. (11) with that the given conditions $\{e\}^k$ and $\{e_m\}^k$ are global level $\{e_g\}^1, \{e_{mg}\}^1$ and local level $\{e_l\}^n, \{e_{ml}\}^n$, respectively. Notice that $\{e_{mg}\}^1$ is hard-negative sentence constructed by HNSC and $\{e_{ml}\}^n$ are mismatched aspects from the batch. q and p are hyper-parameters set to 2.0 and 6 in our experiments. The adversarial loss of G is defined as:

$$
\begin{aligned}
\mathcal{L}_a = & -\mathbb{E}_{G(z) \sim \mathbb{P}_g}[D_g(G(z), \{e_g\}^1)] \\
& -\mathbb{E}_{G(z) \sim \mathbb{P}_g}[D_l(G(z), \{e_l\}^n)] .
\end{aligned}
$$
(12)

To enhance semantic information at different levels in G, we adopt global and local semantic perceptual losses $\mathcal{L}_g, \mathcal{L}_l$:

$$
\mathcal{L}_g = f_C(G(\tilde{z}))^T \cdot \tilde{e}_g , \quad \mathcal{L}_l = \frac{1}{n}\sum_{i=1}^{n} f_C(G(\tilde{z}))^T \cdot \tilde{e}_l^i .
$$

Algorithm 1 Semantic Projection in One Step

Require: Training data($\{e_g\}^1, \{e_l\}^n$), generator G, discriminators D_l, D_g, and $z \sim N(0,1)$
1: $x' = G(z \mid \{e_g\}^1.\{e_l\}^n)$
2: $y_g' = D_g(x' \mid \{e_g\}^1), \quad y_l' = D_a(x' \mid \{e_l\}^n)$
3: Calculate loss $\mathcal{L}_{D_g}, \mathcal{L}_{D_l}$ for D_g, D_l ▷ See Eq. (11).
4: $\delta_{D_g} \leftarrow \bigtriangledown \mathcal{L}_{D_g}, \delta_{D_l} \leftarrow \bigtriangledown \mathcal{L}_{D_l}$
5: $\theta_{D_g} \leftarrow \theta - \alpha\delta_{D_g}, \theta_{D_l} \leftarrow \theta - \alpha\delta_{D_l}$
6: Calculate $\mathcal{L}_a, \mathcal{L}_s$ ▷ See Eq. (14).
7: $\delta_a \leftarrow \bigtriangledown \mathcal{L}_a, \delta_s \leftarrow \bigtriangledown \mathcal{L}_s$
8: $\bar{\delta}_s \leftarrow PROJECT(\delta_a, \delta_s)$ ▷ See Eq. (15).
9: $\theta_G \leftarrow \theta_G - (\alpha_a\delta_a + \alpha_s\bar{\delta}_s)$
   Return: x'

A contrastive loss $\mathcal{L}_c$ is introduced to further repel mismatched samples:

$$
\mathcal{L}_c = \frac{f_C(G(\tilde{z}))^T \cdot \tilde{e}_{mg}}{f_C(G(\tilde{z}))^T \cdot \tilde{e}_{mg} + f_C(G(\tilde{z}))^T \cdot \tilde{e}_g} .
$$
(13)

The final generative loss $\mathcal{L}_G$ combines above four losses:

$$
\mathcal{L}_G = \underbrace{\lambda(\mathcal{L}_g + \mathcal{L}_l + \mathcal{L}_c)}_{Semantic\ loss\ \mathcal{L}_s} + \underbrace{\mathcal{L}_a}_{Adversarial\ loss}
$$
(14)

where we set $\lambda = 10$ empirically in our experiments.

D. Semantic Projection

Since the semantic gap in CLIP space causes conflicts in optimization directions between $\mathcal{L}_a$ and $\mathcal{L}_s$, we design SProj to overcome the conflicts (see Fig. 3(a)). Inspired by GEM for continuous learning [36], we treat minimizing $\mathcal{L}_a$ and $\mathcal{L}_s$ as two tasks. Instead of training on two tasks alternately, we optimize them simultaneously. Algorithm 1 shows the training and protocol of SProj in one step. In each step, after we calculate the gradients δ_a, δ_s for $\mathcal{L}_a$ and $\mathcal{L}_s$, we conduct *PROJECT* on δ_s before we process backpropagation on both tasks. If there is a direction conflict, the semantic optimization direction δ_s will be re-projected to a new direction $\bar{\delta}_s$ in which it can optimize for $\mathcal{L}_s$ while not enlarging $\mathcal{L}_a$. *PROJECT* is defined as: For gradients $\Delta := -(\delta_a, \delta_s)$, if $\langle \delta_a, \delta_s \rangle \geq 0$, we solve the Quadratic Program to get solution ϑ^* for ϑ:

