Supplementary Materials of Semantic-aware Data Augmentation for Text-to-image Synthesis

Anonymous Author(s)
Submission Id: 890

ACM Reference Format:
Anonymous Author(s). 2023. Supplementary Materials of Semantic-aware Data Augmentation for Text-to-image Synthesis. In *Proceedings of ACM International Conference on Multimedia (ACM Multimedia '2023).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

A MORE MATHEMATICAL DETAILS

Here we provide more details for our derivations and proofs.

A.1 Derivation Details of Training Objectives for G with ITA

Based on empirical risk minimization (ERM), the empirical risk for generator G is defined as:

$$R_k(\theta) := \frac{1}{k}\sum_{i=1}^{k} L(\theta, X). \tag{1}$$

Its standard augmented version and corresponding augmented loss are defined as:

$$\hat{R}_k(\theta) := \frac{1}{k}\sum_{i=1}^{k}\int_{\mathcal{A}} L(\theta, f(X))dQ_{ITA}(f), \tag{2}$$

where $\mathbf{Q}_{ITA}$ is a probability distribution on a group $\mathcal{A}$ of ITA transforms from which f is sampled. Since only one ITA will be used, the general sample objective with ITA is defined as:

$$\min_{\theta} \hat{R}_k(\theta) := \frac{1}{k}\sum_{i=1}^{k} L(\theta, ITA(X_i)). \tag{3}$$

We then define the solution of Eq. (3) as:

$$\theta^*_{ITA} \in arg\min_{\theta\in\Theta} \frac{1}{k}\sum_{i=1}^{k} L(\theta, ITA(X_i)), \tag{4}$$

where Θ is defined as some parameter space.

A.2 Proof Details

Proposition A.1 (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:*

$$\mathbb{E}[S(\theta, \hat{\theta}(X))] \geq \mathbb{E}[S(\theta, \hat{\theta}_{ITA}(X)]), \tag{5}$$

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

Proof. From Group-Theoretic Framework for Data Augmentation [2], we obtain a direct consequence that:

$$Cov[S(\hat{\theta}_{ITA}(X))] \preceq Cov[S(\hat{\theta}(X)] \,, \tag{6}$$

where $Cov[\cdot]$ means the covariance matrix decreases in the Loewner order. Therefore for any real-valued convex loss function $S(\theta,\cdot)$, we have:

$$\mathbb{E}_{X\sim\mathcal{X}}[S(G(X))] \geq \mathbb{E}_{X\sim\mathcal{X}}[S(\bar{G}(X))], \tag{7}$$

where $\bar{G}$ is the conditional expectation of G and $S(\theta,\cdot)$ is any real-valued convex loss function. Empirically, $S(\theta,\cdot)$ can be a real-valued convex loss produced by discriminators, perceptual semantic loss produced by pretrained models, and others. It also suggests that ITA can be considered as an algorithmic regularization like other data augmentations, and augmented G can obtain better text-image consistency. □

Proposition A.2. *ITA_C can be considered a closed-form solution for general textual semantic preserving augmentation methods of T2Isyn.*

Proof. Assume exact invariance holds. Captions offered in the dataset are based on real images, thus: $e_{s|r} \triangleq e_s$. Assume all mentioned following models are well-trained. We consider two situations:

(1) For methods that use extra models to generate more textual data based on real images r (such as I2T2I [5] which uses a pretrained captioning model), we have:

$$e_{s|r} \sim \mathcal{N}(m_r, \mathbb{C}_{ss|r}\mathbb{I}) = Q_{s|r}, \tag{8}$$
$$e'_{s|r} \sim \mathcal{N}(m_r, \mathbb{C}'_{ss|r}\mathbb{I}) = Q'_{s|r}. \tag{9}$$

When the extra models are trained on the dataset used for T2Isyn, exact invariance holds. We have:

$$Q_{s|r} =_d Q'_{s|r}, \tag{10}$$
$$e'_{s|r} \sim \mathcal{N}(m_r, \mathbb{C}_{ss|r}\mathbb{I}). \tag{11}$$

(2) Consider methods that use extra models that generate synonymous texts based on real texts (such as retrieving texts from the dataset and refining the conflicts like RiFeGan [3], using extra pretrained synonymous text generating model, and our proposed ITA_T). Assume exact invariance holds. Captions offered in the dataset are based on real images,

ACM Multimedia '2023, October 28– November 03, 2023, Ottawa, Ontario, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

thus: $e_{s|r} \triangleq e_s, e_{s|r} \sim Q_{ss|r}$. Augmented texts e'_s are retrieved from the dataset and refine the semantic conflicts between $e'_{s|r}$ and $e_{s|r}$ based on the main semantics of real images r. Therefore:

$$e_{s|r} \sim N(m_r, \mathbb{C}_{ss|r}\mathbb{I}) = Q_{s|r}, \tag{12}$$

$$e'_{s|r} \sim N(m_r, \mathbb{C}_{ss|r}\mathbb{I}) = Q_{s|r}. \tag{13}$$

Due to $e_{s|r}$ depending on the semantics of r, $e_{s|r}$ should maintain the main semantics of r. Therefore we have:

$$m_{s|r} \approx m_r, \tag{14}$$

$$e'_{s|r} \sim N(m_{s|r}, \mathbb{C}_{ss|r}\mathbb{I}), \tag{15}$$

where ITA_C is a closed-form solution. Therefore, ITA_C can be considered a closed-form solution for general textual semantic preserving augmentation methods of T2Isyn. □

