Dual Orthogonal Guidance for Robust Diffusion-based Handwritten Text Generation

Konstantina Nikolaidou[1]* George Retsinas[2]* Giorgos Sfikas[3]
Silvia Cascianelli[4] Rita Cucchiara[4] Marcus Liwicki[1]

[1]Luleå University of Technology [2]National Technical University of Athens
[3]University of West Attica [4]University of Modena and Reggio Emilia

Abstract

Diffusion-based Handwritten Text Generation (HTG) approaches achieve impressive results on frequent, in-vocabulary words observed at training time and on regular styles. However, they are prone to memorizing training samples and often struggle with style variability and generation clarity. In particular, standard diffusion models tend to produce artifacts or distortions that negatively affect the readability of the generated text, especially when the style is hard to produce. To tackle these issues, we propose a novel sampling guidance strategy, Dual Orthogonal Guidance (DOG), that leverages an orthogonal projection of a negatively perturbed prompt onto the original positive prompt. This approach helps steer the generation away from artifacts while maintaining the intended content, and encourages more diverse, yet plausible, outputs. Unlike standard Classifier-Free Guidance (CFG), which relies on unconditional predictions and produces noise at high guidance scales, DOG introduces a more stable, disentangled direction in the latent space. To control the strength of the guidance across the denoising process, we apply a triangular schedule: weak at the start and end of denoising, when the process is most sensitive, and strongest in the middle steps. Experimental results on the state-of-the-art DiffusionPen and One-DM demonstrate that DOG improves both content clarity and style variability, even for out-of-vocabulary words and challenging writing styles.

1. Introduction

Handwritten Text Generation (HTG), or Styled HTG, is a task that has only relatively recently gained traction compared to the more “traditional” Document Imaging tasks, like Handwritten Text Recognition (HTR) [6, 22, 39–41] or Keyword Spotting (KWS) [19, 37, 38], in terms of effective models and methods. One motivation for HTG systems is user personalization in digital applications, where it can be useful in aiding individuals with physical impairments to produce handwritten notes in a personalized style. Another crucial goal for HTG is to play the role of an efficient tool for data augmentation, acting in an auxiliary way for other main, downstream Document Analysis tasks [10]. This is especially useful in low-resource contexts [31], where inadequately documented scripts or languages with few writers or with few digitized, annotated training samples constitute a serious impediment to creating automated document-imaging tools.

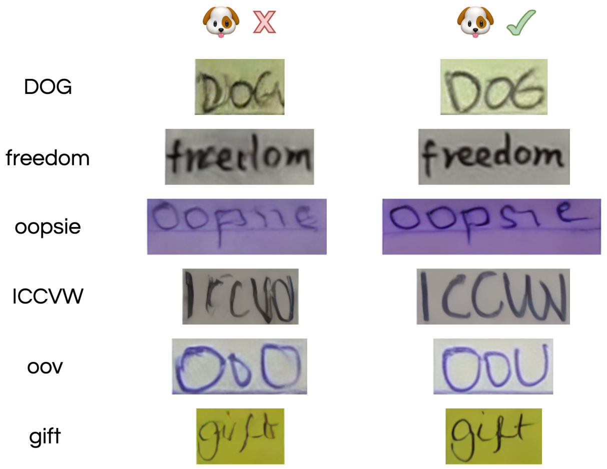

Figure 1. Qualitative examples of generation without (left) and with (right) our proposed *DOG* guidance strategy.

Diffusion-based HTG models have recently shown promising results in producing readable handwritten images replicating an existing style. However, two major issues persist:

1. They frequently generate artifacts that degrade content clarity, even for common, in-distribution examples.

[*]equal contribution

2. They tend to memorize style-content pairs seen during training, making them brittle in low-data regimes or for unseen combinations.

To partially address the first issue, Classifier-Free Guidance (CFG) [17] has been proposed as a sampling method that interpolates between unconditional and conditional predictions, typically improving alignment with the target condition. However, CFG introduces a trade-off: high guidance scales may help with clarity but often result in oversaturation or degraded detail.

In this work, we introduce Dual Orthogonal Guidance (DOG), a guidance strategy that pushes generation along an orthogonal direction derived from a negatively perturbed version of the conditioning prompt. The idea is to encourage clearer generation by suppressing entangled distortions from a perturbed condition (see Fig. 1), while still enabling structured variation. Unlike CFG, which relies on unconditional sampling, DOG uses a negative prompt derived from the actual condition (*e.g.*, noised style or content), which is then orthogonalized with respect to the original. This produces a more targeted and stable modification of the sampling trajectory. To control the influence of this guidance throughout the diffusion process, we use a triangular schedule, which limits the effect at early and late steps, where the process is most sensitive, and maximizes it in the middle steps, where it can influence the structure without causing instability.

In particular, our main contributions are the following:

- *Dual Orthogonal Guidance (DOG):* A test-time sampling strategy that introduces an orthogonal direction derived from a negative prompt, enabling clearer generation and controlled variability.
- *Stability through triangular scheduling:* We modulate the guidance scale across timesteps, peaking mid-process to avoid distorting global structure or introducing artifacts.
- *Plug-and-play integration:* DOG is model-agnostic, requires no retraining, and can be applied directly to pretrained diffusion-based HTG models.
- *Compelling Performance:* We beat the state-of-the-art in terms of posterior sampling variety, with tangible benefits in terms of qualitative and quantitative results.