$$
\min_{\vartheta} 1/2 \left[\vartheta^T \Delta\Delta^T + \delta_s^T \Delta^T \vartheta\right], \quad s.t. \quad \vartheta \geq 0 .
$$
(15)

The projected gradients can then be updated as $\hat{\delta} = G^T\vartheta^* + g_s$. With SProj, the model can converge to the closest point

Parameters	CUB	COCO	Parameters	CUB	COCO
TRUNCATED_NOISE	False	True	MAX_EPOCH	600	120
MAX_ATTR_NUM	3	3	DISCRIMINATOR_LR	e-4 * 2	e-4 * 2
MAX_ATTR_LEN	5	5	GENERATOR_LR	2e-4 / 2	2e-4 / 2
BASE_SIZE	64	64	CAPTIONS_PER_IMAGE	10	5
BATCH_SIZE	32	32	EMBEDDING_DIM	512	512
WORDS_NUM	18	18			

TABLE I: Parameters for our experiments for PDF-GAN.

	CUB				COCO			
	$SSD\downarrow$	CS$\uparrow$	$R\uparrow$	CFID$\downarrow$	$SSD\downarrow$	CS$\uparrow$	$R\uparrow$	CFID$\downarrow$
GT	72.07	69.83	21.26	0.00	72.68	68.30	69.26	0.00
AttnGAN	76.10	68.00	67.82	6.89	88.01	50.48	85.47	7.75
DM-GAN	75.51	69.03	72.31	**3.04**	86.62	51.10	88.56	**6.25**
DAE-GAN	76.61	68.93	**85.45**	4.16	87.98	51.88	**92.61**	17.11
DF-GAN	76.82	67.80	39.50	4.55	82.67	60.85	51.75	7.43
PDF-GAN	**72.72**	**80.28**	11.75	12.86	**75.29**	**72.71**	49.79	8.68

TABLE II: Text-image consistency results of SSD, CS, R and $CFID$ on CUB and COCO. Best results for each metric are highlighted.

between texts and image features in the semantic space (as shown in Fig. 3 (a)), attaining a mutual balance between text-image consistency and image quality.

V. EXPERIMENTS

We conduct two groups of experiments: 1) The first group of experiments are performed on PDF-GAN to showcase the feasibility of SSD for text-to-image task. It also suggests that our SSD can be used to guide text-to-image model designing. 2) The second group is a series of simple experiments that show the feasibility of SSD_T for image captioning. These experimental results reflect that the limitations of the existing text-image consistency metrics can be alleviated by our SSD.

A. Experimental Setup for PDF-GAN

In our experiments, we evaluate our metric and method on two datasets, CUB [10] and COCO [9]. The generated images are judged by both text-image consistency metrics and image quality metrics. For qualitatively evaluating the text-image consistency, the proposed SSD is used and compared with other popular metrics R [8], CS [12], and CFID [13]. For a fair comparison, CFID is calculated from CLIP embeddings, and all CLIP-based metrics, SSD, CS and CFID, are scaled by 100 in our experiment. The standard metrics, Inception Score (IS) [37] and Fréchet Inception Distance (FID) [38], are used to quantitatively evaluate the generated image quality. IS was not used to evaluate COCO because it works not well as indicated in [7], [39]. Specially, DF-GAN is also trained with CLIP as encoders (DF-GAN+CLIP) by using its original settings for comparison.

All the metrics are computed over 30K generated images. We use the released models from competitors for the metric calculation of text-image consistency, and directly adopt their reported image quality results. The number of the aspect per caption is set to 3, and the maximum number of words per caption is set to 18.

The parameters used for our experiments on CUB and COCO datasets are Table I.