Proposition A.3. *Assume that E_I is linear. Constraining the distribution $Q_{\mathcal{E}}$ of $e_{f|s}$ can additionally constrain the distribution $\mathcal{F}$ of f.*

Proof. There are two situations:

(1) If E_I is inevitable, Proposition A.3 is obvious.
(2) If E_I is not inevitable, constraining $\mathcal{F}$ can affect $\mathcal{E}$ in not Nullspace of E_I:$\neg Null(E_I)$:

$$C(\mathcal{E}) \propto C(\neg Null(E_I)(\mathcal{F})), \tag{16}$$

where $C(\cdot)$ is a certain constraint. For Nullspace, there will be no effect. If not all the mass of $\mathcal{F}$ locates in the $Null(E_I)$, Proposition A.3 holds. If $\mathcal{F}$ all locates in the $Null(E_I)$ while E_I is well trained, it means $\mathcal{F}$ does not contain any semantics that matches textual semantics, inferring a total collapse of G. Since we assume the G can learn the representation, it is impossible that $\mathcal{F}$ all locates in the $Null(E_I)$.

Therefore, Proposition A.3 holds.

□

Proposition A.4. *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. (**??**), we have following constrain for $e'_{f|s}$ and $e_{f|s}$:

$$|e'_{f|s} - e_{f|s}| \leq |e''_{f|s} - e_{f|s}| = |\epsilon^*| \odot \beta \cdot d(\mathbb{C}_{rr|s}) \ . \tag{17}$$

For each dimension of semantic embeddings, we have:

$$\begin{aligned} |{e'_{f|s}}^d - {e_{f|s}}^d| &= \beta \cdot \mathbb{E}[({e'_{s|r}}^d - {e_{s|r}}^d)^2] \ , \\ &\leq \beta \cdot \max[({e'_{s|r}}^d - {e_{s|r}}^d)^2] \\ &= \beta \cdot [({e''_{s|r}}^d - {e_{s|r}}^d)^2] \ , \\ &= |{\epsilon^*}^d| \cdot \beta \cdot d(\mathbb{C}_{rr|s})^d \\ &= \beta \cdot d(\mathbb{C}_{rr|s})^d, \end{aligned} \tag{18}$$

$$|{e'_{f|s}}^d - {e_{f|s}}^d| \leq = \beta \cdot d(\mathbb{C}_{rr|s})^d, \tag{19}$$

where $d = \{1, ..., n\}$ and n is the dimension of the semantic embedding; $d(.)$ represents diagonal part of a matrix; β is a positive constant. Due to the fact of the many-to-many relationship between texts and images, we have $d(\mathbb{C}_{rr|s})^d > 0$. Assume exact invariance holds, $|{\epsilon^*}^d| = 1$; $\beta \cdot d(\mathbb{C}_{rr|s})^d > 0$ is a constant. Thus:

$$|e'_{f|s} - e_{f|s}| \leq \Lambda. \tag{20}$$

If we use $e''_{s|r}$ to generate images, we can alter Eq. (20) to:

$$|e''_{f|s} - e_{f|s}| = \Lambda. \tag{21}$$

Similar to Eq. 21, we can have:

$$|e''_{s|r} - e_{s|r}| = \Lambda_s, \tag{22}$$

where Λ_s is also a sequence of positive constants. Then we have:

$$\frac{|e''_{f|s} - e_{f|s}|}{|e''_{s|r} - e_{s|r}|} = \frac{\Lambda}{\Lambda_s} = M. \tag{23}$$

Due to the findings that semantic features in deep feature space are usually linearized [1, 11, 13], we assume semantic features for texts and images are linearized. Following Eq. (**??**), we can further have that:

$$|\delta| \frac{|e'_{f|s} - e_{f|s}|}{|e'_{s|r} - e_{s|r}|} = \frac{|e''_{f|s} - e_{f|s}|}{|e''_{s|r} - e_{s|r}|} = \frac{M}{|\delta|} \leq K, \text{ s.t. } e'_{s|r} \neq e_{s|r}, \tag{24}$$

where δ is a non-zero coefficient. Finally, $e'_{f|s} = E_I(G(e'_{s|r}))$, $e_{f|s} = E_I(G(e_{s|r}))$ where E_I is the image encoder, we have:

$$\frac{|E_I(G(e'_{s|r})) - E_I(G(e_{s|r}))|}{|e'_{s|r} - e_{s|r}|} \leq K, \text{ s.t. } e'_{s|r} \neq e_{s|r}, \tag{25}$$

where it meets Lipschitz condition. □

Proposition A.5. *L_r provides tighter image semantic constraints than L_{db} [6] which is defined as:*

$$L_{db} = 1 - \frac{(e'_{s|r} - e_{s|r}) \cdot (e'_{f|s} - e_{f|s})}{||(e'_{s|r} - e_{s|r})||^2 \cdot ||(e'_{f|s} - e_{f|s})||^2} \ , \tag{26}$$