The remainder of this paper is structured as follows. In Section 2, we review the related work. In Section 3, we describe the proposed Dual Orthogonal Guidance. Section 4 presents numerical experiments and showcases illustrative qualitative results. Finally, we conclude the paper with Section 5, where we summarize our contributions and discuss future work.

2. Related work

Styled HTG & Classifier-free Guidance. *Styled HTG* is the task of generating handwritten images given a style and a content condition. Early HTG methods relied on extensive hand-crafted feature extraction and text resynthesizing using rule-based methods [24, 46, 48]. Later, recurrent architectures demonstrated the ability to generate handwriting [15]. Building on early methods, GAN-based techniques [1, 9, 14, 20, 21, 28, 43], including several that integrate transformer architectures [4, 35, 47], have enhanced handwriting generation. However, these approaches still face challenges such as training instability and limited diversity [2, 13, 30]. More recently, Diffusion Models [5, 8, 16, 32, 33, 49] have been put forward as an alternative to GAN-based models. While the paradigm of diffusion forms a cohesive foundation for styled HTG, several problems have proved to be non-straightforward to solve in a satisfactory manner, including style variability, rare character combinations, and training data memorization. *Classifier-free guidance (CFG)* [17] has shown improvement in the quality of the generation by performing a linear interpolation between the conditional and unconditional estimations, $\lambda\epsilon(x|c) + (1 - \lambda)\epsilon(x|\emptyset)$. A few works have showcased the effect of CFG in HTG by exploring the standard approach [5, 8, 11, 29]. However, only [5] explores the effect of CFG on the generation.

Negative Prompting in Diffusion Models. Diffusion Models are state-of-the-art latent variable generative models that are consistently setting new benchmarks in numerous and diverse tasks [44], including those that pertain to one or another form of prompting, in the sense of a textual condition to the model [7, 18]. *Negative Prompting* seeks to steer the guidance away from unwanted attributes. In [25], an objective that may include multiple conditions is broken down to a sum of composing directions in the latent space. Unwanted conditions are then assigned a negative weight, in principle allowing the end-user to specify the set of conditions at will. This translates to a simple but very effective test-time algorithm. Perp-Neg [3] is another sampling algorithm for standard Text-to-Image Diffusion Models, which computes a negative gradient that is perpendicular to the main prompt. The rationale related to this choice is that two conditions should not be taken *a priori* to be conditionally independent; moving in the orthogonal direction ensures that unwanted details are suppressed without interfering with the primary semantic content. Both works underscore the importance of carefully adjusting the positive and negative cues to enhance the fidelity and controllability of the generated images. Our proposal is inspired by this previous work, and puts forward a technique that is also based on test-time guidance of sampling.

Interestingly, a very recent and independent work also explores the use of orthogonal projections in diffusion guidance [42]. This method, named Adaptive Projected Guidance (APG), applies a perpendicular projection of the unconditional CFG term to improve saturation in text-to-