B. Experimental Setup for Text Modification

For testing SSD_T, we modify the original texts from CUB dataset and use modified texts as toy examples for image captioning task. The group of experiments "Cut Words" means that we slice original captions and take them as fake captions for Eq. (9) calculation and $[* : *]$ indicates the slice we take. The group of experiments "Replaces Words" mean that we replace words in the caption and use them as fake captions. We engage our HNSC to produce the replaced captions. Parameter r means the ratio of the caption that we replace. We compare our SSD_T with image captioning metric CLIPScore [12] (CS). We also deploy 30K text-image pairs for calculating SSD_T and CS.

C. Appraise SSD

We appraise our SSD by comparing it with CS, R and CFID on AttnGAN [8], DM-GAN [20], DAE-GAN [22], DF-GAN [7], DF-GAN+CLIP and our PDF-GAN. Quantitative results are presented in Table II [3], and the generated examples are visualized in Figs. 9-10.

As shown in Table II, R's indirect measurement and highly-dataset biased sampling generate even better scores for most methods than GT, indicating its severe limitations. Leveraging CLIP embeddings, CS can better reflect text-image consistency. However, since CLIP may suffer from binding attributes to objects [33], CS struggles to reflect precise semantics consistency. E.g., it might lead that our PDF-GAN exceeds GT in CS. Such drawback of CS can also be seen in Fig. 7 from our ablation studies, where the model failing to capture precise semantics like 'orange and white' but obtains higher CS. It can also been seen in Table. IV where modified captions exceed CS on real captions.

CFID aims to measure the distance between images' distributions approximately. If we compare the CFID scores in Table II with the corresponding examples in Fig. 9, Fig. 10, it is evident that CFID is not in line with the consistency. This is

---

[3] SSD, SS, dSV and $TrSV$ are by 100 for better readability in this paper.

	CUB				**COCO**			
	$SSD\downarrow$	$SS\downarrow$	$dSV\downarrow$	$TrSV\downarrow$	$SSD\downarrow$	$SS\downarrow$	$dSV\downarrow$	$TrSV\downarrow$
GT	72.07	72.07	0.00	0.00	72.68	72.68	0.00	0.00
AttnGAN	76.10	72.80	3.30	0.26	88.01	79.81	8.20	1.26
DM-GAN	75.51	72.39	**3.12**	**0.19**	86.62	79.56	7.06	0.90
DAE-GAN	76.61	72.43	4.18	0.36	87.98	79.25	8.73	1.55
DF-GAN	76.82	72.88	3.94	0.32	82.67	75.66	7.01	0.89
DF-GAN+CLIP	76.63	72.00	4.63	0.11	78.75	76.53	**2.22**	**0.20**
PDF-GAN	**72.72**	**67.89**	4.83	0.59	**76.49**	**70.81**	5.68	0.60

TABLE III: Results of SSD, SS, dSV and $TrSV$. Best results are highlighted as **bold**.

Cut	$SSD_T\downarrow$	$SS\downarrow$	$dSV\downarrow$	CS↑	CFID↓	**Rep**	$SSD_T\downarrow$	$SS\downarrow$	$dSV\downarrow$	CS↑	CFID↓
GT	72.08	72.08	0.00	69.80	0.00	GT	72.08	72.08	0.00	69.80	0.00
[1:]	**72.81**	71.55	1.26	**71.13**	2.01	r=0.05	78.46	73.77	4.69	65.58	9.22
[1:-1]	**72.92**	71.61	1.31	**70.98**	2.12	r=0.1	80.39	74.04	6.35	64.90	10.66
[2:]	76.05	72.50	3.55	68.75	2.01	r=0.6	92.17	76.43	15.74	58.93	25.93
[3:]	79.03	72.84	6.19	67.90	7.78						

TABLE IV: Results of 'Cut Words' ('Cut' at right-hand side) and 'Replace Words' ('Rep' at left-hand side) experiments of SSD_T for CUB. SSD_T, dSV, SS, and CS are scaled by 100 for readability. Results exceed GT are highlighted as **bold**.
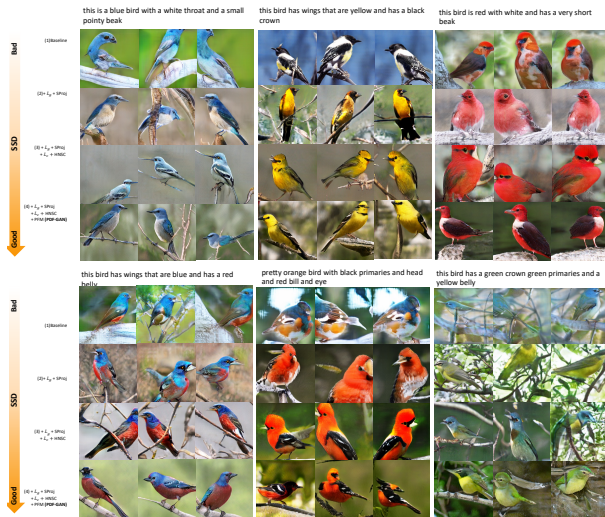