Proof. For Eq. (**??**), assume $L_{db} = 0$ and use ϵ^*, combining with Eq. (**??**), we have:

$$\frac{(e''_{f|s} - e_{f|s})}{||e''_{f|s} - e_{f|s}||^2} = \frac{||e''_{s|r} - e_{s|r}||^2}{(e''_{s|r} - e_{s|r})} \tag{27}$$

$$= \frac{||\beta \cdot \epsilon^* \odot d(\mathbb{C}_{ss|r})||^2}{\beta \cdot \epsilon^* \odot d(\mathbb{C}_{ss|r})} \ . \tag{28}$$

Therefore:

$$|e''_{f|s} - e_{f|s}| = ||e''_{f|s} - e_{f|s}||^2 \cdot \frac{||\beta \cdot \epsilon^* \odot d(\mathbb{C}_{ss|r})||^2}{|\beta \cdot \epsilon^* \odot d(\mathbb{C}_{ss|r})|} \geq 0 \ . \tag{29}$$

where preservation of semantic collapse is not guaranteed due to the distance between $e'_{f|s}$ and $e_{f|s}$ is not contained. This infers that when two slightly semantic distinct textual embeddings are given, the generated images' semantics can also be the same.

Assume $L_r = 0$, we have:

$$|e''_{f|s} - e_{f|s}| = |\epsilon^*| \odot \beta \cdot d(\mathbb{C}_{rr|s}) \tag{30}$$

$$= \beta \cdot d(\mathbb{C}_{rr|s}) \tag{31}$$

$$> 0 \ , \tag{32}$$

where provides tighter constraints than L_{db}. Note that the proof details of $\beta \cdot d(\mathbb{C}_{rr|s}) > 0$ can refer to Appendix Proposition A.5. □

B MORE DETAILS OF APPLICATION

If there is no specification, for all frameworks, we use their original losses $L(\theta, X)$ and $L(\theta, X')$ with *GisC*: L_{db} or L_r. See specified parameter settings in Table 1. We then demonstrate detailed implementations for tested frameworks. Note that since ITA_C needs no more training, thus model with ITA_C requires no more implementation of L_S in Eq. (??) and L_{ITA_T} in Eq. (??).

Section B.1 are implementations of L_S in Eq. (??) and L_{ITA_T} in Eq. (??) with different backbone: AttnGaN [15], DF-GAN [10] and VQ-GAN + CLIP [14].

B.1 Applying ITA_T

B.1.1 Applying ITA_T to DF-GAN. DF-GAN [10] is a currently proposed one-way output T2Isyn backbone. For $L_{DF_D} = (\theta_{DF}, \cdot)$ of DF-GAN's Discriminator D_{DF}, we use it as L_S for DF-GAN:

$$\begin{aligned} L_{S-DF_D} = & L_{DF_D}(\theta_{DF_D}, (e_{s|r}, \mathcal{G}_{DF}(e_{s|r}))+ \\ & L_{DF_D}(\theta_{DF_D}, (e'_{s|r}, \mathcal{G}_{DF}(e'_{s|r}))+ \\ & L_{DF_D}(\theta_{DF_D}, (e_{s|r}, \mathcal{G}_{DF}(e'_{s|r})) \ , \end{aligned} \tag{33}$$

where L_{DF} is the simplified representation for DF-GAN's original Discriminator losses; $\mathcal{G} = h_{DF}(G_{DF}(.))$ where $(.)$ takes a textual embedding, h_{DF} maps generated images of (G_{DF} on the textual embedding. Notations in the following frameworks are similar. All embeddings used in DF-GAN are gained from DAMSM images and text encoders.

Then for generator G loss $L_{DF_G}(\theta_G, \cdot)$, we have loss:

$$\begin{aligned} L_{S-DF_G} = & L_{DF_G}(\theta_{DF_G}, (e_{s|r}, \mathcal{G}_{DF}(e_{s|r})))+ \\ & L_{DF_G}(\theta_{DF_G}, (e'_{s|r}, \mathcal{G}_{DF}(e'_{s|r})))+ \\ & L_{DF_G}(\theta_{DF_G}, (e_{s|r}, \mathcal{G}_{DF}(e'_{s|r}))). \end{aligned} \tag{34}$$

Since DF-GAN only uses one discriminator D_{DF} for both semantic matching and image quality supervision. Therefore, we can use $L_{DF_D}(\theta_{DF}, \cdot)$ to force $\mathcal{G}_{DF}(t'_s)$ be consistent with t_s by optimizing parameters r of ITA_T :

$$\begin{aligned} L_{ITA_T-DF} = & r \cdot L_{iemse}(e_{s|r}, e'_{s|r})+ \\ & (1-r) \cdot L_{DF_G}(\alpha, (e_{s|r}, \mathcal{G}_{DF}(e'_{s|r})) \ . \end{aligned} \tag{35}$$

B.1.2 Applying ITA_T to AttnGAN. AttnGAN [15] is a widely used backbone for GAN-based text-to-image generation baseline. Since AttnGAN uses both sentence-level $e_{s|r}$ and word-level semantics e_w embeddings, we implement augmented sentence and words $e'_{s|r}, e'_w$ as $e'_{s|r} = ITA_T(e_{s|r})$, $e'_w = e_w + (e'_{s|r} - e_{s|r})$. Other implementations refer to Section B.1.1.

All embeddings used in AttnGAN are gained from DAMSM images and text encoders.