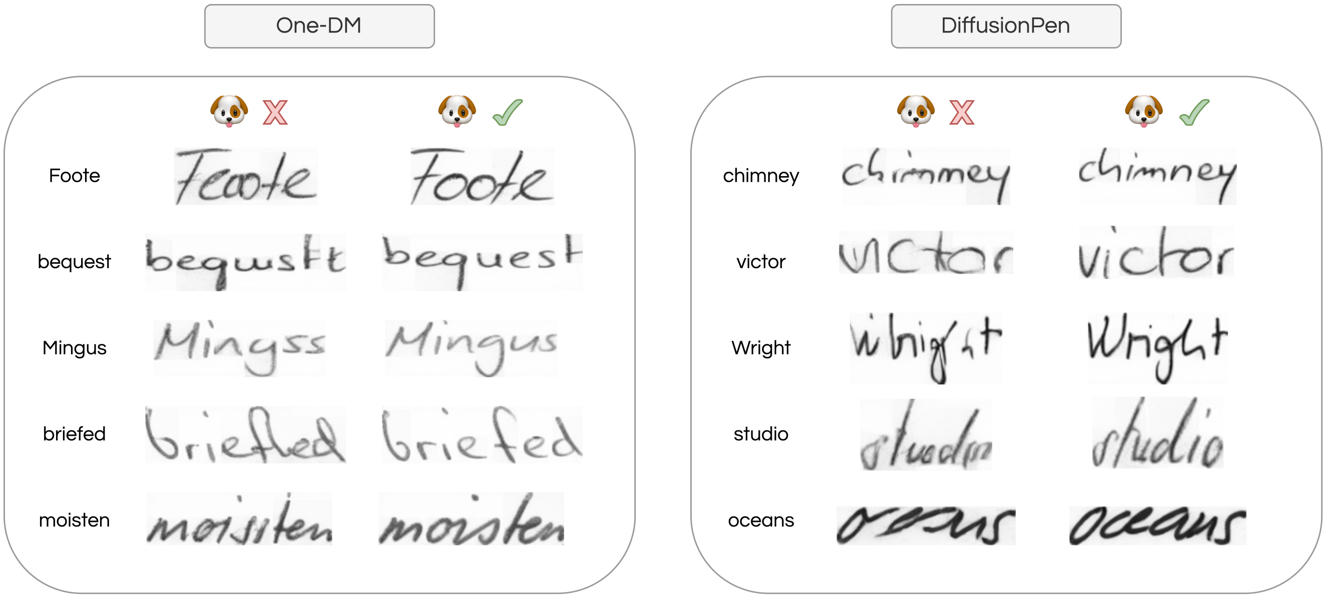

Figure 2. Qualitative results of the application of the proposed DOG guidance strategy at inference time to the Diffusion-based OneDM and DiffusionPen HTG approaches.

image generation for the case of high guidance scales. While our approach was developed independently and is specifically tailored to HTG, both our work and [42] highlight the benefit of introducing orthogonal components to steer the generative process without conflicting with the main conditioning. However, unlike DOG, APG further assigns a small weight to the parallel component and lacks a scheduling mechanism to modulate the influence of guidance across denoising steps. As we show, the absence of such scheduling can make guidance scale selection delicate, introducing a trade-off between control and artifact-free generation. In our work, we compare both existing guidance strategies with our proposed DOG.

3. Dual Orthogonal Guidance (DOG)

3.1. Preliminaries on Diffusion Models

Diffusion models are a class of generative models that synthesize data by learning to reverse a fixed noise process through iterative denoising steps. Denoising Diffusion Probabilistic Models (DDPM) [44], a widely used instantiation of this idea, define a latent variable model $p(x) = \int p(x,z)\,dz$, where $z = \{z_1, z_2, \dots, z_T\}$ is a latent Markov chain over T timesteps. The forward process $q(z|x)$ gradually corrupts data with Gaussian noise in a variance-preserving manner, and the reverse process is learned to approximate $p(x|z)$. This formulation resembles a hierarchical Variational Autoencoder [26]. In our case, we adopt the DDIM formulation [45], a deterministic alternative to DDPM, while keeping the same core noise prediction structure.

In the Styled HTG setting, the conditioning c typically consists of the desired content c_t (*e.g.*, the target text) and the style c_s (*e.g.*, writer identity or visual features). Sampling is performed via ancestral denoising, starting from pure noise $z_0 \sim \mathcal{N}(0, I)$ and progressively refining it through a learned reverse process:

$$x \sim p(x|c) = \int p(x|z_1)\,p(z_1|z_2)\cdots p(z_T|z_0)\,p(z_0)\,dz_{1:T} \quad (1)$$

We use the notation $p(x|c)$ to denote this generative process conditioned on the *dual* c. During training, the model learns to predict the noise ϵ_θ added to a sample x_0 at timestep t, by using a noised input x_t and conditions c. At inference, this prediction is used to iteratively reconstruct the image from noise.

3.2. Motivation and Synopsis

In HTG, modifying the conditioning inputs can introduce variation in the generated image, but often in an unstructured and entangled way. Since content and style representations are not fully disentangled in practice, perturbing the style component c_s can inadvertently degrade the fidelity of the content c_t, resulting in unclear or semantically corrupted outputs.

To address this, we move beyond naive perturbation. Rather than using the perturbed condition directly, which may conflate content and style, we define a guidance mechanism that encourages the generation to be faithful to the intended (c_t, c_s) condition, while actively steering it away from a negative pairing. This negative pairing is formed by corrupting one of the conditions (or both of them) and acts as a counterexample. This builds naturally on the CFG framework, which interpolates between unconditional and conditional predictions to enhance content clarity. Instead of using an unconditional signal, we use a corrupted condition and isolate its influence by projecting out the component aligned with the original (positive) prediction. The remaining orthogonal direction provides a controlled, content-preserving signal that still allows for variability.

The core challenge is how to construct this orthogonal direction from a noisy negative prompt in a way that maintains meaningful structure and avoids pushing the model toward unrealistic outputs. In the next section, we define this construction formally and explain how it is integrated into the diffusion trajectory.

3.3. Subspace Projection

We define the positive and negative dual prompts based on the content-style condition. Let r_t and r_s denote the representations for c_t and c_s, respectively. The clean pair (r_t, r_s) serves as the positive prompt.

To simulate a negative pair, we generate noisy variants of the content and style representations by applying element-wise dropout and scaled Gaussian noise:

$$\tilde{r}_s = \lambda_s \cdot \eta_s \cdot \mathcal{N}(0, I), \tag{2}$$
$$\tilde{r}_t = \lambda_t \cdot \eta_t \cdot \mathcal{N}(0, I), \tag{3}$$
$$\eta_s, \eta_t \sim \text{Bernoulli}(p),$$

where λ_s and λ_t control the magnitude of the noise, and p determines the sparsity of the active dimensions through dropout masks. This formulation enables selective corruption of latent attributes, encouraging the model to explore attribute-specific guidance paths rather than uniformly noisy directions. Empirically, we find that this stochastic masking mechanism yields more informative contrastive gradients. Depending on the intended contrastive setup, one may perturb either the style or content representation independently, or both jointly, while keeping the other component fixed from the original (positive) pair.