Fig. 7: Generated examples of ablation studies. Generated examples organized as SSD-descending order–Continuous. Note that ablation task $+\mathcal{L}_c$ (CS: 71.76, SSD: 74.49) obtains better CS but worse SSD than +HNSC (CS: 71.08, SSD: 73.68).

Fig. 8: SSD of DM-GAN [20] and PDF-GAN vs. different number of samples.

of ablation studies are shown in Table VI and qualitative results are shown in Fig. 7.

It can be observed that when our SSD is reduced, the generated images tend to contain more consistent semantics with the given texts. Fig. 7(1) does not obtain good text-image consistency due to high semantic bias even with a low dSV. When SS achieves a better level of about 69, dSV makes a difference. As shown in Fig. 7(2)-(4), even with a similar SS, the models with better SSD can lead to significantly better text-consistent images. Note that task (3) with higher CS but lower SSD and (4) with lower CS and higher SSD in Fig. 7, task (4) are more consistent with the given texts, indicating the importance of the correction brought by dSV.

2) dSV Vs. $TrSV$: We show additional experiments to support that our SSD's second-moment term (dSV) is more indicative of the major semantic variations than $CFID$'s second-moment term ($TrSV$). Notice that we exploit the diagonal part of $\mathbb{C}$ to calculate $TrSV$. We also display the first-moment term (SS) of SSD for better readability. In Table III, it clearly shows that $TrSV$ and dSV have the same tendency to reflect semantic changes, as they are equivalent. But, $TrSV$ cannot obviously represent semantic changes, while our dSV can produce more significantly different indications.

3) Appraise SSD_T: Table IV exhibits results of SSD_T In the first group of experiments, when we cut out the first word and both first and last words, they have better CS scores than ground truth (GT), suggesting that CS is quite limited. Our

mainly because real images usually contain redundancies not mentioned by texts which may however be not reacquired in the generated images, causing unsuitable fist-moment term in Eq. (7).

Our SSD mitigates the drawbacks of both CS and CFID. As shown in Fig. 9 and Fig. 10 that are organized in SSD descending-order horizontally, text-image consistency can indeed be observed better from left to right. In contrast to CS in Fig. 7, our SSD demonstrates clearly better semantic consistency. Furthermore, since our SSD alleviates the dataset bias, we can compare SSD across different datasets.