B.1.3 Applying ITA_T to VQ-GAN + CLIP. We use the released checkpoint and code of [14] for tuning. Notice the [14] is originally trained on the clip embeddings of images; we directly altered it by using textual CLIP embeddings. We only tune the transformer part for organizing the discrete code, while the image-generating decoder part is fixed. Due to the long training time, we only tune the model for 20 epochs with L_r and use its original $L_{vqclip}(\theta, X)$ for our augmented X' as $L_{vqclip}(\theta, X')$. Other settings follow the original settings. All embeddings used in VQ-GAN + CLIP are gained from CLIP images and text encoders. We only test $ITA_C + L_r$ with VQ-GAN + CLIP due to its long training time.

B.1.4 ITA_T Implementation Suggestions. Training ITA_T with adversarial models needs concern about how to avoid exploding gradient. Because in the early stage, the discriminators may not provide meaningful semantic bounding on ITA_T, causing the augmented $e'_{s|r}$ located too far from $e_{s|r}$ and then a too large loss for generators which cannot be optimized.

Thus we suggest a warmup phase before training ITA_T. For AttnGAN, we set a warmup phase to avoid this kind of crush. Due to DF-GAN using hinge losses, which cannot be larger than one, it can have no warmup phase. Refers to Table 1 for more parameter details.

We also suggest scaling the learning rate up when training with ITA with L_{db} or L_r due to their regularity.

B.2 Applying ITA_C and L_r

ITA_C and L_r are based on $\mathbb{C}_{ss|r}$ and $\mathbb{C}_{rr|s}$ defined as:

$$\mathbb{C}_{ss|r} = \mathbb{C}_{ss} - \mathbb{C}_{sr}\mathbb{C}_{rr}^{-1}\mathbb{C}_{rs}, \tag{36}$$

$$\mathbb{C}_{rr|s} = \mathbb{C}_{rr} - \mathbb{C}_{rs}\mathbb{C}_{ss}^{-1}\mathbb{C}_{sr} \ . \tag{37}$$

We only used 30K random samples from CUB and COCO training sets, respectively, to obtain $\mathbb{C}_{ss|r}$ and $\mathbb{C}_{rr|s}$ for our experiments. The number of samples follows it of calculating FID [7]. It is rational to scale the number of samples up according to the size of the dataset. Nevertheless, we do not recommend using the whole training set for the calculation due to its memory consumption. Our calculated $\mathbb{C}_{ss|r}$ and $\mathbb{C}_{rr|s}$ will be released with our code.

B.3 Applying SADA to Diffusion Models

For Diffusion models, *ITA* should be applied to conditional embeddings (including textual conditional embeddings). The *GisC* should be applied to features of generated images at each step.

C MORE EXPERIMENTAL DETAILS

C.1 Experimental Settings

We compare Mixup [16], Random Mask, Add Noise, with our proposed ITA_T, ITA_C and L_r by applying them to AttnGAN [15], DF-GAN [10] with DAMSM encoders [15] We evaluate each framework on two benchmark datasets, CUB [12] and COCO [9]. To delve deeper, we apply ITA_C with L_r to VQ-GAN+CLIP [14] with CLIP on COCO and the conditional DDPM [8] by specifically utilizing the MNIST dataset [4].

Specifically, experiments based on the conditional DDPM [8], specifically utilizes the MNIST dataset [4]. The methodology applied involved incorporating our proposed ITA_C on condition embeddings, with further integration of L_r on calculated feature shift of generated images from U-Net's bottleneck. We first train the bare diffusion model and then use its bottleneck's hidden feature as $e_{s|r}$ and the bottleneck's hidden feature of the next step as $e_{f|s}$. Then other details will be as same as aforementioned.

Table 1: Parameters for experiments.

Dataset: CUB	+ ITA_T	
Backbone	Warm-up	r
AttnGAN	50	0
DF-GAN	100	0.2
Dataset: COCO	+ ITA_T	
Backbone	Warm-up	r
AttnGAN	0	0
DF-GAN	0	0.2
Dataset: CUB	+ ITA_C	
Backbone	β	r
AttnGAN	0.05	0
DF-GAN	0.05	0.2
Dataset: COCO	+ ITA_C	
Backbone	β	r
AttnGAN	0.01	0
DF-GAN	0.01	0.2
VQ-GAN + CLIP	0.05	0.2
Dataset: CUB	+ ITA_C + L_r	
Backbone	φ	
AttnGAN	0.01	
DF-GAN	0.01	
Dataset: COCO	+ ITA_C + L_r	
Backbone	φ	Learning Rate
AttnGAN	0.001	As original
DF-GAN	0.001	Doubled
VQ-GAN + CLIP	0.05	As original

Parameter settings follow the original models (including augmentations they used) of each framework for all experiments unless specified. For training settings, we train the model from scratch and also use their released model for tuning experiments. The discriminators are retrained during the tuning process for AttnGAN and DF-GAN since no released checkpoints are available; and we only tune the transformer part of experimental settings, which are specified in Table **??**. Notice that we do not conduct ITA_T with L_r because $e'_{s|r} - e_{s|r} \leq \epsilon \odot \beta \cdot d(\mathbb{C}_{ff|s})$ in Eq. (**??**) is required for L_r. See more training details and model parameter settings in Appendix C.

Especially, $\mathbb{C}_{ss|r}$ for ITA_C and $\mathbb{C}_{rr|s}$ for L_r are calculated on the training set using the encoders of the framework. We use 30K random samples from each dataset in our experiments. Limited sampling also leads to the possible implementation of ITA_C and L_r on super-large datasets.