Given the perturbed condition, we compute two noise predictions:

$$\epsilon_p = \epsilon(x_t, t, r_t, r_s), \tag{4}$$
$$\epsilon_n = \epsilon(x_t, t, \tilde{r}_t, \tilde{r}_s). \tag{5}$$

While ϵ_n introduces perturbations, it likely contains both structured and unstructured deviations from ϵ_p. To isolate a direction that influences generation without corrupting the core signal, we subtract the projection of ϵ_n onto ϵ_p:

$$\epsilon^* = \epsilon_n - \text{proj}_{\epsilon_p}(\epsilon_n). \tag{6}$$

where the projection term is given by:

$$\text{proj}_{\epsilon_p}(\epsilon_n) = \frac{\langle \epsilon_n, \epsilon_p \rangle}{\|\epsilon_p\|^2} \cdot \epsilon_p. \tag{7}$$

This orthogonal component ϵ^* captures contrastive variation while remaining disjoint from the intended generation direction.

To ensure numerical stability and prevent degenerate behavior at large magnitudes, we clip the norm of ϵ_n, before the projection step, using a threshold τ:

$$\epsilon_n \leftarrow \min\left(1, \frac{\tau}{\|\epsilon_n\|}\right) \cdot \epsilon_n. \tag{8}$$

The final denoising prediction becomes:

$$\hat{\epsilon} = \epsilon_p + g(t) \cdot (\epsilon_p - \epsilon^*), \tag{9}$$

where $g(t)$ is the time-dependent guidance scale, described in Section 3.4.

3.4. Guidance Scale with Scheduling

The denoising behavior naturally follows a coarse-to-fine progression. The first timesteps of the denoising process are dominated by noise and lack a clear structure, making it undesirable to enforce a strong guidance signal and have a very heavy influence on the decisions. Conversely, in later "cleaner" timesteps, the sample is refined with intricate details, and over-conditioning may hinder the preservation of subtle features and cause the appearance of artifacts. To address this, we influence the guidance signal across timesteps by using a triangular schedule. For every given timestep $t \in [0, T]$, a threshold u_T is defined such that:

$$\gamma(t) = \begin{cases} \dfrac{t}{u_T}, & \text{if } t \leq u_T, \\ 1 - \dfrac{t - u_T}{T - u_T}, & \text{if } t > u_T. \end{cases} \tag{10}$$

Here, u_t controls the location of the peak in the triangle. The overall guidance scale at timestep t is then given by multiplying $\gamma(t)$ by a base guidance factor gs:

$$g(t) = gs \cdot \gamma(t). \tag{11}$$

3.5. Sampling with Dual Orthogonal Guidance

The proposed guidance is applied during each denoising step of the reverse diffusion process. At step t, we compute the standard conditional prediction ϵ_p using the given content-style pair (c_t, c_s), and the negative prediction ϵ_n using the perturbed counterpart (either $\tilde{c}_s$, $\tilde{c}_t$, or both). The orthogonal direction ϵ^* is derived by subtracting the projection of ϵ_n onto ϵ_p, as discussed previously.

The final residual from Eq. (9) is used within the DDIM sampling rule to update the sample x_t:

$$x_{t-1} = \text{DDIMStep}(x_t, \hat{\epsilon}, t), \tag{12}$$

where DDIMStep denotes the deterministic transition rule at timestep t. This operation can be implemented directly using any scheduler that supports DDIM-style updates.

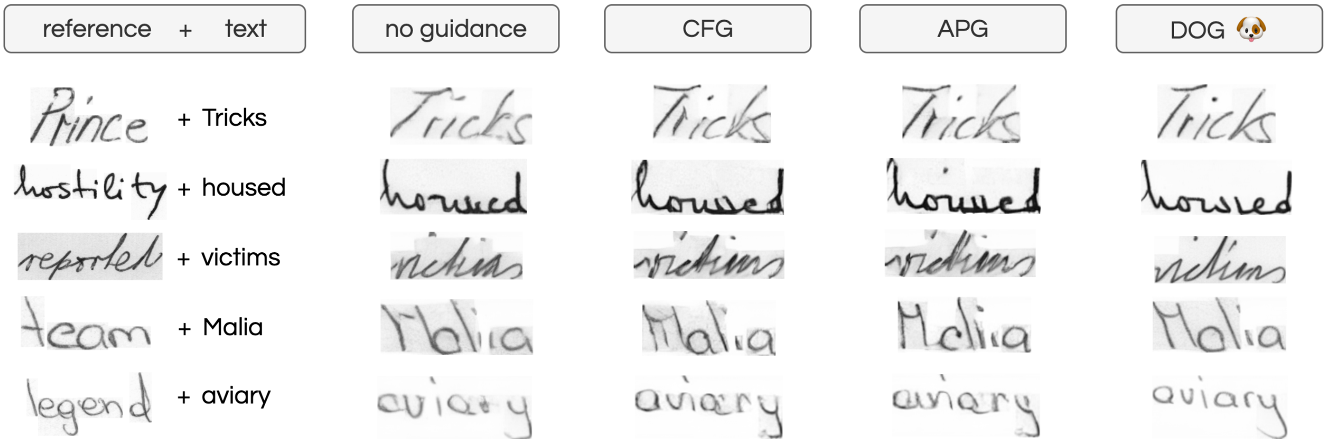

Figure 3. Qualitative comparison between guidance strategies applied to DiffusionPen when generating a target text in different styles.