1) SS and dSV: We show that dSV has a significant contribution when SS scores are similar. Since the other methods may lead to high semantic bias, we take the results from our ablation studies for a better demonstration. Quantitative results
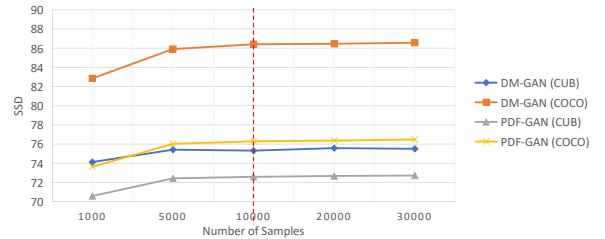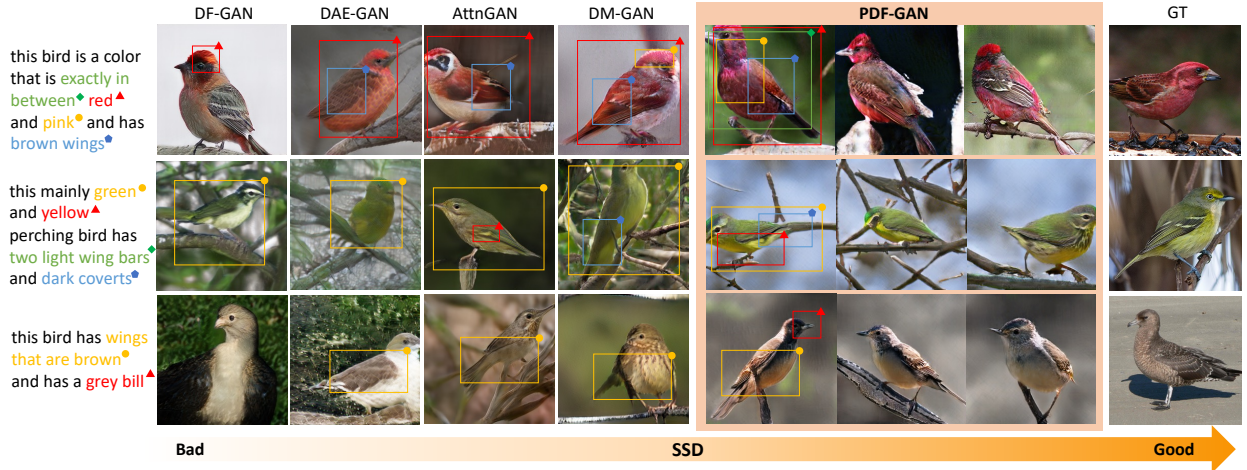
Fig. 9: Examples for T2I synthesis by AttnGAN, DM-GAN, DAE-GAN, DF-GAN and our PDF-GAN on CUB.
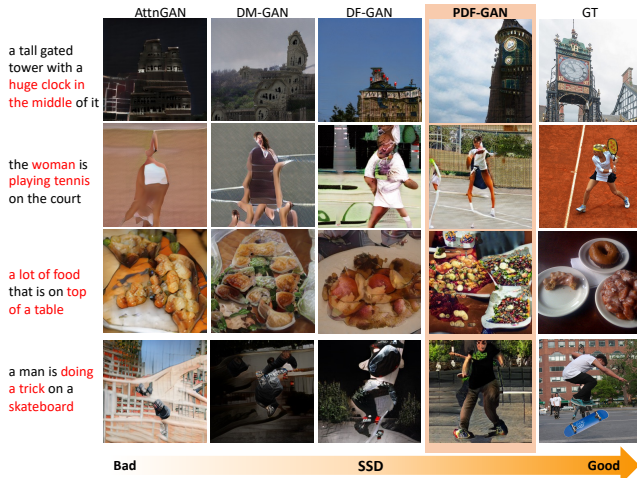
Fig. 10: Examples for T2I synthesis by AttnGAN, DM-GAN, DF-GAN and our PDF-GAN on COCO.

| | **CUB** | | **COCO** |
	IS ↑	FID ↓	FID ↓
AttnGAN	4.36	23.98	35.49
DM-GAN	4.75	16.09	32.64
DAE-GAN	4.42	15.19	28.12
MirrorGAN	4.56	18.34	34.71
SD-GAN	4.67	-	-
TIME	<u>4.91</u>	<u>14.30</u>	34.14
DF-GAN	**5.10**	14.81	**19.32**
DF-GAN + CLIP	4.63	24.16	21.84
PDF-GAN	4.59	**12.30**	<u>21.01</u>

TABLE V: IS on CUB and FID on CUB and COCO. Best and second best results are highlighted as **bold** and by <u>underlines</u>, respectively.

SSD_T's dSV can overcome this issue. It can be seen that our SSD_T gets worse as the modification progresses. CFID gets the same tendency as SSD_T because there are quite fewer redundancies between texts, i.e., the images usually contain all textual described information. This also indicates that the text-image consistency evaluation is more difficult than image-text consistency evaluation.

4) Stability of SSD: As shown in Fig. 8 we use different numbers of samples for SSD calculation of DM-GAN [20] and our PDF-GAN. It turns out that our SSD becomes stable after more than 10K samples are used. Thus we recommend using more than 10K samples for the SSD calculation.

D. Appraise PDF-GAN

If we examine SSD scores in Table II, all the methods show better text-image consistency on CUB than that on COCO, while our PDF-GAN achieves the best text-image consistency. Meanwhile, PDF-GAN can maintain decent image quality, achieving the best (12.30) and second best (21.01) in FID on CUB and COCO, respectively.