C.2 Implementing Mixup with AttnGAN and DF-GAN

We use the official code of Mixup [16] for our experiments. Augmented images and mixed textual embeddings are used for model training. All model settings follow the original settings of AttnGAN and DF-GAN.

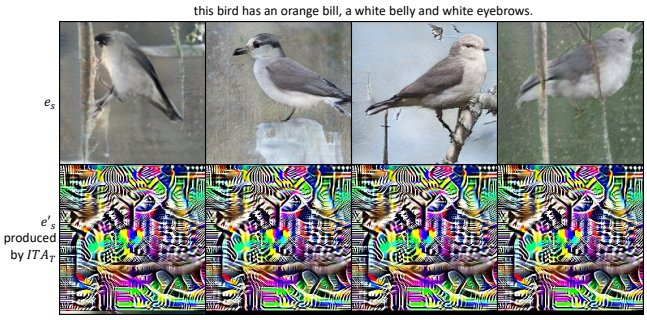

Figure 1: Collapse examples of DF-GAN with $ITA_T + L_{db}$ using $\alpha = 0.3$ generated on $e_{s|r}$ and augmented $e'_{s|r}$.

C.3 Implementing with AttnGAN and DF-GAN

We randomly mask 15% tokens and use the original settings of AttnGAN and DF-GAN. AttnGAN with Random Mask collapsed multiple times during the training. We use the checkpoints that were saved before the collapse to resume the training.

C.4 Parameter Settings

We train each backbone from the start on the CUB dataset and tune their released checkpoint on the COCO dataset. Due to no released checkpoints for discriminators of AttnGan and DF-GAN, we retrain discriminators during the tuning phase. If there is no specification, we follow the original experimental settings of each backbone. Specified parameters used for producing final results in the paper are shown in Table 1. Notice that r for ITA_T can be set to zero due to the weak supervision of generative adversarial networks. Specifically, we double the learning rate for $ITA_C + L_r$ tests due to their regularity.

Table 2: Results of DF-GAN with $ITA_T + L_{db}$ using different r values on the CUB dataset, training within 600 epochs.

r	CS	FID
0	0.5791	13.96
0.1	0.5793	12.7
0.2	**0.5807**	**11.74**
0.3	ITA_T collapses	

D MORE RESULTS

D.1 More Qualitative Results

D.1.1 ITA_T with different r . As stated, r in Eq. (**??**) can control the augmentation strength. Larger r in ITA_T leads to more intensive augmentation. However, as shown in Table 2, an inappropriate large r can cause model collapse because $S(\alpha, (e_{s|r}, \mathcal{G}(e'_{s|r}))$ will lose its constraint, causing that $e'_{s|r}$ is too different from $e_{s|r}$ (i.e., $e'_{s|r}$ cannot maintain the main semantics of $e_{s|r}$). Collapse examples are shown in Figure 1. It can be seen that using $\alpha = 0.3$ ITA_T cannot produce a semantic maintaining $e'_{s|r}$ for G. Within the appropriate range, larger r offers better text-image consistency and image quality.

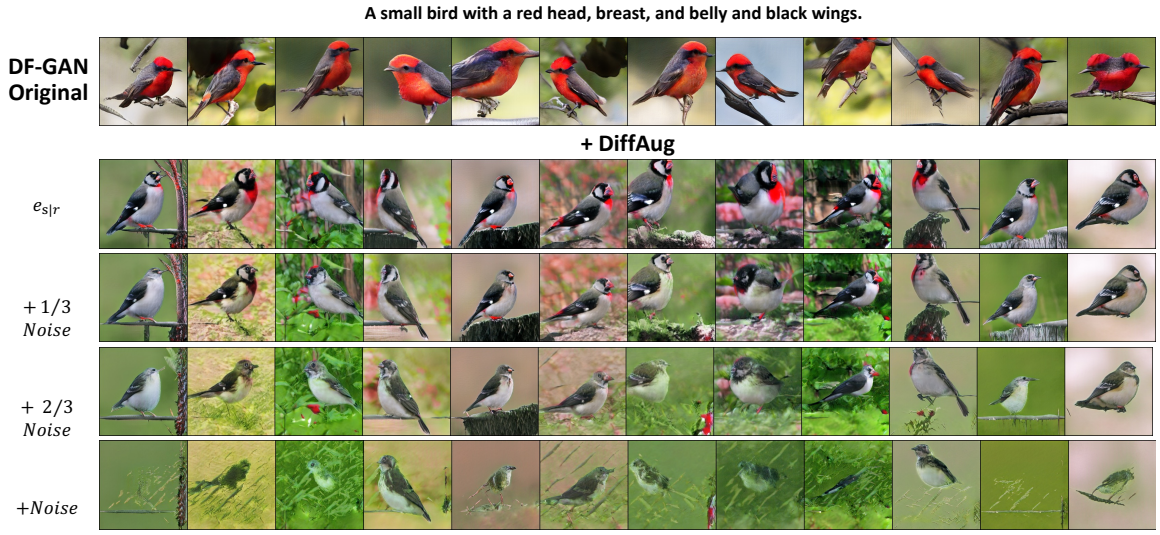

Figure 2: Generated images of DF-GAN with DiffAug on CUB dataset with ascending scales of ϵ are added.