By leveraging both a faithful prompt and a structured negative variant, the proposed guidance effectively encourages content-clarity and controlled stylistic variation. The orthogonal decomposition ensures that generation is nudged along contrastive directions that preserve semantic fidelity.

4. Experiments

In this section, we evaluate our proposed guidance strategy, DOG, by applying it to pre-trained, off-the-shelf diffusion-based HTG models. We present both *qualitative* and *quantitative* results that demonstrate the effectiveness of DOG in improving generation quality. In addition, we provide ablation studies analyzing the impact of key components and hyperparameters of our method.

4.1. Experimental Setup

We conduct experiments with the two main existing pre-trained diffusion-based HTG backbones, DiffusionPen [33] and One-DM [8], trained on the IAM offline handwriting database [27]. We also use a version of DiffusionPen [33] that is pre-trained on the GNHK dataset [23]. DiffusionPen is a latent-diffusion HTG that deploys a hybrid metric- and classification-style encoder to embed style features in a few-shot setting, while One-DM operates on pixel space in a one-shot style encoding setting.

For comparison with the guidance literature, we further set up CFG [17] and APG [42] on DiffusionPen. To this end, we re-train DiffusioPen [33] with style and content conditions dropped with a probability of 0.2 in order to use the unconditional components necessary for CFG and APG that were not included in the original training. We keep the HTG models in evaluation mode and use them for sampling in their default settings, integrating ours and the competitor guidance strategies as described in Sec. 3. Hence, except for the adaptation of DiffusionPen for comparative reasons, no training process is included. The hyperparameters for DOG are set as follows. For the triangular scheduling, we use a peak threshold timestep of $u_T = 700$. Throughout all experiments, we fix the noise magnitudes to $\lambda_s = \lambda_t = 1000$ and use a keep-probability of $p = 0.75$ for dropout masks.

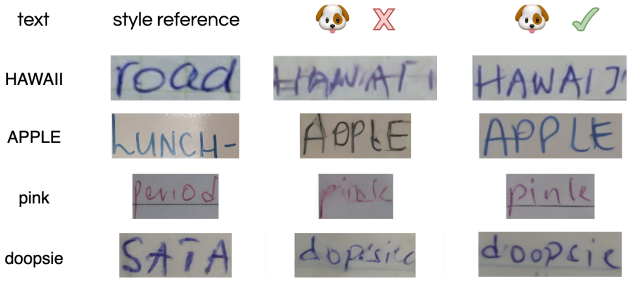

Figure 4. Qualitative examples of the proposed DOG on DiffusionPen when replicating unseen styles from the GNHK dataset.

4.2. Qualitative Results

We present qualitative examples in different scenarios showcasing the effect of our proposed method, focusing on aspects such as content robustness, style replication, and variability, while comparing with other guidance strategies.

Content and Style Robustness. DOG appears to stabilize the generation and produce more accurate text in both seen and unseen style cases. Fig. 2 showcases the effect of DOG on both DiffusionPen [33] and One-DM [8] when generating Out-of-Vocabulary (OOV) words using seen writer styles of the IAM database. Furthermore, Figs. 1 and 4 show how the adaptation of DOG on DiffusionPen for the GNHK dataset [23] fixes hard cases of unseen styles while preserving the intended style. It is clearly observed that DOG enhances the quality of the generation by consistently improving the content of the generated words while keeping the style characteristics close to the initial generation.

Notably, DOG is model-agnostic, yielding similar content improvements when applied to both DiffusionPen [33] and One-DM [8].

To properly assess our method, we also compare it against the alternative guidance strategies CFG [17] and the recently introduced APG [42]. Fig. 3 shows comparative qualitative examples of DiffusionPen in its original form, *i.e.*, without any guidance, and with the additional guidance strategies. It is clear that all guidance strategies assist the content preservation. However, our proposed method is able to generate "cleaner" images with less noisy artifacts. In addition to preserving content and style, our method supports substantially higher guidance scale values, as shown in Fig. 5. This robustness is enabled by the proposed scheduling strategy, which allows for more effective guidance, even in challenging cases, while reducing sensitivity to the hyperparameter gs. We present further proof of this property in Sec. 4.3.

Finally, while the existing diffusion HTG models have the ability to generalize to datasets like IAM [27], in harder cases, such as GNHK [23], an unseen style might be hard to reproduce with faithful content. In Fig. 4, we observe that using DOG, the generation process manages to approach harder cases of unseen styles during training, while preserving the content.