As shown in Figs. 9-10, PDF-GAN can capture complex and precise semantics while maintaining decent quality. For a less complex dataset like CUB, our PDF-GAN surpasses 2.80 in SSD over the second best method while DF-GAN produces high-quality but text-mismatched images; other attention-based methods fail to capture detailed semantics. Moreover, PDF-GAN can stably generate complex semantics in visual (e.g. in Fig. 9 where the bird matches 'exactly in red and pink' and other highlighted phrases). The superiority of our PDF-GAN is more evident on COCO. In comparison, DF-GAN and DF-GAN+CLIP have more balanced results on both datasets and may also produce high-quality images, but the overall text-image consistency is not promising. Our proposed PDF-GAN can capture the complex semantics and have 6.18 improvement in SSD on COCO, suggesting our model is able to capture complex semantics (see Fig. 10).

For better visualization, we illustrate semantic distributions of fake images generated by our PDF-GAN and other models by comparing with real images and texts as Fig. 11 and Fig. 12. Our generated images obtain the semantic distribution closer to given texts than other methods.

1) Ablation Study: We take CUB as one typical example for ablative analysis. Our ablation study settings and results can be seen in Table VI whilst examples are shown in Fig. 3 (b). Two findings are verified: 1) The semantic gap can cause optimization conflicts between adversarial and semantic losses.

Task	CLIP	$D_g + \mathcal{L}_g$	SProj	$\mathcal{L}_c$	HNSC	PFM	$D_l + \mathcal{L}_l$	$SSD \downarrow$	$SS \downarrow$	$dSV \downarrow$	CS ↑	IS ↑	FID ↓
Baseline	✓							76.63	72.00	**4.63**	59.94	4.50	24.16
+ $\mathcal{L}_g$	✓	✓						75.59	69.55	6.04	68.10	4.63	14.69
+ SProj	✓	✓	✓					74.10	68.77	5.33	70.70	**4.81**	15.55
+ $\mathcal{L}_c$	✓	✓	✓	✓				74.49	68.45	6.04	71.76	4.58	14.78
+ HNSC	✓	✓	✓	✓	✓			73.68	68.67	5.01	71.03	4.61	14.41
+ PFM **PDF-GAN**	✓	✓	✓	✓	✓	✓	✓	**72.72**	**67.89**	4.83	**73.63**	4.59	**12.30**

TABLE VI: Ablation settings and ablation study results of different components on CUB. All components are added cumulatively as shown in Table VI. Note experiment $+L_c$ uses random samples for L_c calculation.
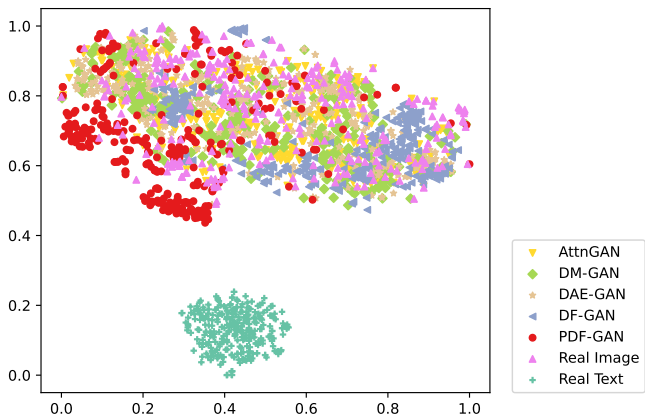

Fig. 11: TSNE map of CLIP embeddings of generated images by AttnGAN, DM-GAN, DAE-GAN, DF-GAN, and our PDF-GAN compared with real images and texts on CUB.
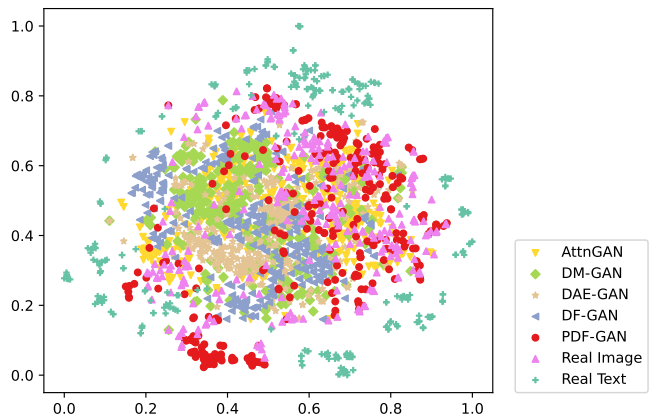

Fig. 12: TSNE map of CLIP embeddings of generated images by AttnGAN, DM-GAN, DAE-GAN, DF-GAN, and our PDF-GAN compared with real images and texts on COCO.