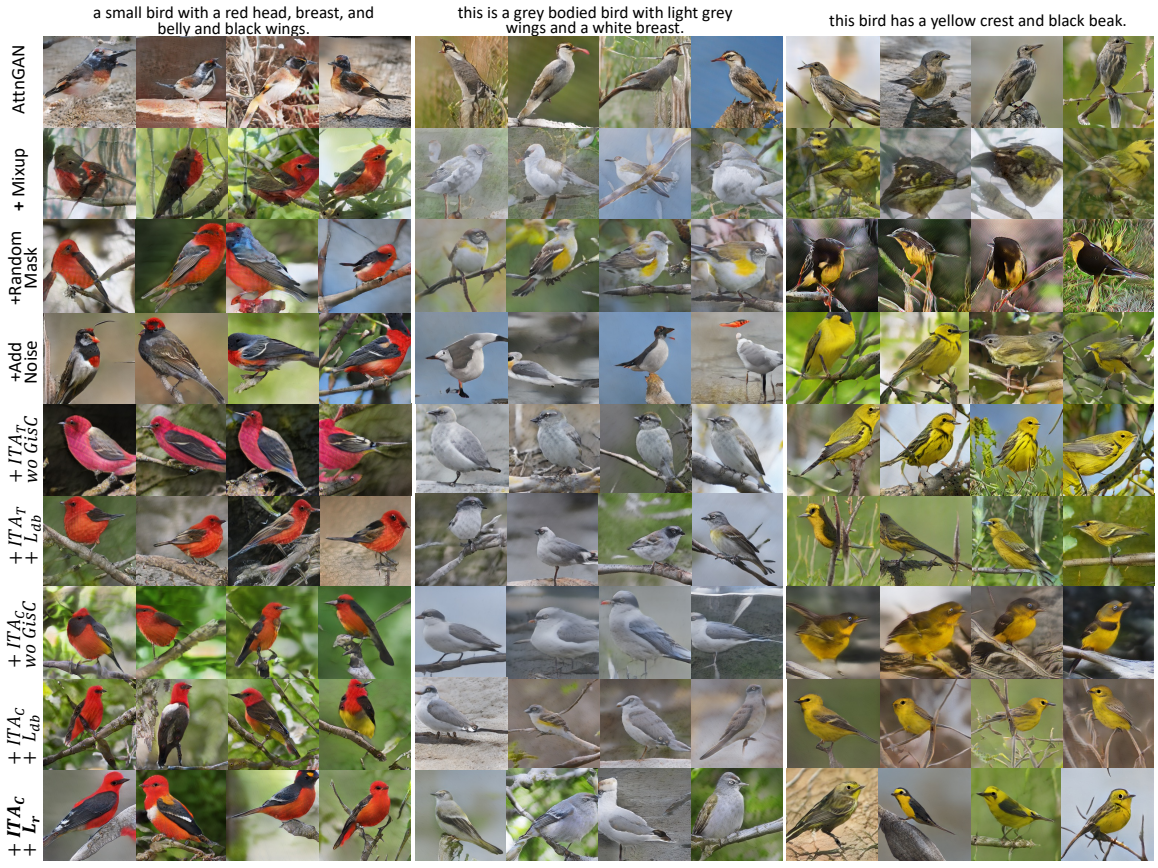

Figure 3: Generated results of AttnGAN on CUB.

D.1.2 Semantic Collapse in DF-GAN with DiffAug. We exhibit generated images of DF-GAN with DiffAug on the CUB dataset. As the ascending scales of ϵ are added, the images illustrate the occurrence of semantic collapse.

D.2 More Quantitative Results

We show more generated examples of AttnGAN [15] as Figure 3 and Figure 4, DF-GAN [10] as Figure 5 and Figure. 6, and VQ-GAN + CLIP [14] on COCO as Figure 7. The diversity improvement is more obvious for AttnGAN than other backbones.

REFERENCES

[1] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. 2013. Better mixing via deep representations. In *International Conference on Machine Learning*. PMLR, 552–560.

[2] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. 2020. A group-theoretic framework for data augmentation. *The Journal of Machine Learning Research* 21, 1 (2020), 9885–9955.

[3] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. 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.

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

[5] Hao Dong, Jingqing Zhang, Douglas McIlwraith, and Yike Guo. 2017. I2t2i: Learning text to image synthesis with textual data augmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015–2019.

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

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[10] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 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.

[11] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. 2017. Deep feature interpolation for image content changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7064–7073.

[12] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[13] Yulin Wang, Gao Huang, Shiji Song, Xuran Pan, Yitong Xia, and Cheng Wu. 2021. Regularizing deep networks with semantic data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[14] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. 2022. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP. *arXiv preprint arXiv:2203.00386* (2022).

[15] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 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.