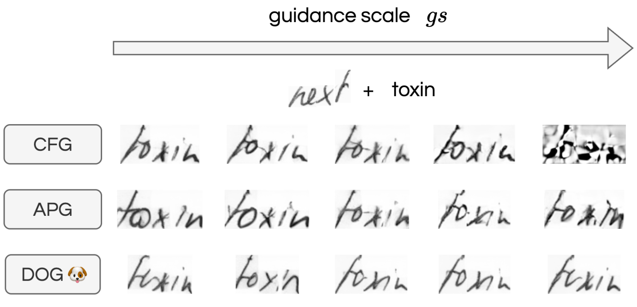

Figure 5. Comparison between different guidance strategies as guidance scale gs increases. We showcase gs values of 2, 5, 10, 20, and 30 (left to right column).

Variability. Diffusion-based HTG models tend to memorize the training set and often struggle to generate variations of the same words written by a specific writer, especially when only a single example of this instance exists. This means that the model has not learned to produce different instances of a given query word in a specific writer's style. This issue becomes evident when comparing DOG's variation capabilities with CFG and APG in Fig. 6. DOG consistently generates diverse instances of the same word across multiple runs, whereas CFG and APG exhibit only limited visual variation, often producing nearly identical outputs, even from random noise initialization, where different sampling output is expected.

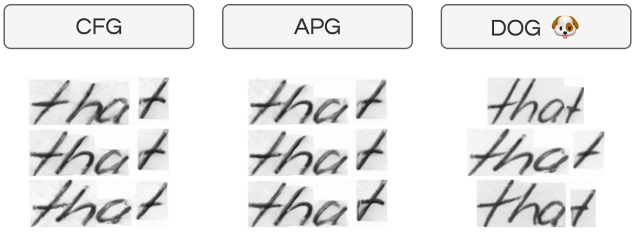

Figure 6. Comparison between CFG, APG, and the proposed DOG in terms of variance obtained from multiple samplings with different noise initializations.

4.3. Quantitative Results

We evaluate the generated data by including Handwriting Text Recognition (HTR) experiments using the system proposed in [40], in similar ways as [34]. We mainly focus on an OOV-generated set of seen writer styles from IAM to quantify the ability to produce variable styles and preserve text content. Moreover, we compute commonly used scores such as HWD [36] and FID [12] on generated IAM test sets using the guidance strategies. For this section, we proceed with DiffusionPen as a backbone, as it performs faster generation than One-DM due to its latent space operation.

Table 1. $\mathrm{HTG_{OOV}}$ (CER ↓) for DiffusionPen with various guidance types and strengths. The lower the better. Bold refers to the best score per gs value and underlined per guidance strategy.

	Guidance			
g_s	none	CFG	APG	DOG (ours)
–	22.8	–	–	–
2	–	20.1	20.8	**18.3**
10	–	19.9	20.8	**18.1**
20	–	39.4	24.0	**18.4**
30	–	90.8	29.4	**18.5**

OOV Words. We showcase how DOG improves content accuracy by creating a small set of ∼ 18.6K OOV words generated with random seen writer styles from IAM. We quantify the effect of CFG, APG, and DOG by presenting $\mathrm{HTR_{OOV}}$ for the best guidance scale gs for each strategy in Tab. 1. $\mathrm{HTG_{OOV}}$ refers to using an HTR trained on the real IAM corpus and testing the recognition of the OOV set, which is a Character Error Rate (CER), hence evaluating the readability. We can see that DOG achieves better performance in terms of content accuracy with the lowest CER in the $\mathrm{HTG_{OOV}}$.

Guidance Scale Range. Our proposed DOG enables larger values of the base guidance scale gs as shown in Fig. 5. To quantify that effect and obtain the best scores

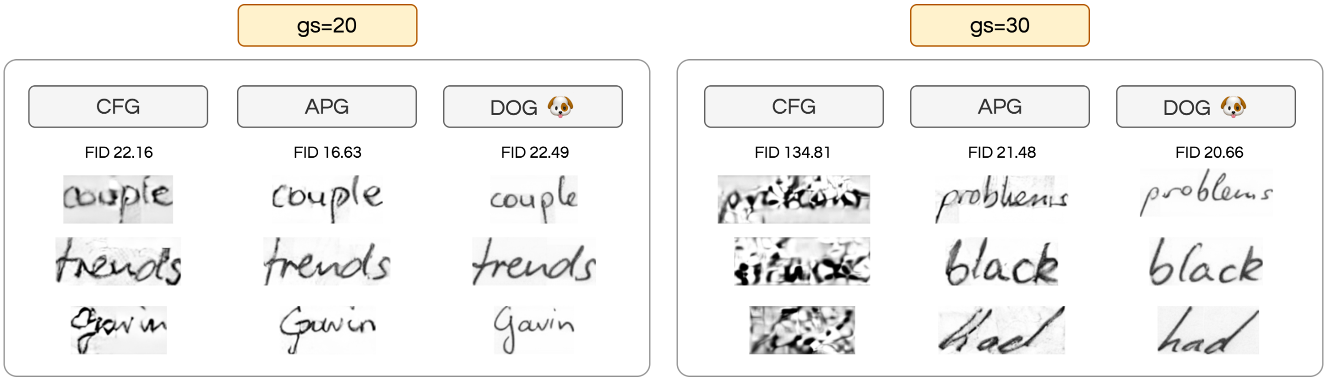

Figure 7. Qualitative results of CFG, APG, and our proposed DOG for $gs = 20$ (left) and $gs = 30$ (right) in correlation with FID score.