2) Using random samples from the dataset or the training batch may prevent discriminators from learning precise semantics. Overall, these two issues could be well tackled by the proposed SProj and HNSC.

Reported in the first group of experiment of Table VI, brutally adding semantic perceptual loss can improve image quality, but may not significantly improve text-image consistency due to the semantic gap. Specifically, the decrease of SS is accompanied by a significant increase of dSV, indicating when the semantic distribution of generated images gets closer to the texts, its distribution form is likely to be distorted. SProj can strengthen the $\mathcal{L}_g$'s supervision, forcing the model to be further optimized in semantics without sacrificing the image quality. As shown in Fig. 3 (b)(2), SProj with semantic perceptual loss benefits capturing more semantic details, easing semantic distribution distortion of generated images.

The second group shows that contrastive loss $\mathcal{L}_c$ may deteriorate text-image consistency if random mismatched texts are given. The scores of SSD, dSV, and IS get worse when this loss is added. However, HNSC can mitigate this issue by offering stable and controllable hard-negative sentences as mismatched samples. It pushes the discriminators to learn more semantic differences between positive and hard negative texts, thus improving the model's capability in capturing precise semantics through reducing semantic distribution distortion of generated images, as seen in Fig. 3 (b)(3) and Fig. 7. It also eases the degradation in IS and improves FID.

2) Sensitivity Analysis of λ: We conduct a sensitivity analysis for λ in the $\mathcal{L}_G$ in Eq. (14). As shown in Table VII,

	$SSD \downarrow$	$S_C \uparrow$	$IS \uparrow$	$FID \downarrow$
$\lambda = 1$	79.26	67.38	4.82	14.13
$\lambda = 10$	**72.72**	80.28	4.59	**12.30**
$\lambda = 20$	72.86	**81.60**	**4.91**	14.73

TABLE VII: Sensitive analysis for λ.

small λ causes degradation of text-image consistency because the optimization step size is too small. We then recommend setting a reasonably large value for λ because our SProj can constrain the optimization of $\mathcal{L}_G$.

VI. CONCLUSION

In this paper, we propose a novel metric SSD to better evaluate text-image consistency. Both theoretical analysis and empirical investigation show that SSD can indeed reflect the semantic consistency in text-to-image generation. We also design a novel framework called PDF-GAN along with two plug-and-play modules that can further enhance the text-image consistency. Experiments on benchmark datasets confirm the effectiveness of SSD as well as the advantages of PDF-GAN both qualitatively and quantitatively.

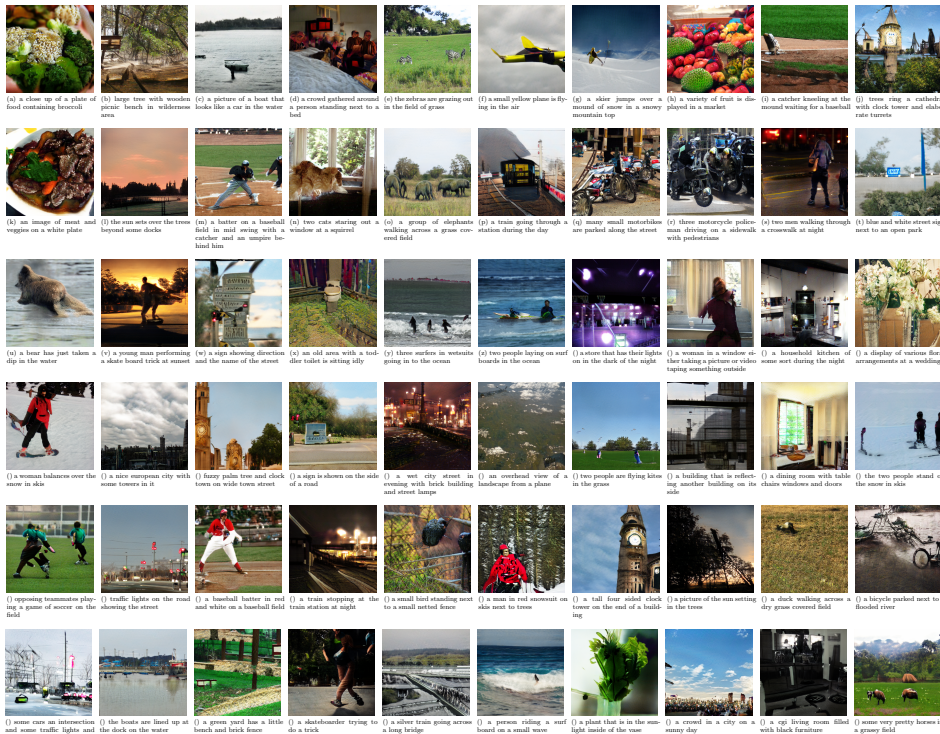Fig. 13: More examples produced by PDF-GAN on CUB.