[16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

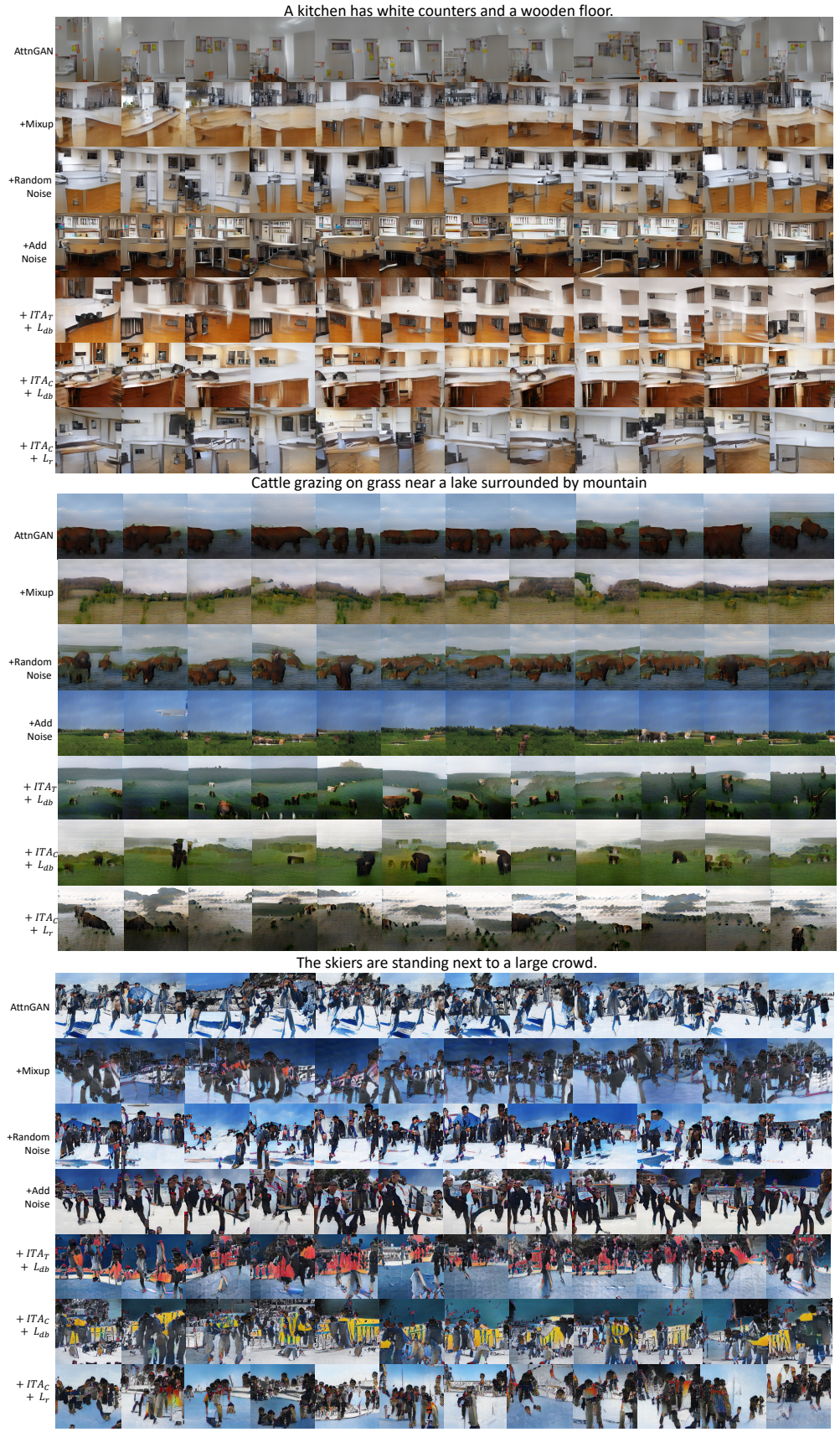

Figure 4: Generated results of AttnGAN on COCO.