Table 2. HTR performance on the real IAM validation and test sets when incorporating large-scale generated data with and without DOG. For DOG, we use $gs = 20$.

Training	CER ↓	
	validation	test
Real IAM	3.58	4.92
DiffPen [33]	2.43	4.17
+DOG (ours)	**2.27**	**3.99**

for every guidance strategy, we present HTG$_{\text{OOV}}$ in various guidance scales in Tab. 1, spanning gs values from 2 to 30. It is clear that our method gives more robust results as gs increases, while CFG and APG output much more noise and artifacts, harming the readability.

Handwritten Text Recognition (HTR). To ensure that the generation using our proposed guidance does not simplify the text, resulting in the high recognition performance obtained in the HTG$_{\text{OOV}}$ results, we incorporate generated data into the training process of an HTR system [40]. To this end, we generate a large corpus of ∼376K samples from IAM training writer styles using DiffusionPen with and without the DOG guidance. We incorporate the large generated sets in the training process of the HTR along with the real data, aiming to boost the performance, and present the CER on the real validation and test sets of the IAM database. The results are presented in Tab. 2. In both generated cases, the performance improves; however, the generated data produced using our guidance strategy improves the recognition even more. We should note that to avoid harming the HTR learning with too noisy data, we filter the generated data as proposed in [34]. This means that with our proposed DOG, in one generation pass, we are able to keep more data that is useful to train an HTR system.

Table 3. HWD, FID, and CER scores of the IAM test set generated by DiffusionPen using no strategy or CFG, APG, and our proposed DOG. For all scores, the lower the better. Bold refers to the best result per score, and underlined refers to the best value of each score per guidance strategy.

Guidance	g_s	HWD↓	FID↓	CER↓
None	–	1.57	**12.05**	10.2
CFG	2	1.56	12.39	8.7
	10	**1.55**	13.95	9.0
	20	1.61	22.16	22.3
	30	2.40	134.81	82.1
APG	2	1.57	12.79	8.9
	10	**1.55**	13.57	9.7
	20	1.56	16.63	11.6
	30	1.62	21.48	16.4
DOG (ours)	2	1.65	22.16	7.8
	10	1.65	22.16	**7.2**
	20	1.65	22.49	7.5
	30	1.64	20.66	7.5

Generation Scores. We present HWD [36], FID [12], and CER scores, comparing the IAM test set generation quality of DiffusionPen, with and without the guidance strategies, in Tab. 3. While our method clearly improves readability according to the CER results, HWD is slightly harmed, which is expected as the style output might drift due to the guidance. FID is increasing by a value of 10 across all gs values for our proposed DOG compared to CFG and APG, which have increased scores for $gs > 20$. However, if we look qualitatively at the outputs, it is clear that FID is not appropriate to evaluate the quality of the generation as shown in Fig. 7. There, we can see that for lower values of FID for $gs = 20$, both CFG and APG output noisy samples, while DOG, which has the worst (highest) FID score, presents the most stable samples. In the case of $gs = 30$, we can observe that CFG outputs completely erroneous samples, which justifies the exploded FID score it obtains, while APG has a worse score, still, though, lower than the $gs = 20$ of DOG that produces the most "clean" results. This confirms that FID is not an appropriate score to measure HTG quality.

4.4. Ablation

We perform ablation studies on the key elements of our proposed method: the orthogonal projection and the scheduling strategy. The importance of both components is evident as shown in Fig. 8. Without the orthogonal projection, the generation collapses under the influence of high-magnitude noise, which overwhelms the meaningful style and content signals. Similarly, removing the scheduling component leads to noticeably noisier outputs. In this case, we have used a value of λ_s and λ_t equal to 100 to better demonstrate the effect of the examined components. This issue becomes increasingly severe at higher gs values, even when orthogonal projection is applied.

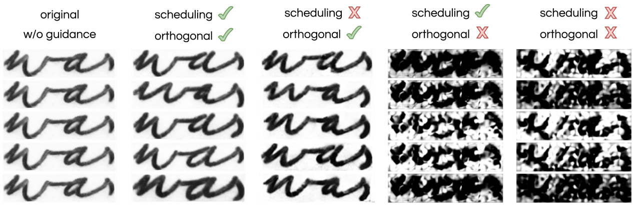

Figure 8. Ablation on the use of our proposed guidance scheduling and orthogonal component. The second column represents the full DOG method. The samples are generated using $gs = 2$.

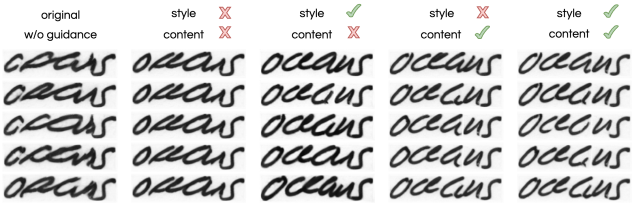

Figure 9. Ablation on the possible negative conditions (style and content) to obtain the negative direction using $gs = 25$. Second column corresponds to using the unconditional prediction as the negative prompt, while the third and fourth column keep the content and the style, respectively, the same as the positive prompt.