Fig. 14: More examples produced by PDF-GAN on COCO.

REFERENCES

[1] M. Yuan and Y. Peng, "Ckd: Cross-task knowledge distillation for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 1955–1968, 2019.

[2] S. Hong, D. Yang, J. Choi, and H. Lee, "Inferring semantic layout for hierarchical text-to-image synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7986–7994.

[3] R. Li, N. Wang, F. Feng, G. Zhang, and X. Wang, "Exploring global and local linguistic representations for text-to-image synthesis," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3075–3087, 2020.

[4] Y. Gou, Q. Wu, M. Li, B. Gong, and M. Han, "Segattngan: Text to image generation with segmentation attention," *arXiv preprint arXiv:2005.12444*, 2020.

[5] J. Cheng, F. Wu, Y. Tian, L. Wang, and D. Tao, "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*, 2020, pp. 10 911–10 920.

[6] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[7] M. Tao, H. Tang, S. Wu, N. Sebe, X.-Y. Jing, F. Wu, and B. Bao, "Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis," *arXiv preprint arXiv:2008.05865*, 2020.

[8] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1316–1324.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[10] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[11] T. Hinz, S. Heinrich, and S. Wermter, "Semantic object accuracy for generative text-to-image synthesis," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

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

[13] M. Soloveitchik, T. Diskin, E. Morin, and A. Wiesel, "Conditional frechet inception distance," *arXiv preprint arXiv:2103.11521*, 2021.

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

[15] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," *arXiv preprint arXiv:2203.02053*, 2022.

[16] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.

[17] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907–5915.

[18] ——, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[19] M. Wang, C. Lang, S. Feng, T. Wang, Y. Jin, and Y. Li, "Text to photo-realistic image synthesis via chained deep recurrent generative adversarial network," *Journal of Visual Communication and Image Representation*, vol. 74, p. 102955, 2021.

[20] M. Zhu, P. Pan, W. Chen, and Y. Yang, "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*, 2019, pp. 5802–5810.

[21] W. Huang, R. Y. Da Xu, and I. Oppermann, "Realistic image generation using region-phrase attention," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 284–299.

[22] S. Ruan, Y. Zhang, K. Zhang, Y. Fan, F. Tang, Q. Liu, and E. Chen, "Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 960–13 969.

[23] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] L. Gao, D. Chen, Z. Zhao, J. Shao, and H. T. Shen, "Lightweight dynamic conditional gan with pyramid attention for text-to-image synthesis," *Pattern Recognition*, vol. 110, p. 107384, 2021.

[25] J. Ma, L. Zhang, and J. Zhang, "Sd-gan: Saliency-discriminated gan for remote sensing image superresolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1973–1977, 2019.

[26] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 833–842.

[27] T. Qiao, J. Zhang, D. Xu, and D. Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514.

[28] A. D. Seshadri and B. Ravindran, "Multi-tailed, multi-headed, spatial dynamic memory refined text-to-image synthesis," *arXiv preprint arXiv:2110.08143*, 2021.

[29] Y. Dong, Y. Zhang, L. Ma, Z. Wang, and J. Luo, "Unsupervised text-to-image synthesis," *Pattern Recognition*, vol. 110, p. 107573, 2021.

[30] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[31] R. Gal, O. Patashnik, H. Maron, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *arXiv preprint arXiv:2108.00946*, 2021.

[32] Z. Wang, W. Liu, Q. He, X. Wu, and Z. Yi, "Clip-gen: Language-free training of a text-to-image generator with clip," *arXiv preprint arXiv:2203.00386*, 2022.

[33] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

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

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[36] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, 2017.

[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[38] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[39] Z. Zhang and L. Schomaker, "Dtgan: Dual attention generative adversarial networks for text-to-image generation," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.