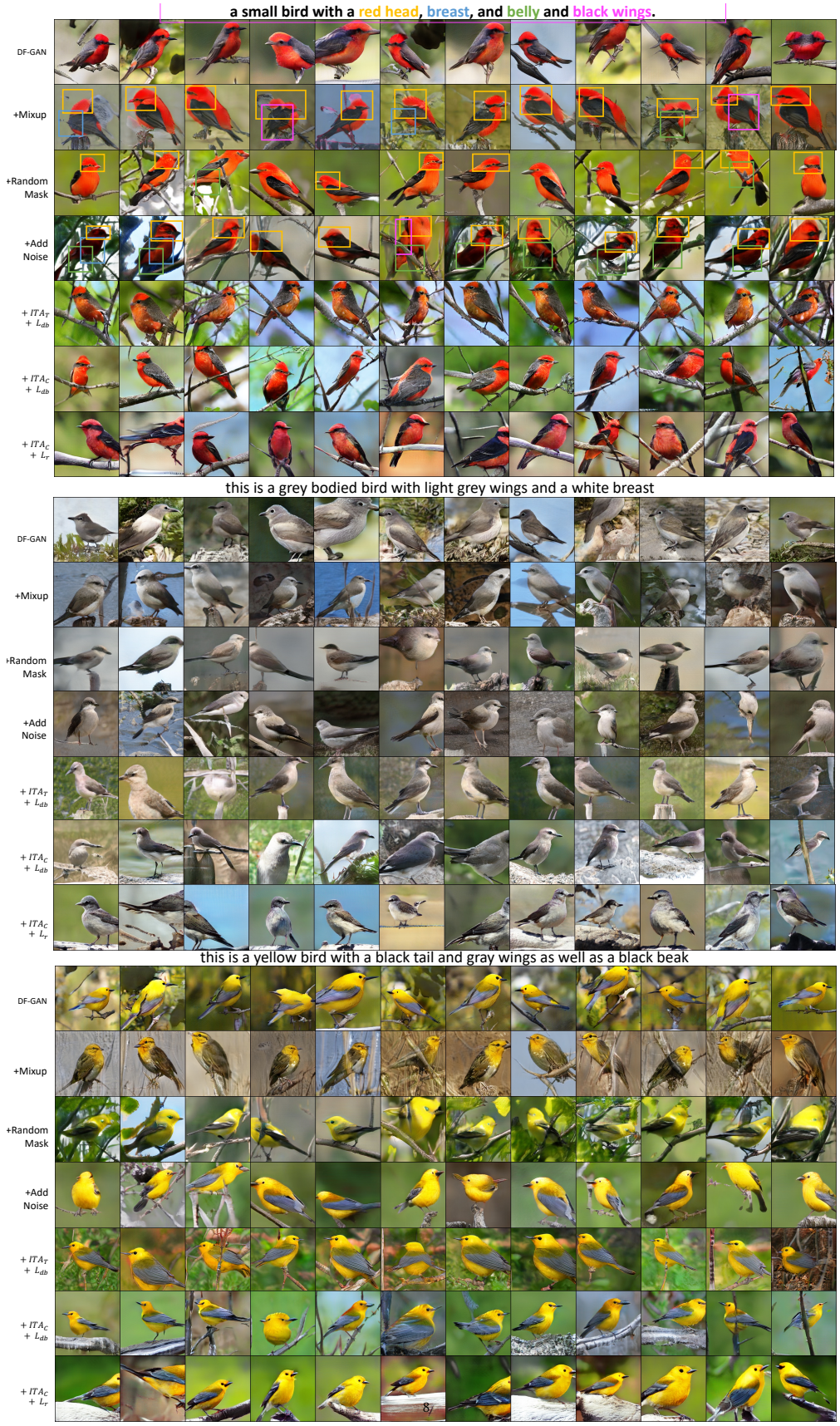

Figure 5: Generated results of DF-GAN on CUB. Semantic mismatches and image quality degradation are highlighted for generated results of DF-GAN w/ Mixup, w/ Random Mask, and w/ Add Noise in the top group.

Figure 6: Generated results of DF-GAN on COCO.

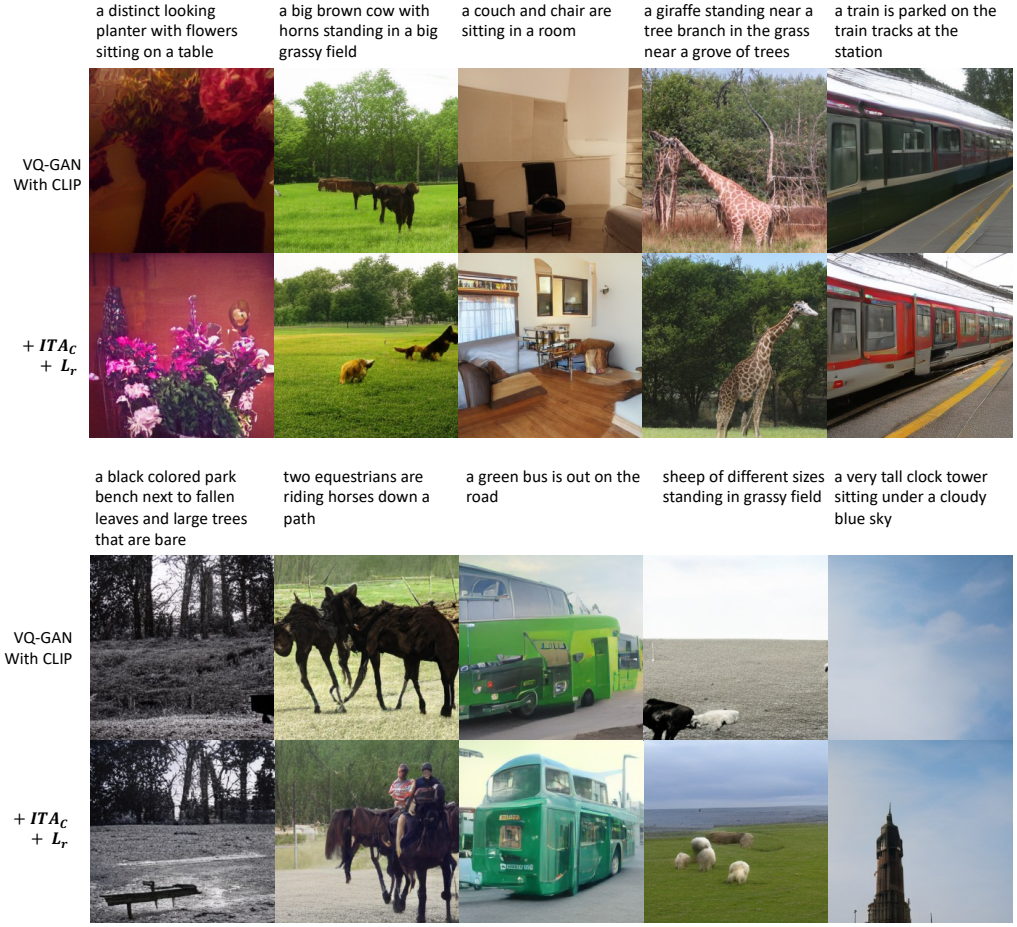

Figure 7: Generated results of VQ-GAN + CLIP on COCO.