Moreover, we examine the effect of using different negative conditions, namely style and content, when constructing the negative direction in Eq. (5), as shown in Fig. 9. We compare several configurations: using only the negative style or negative content while keeping the other fixed, and replacing the negative condition entirely with the unconditional prediction (second column), keeping the projection and scheduling components constant across all cases. The results lead to two key observations. First, while using the unconditional prediction improves over the original system without guidance, it provides less variability and clarity compared to using negative prompts, even when only one condition is altered. Second, applying a negative condition to either the style or content alone appears to be effective, supporting the flexibility of the proposed method depending on the target needs of the generation.

Finally, we experiment with early, middle, and late timestep values of the scheduling peak u_T presented in the triangular scheduling of our proposed DOG. As we can see in Fig. 10, the early peak value choice of 200 generates noisy outputs. A possible reason is that in later (cleaner) steps, the sample has already been formed, hence a high guidance at that point might create erosions. Looking at the results, the earlier (noisier) timestep peak we choose for the scheduling, the more stable results we have, as the guidance is provided in a step where formation choices are still made, and hence, our choice of $u_T = 700$.

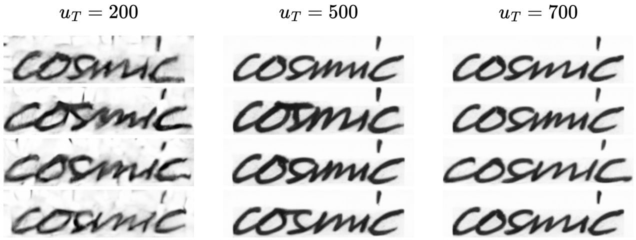

Figure 10. Ablation on the scheduling peak timestep u_T showing an early (left), middle (middle), and later (right) timestep. The samples are generated using $gs = 30$.

5. Conclusion and Future Work

We introduced a simple yet effective sampling strategy for diffusion-based HTG systems named DOG. Instead of the unconditional prediction used in traditional CFG, DOG employs a *structured negative prompt* and derives an orthogonal update that is disentangled from the positive direction, providing a dual-component condition. We couple the guidance with a triangular, time-aware schedule and a scaling component for further robustness. DOG can be plugged into any off-the-shelf diffusion-based HTG model without further training and improves the generation with more faithful content while preserving style variability. We compare DOG with existing CFG and APG sampling strategies, showcasing superior and more stable generation results. Through a combination of qualitative and quantitative results, we also show how problematic the use of FID is in domains outside of natural images, like HTG. While our work proves robust in batch sampling, there is still room for improvement, as each sample may benefit from its own ideal guidance scale. In general, DOG can serve as a preliminary exploration for future research on guiding the generation of handwritten words with more robust results, enhancing content accuracy while preserving the style.

Acknowledgment

The computations and data handling were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre at Linköping University.

References

[1] Eloi Alonso, Bastien Moysset, and Ronaldo Messina. Adversarial Generation of Handwritten Text Images Conditioned on Sequences. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 481–486. IEEE, 2019. 2

[2] Martin Arjovsky and Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks. *arXiv preprint arXiv:1701.04862*, 2017. 2

[3] Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the Negative Prompt Algorithm: Transform 2d Diffusion into 3D, alleviate Janus problem and Beyond. *arXiv preprint arXiv:2304.04968*, 2023. 2

[4] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Mubarak Shah. Handwriting Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1086–1094, 2021. 2

[5] Kai Brandenbusch. Semi-Supervised Adaptation of Diffusion Models for Handwritten Text Generation. *arXiv preprint arXiv:2412.15853*, 2024. 2

[6] Silvia Cascianelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Boosting Modern and Historical Handwritten Text Recognition with Deformable Convolutions. *International Journal on Document Analysis and Recognition (IJDAR)*, 25(3):207–217, 2022. 1

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 2

[8] Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-DM: One-Shot Diffusion Mimicker for Handwritten Text Generation. In *European Conference on Computer Vision*, pages 410–427. Springer, 2025. 2, 5, 6

[9] Brian L. Davis, Chris Tensmeyer, Brian L. Price, Curtis Wigington, B. Morse, and R. Jain. Text and Style Conditioned GAN for the Generation of Offline-Handwriting Lines. In *Proceedings of the* 31*st British Machine Vision Conference (BMVC)*, 2020. 2

[10] Moises Diaz, Andrea Mendoza-García, Miguel A Ferrer, and Robert Sabourin. A survey of handwriting synthesis from 2019 to 2024: A comprehensive review. *Pattern Recognition*, page 111357, 2025. 1

[11] Haisong Ding, Bozhi Luan, Dongnan Gui, Kai Chen, and Qiang Huo. Improving Handwritten OCR with Training Samples Generated by Glyph Conditional Denoising Diffusion Probabilistic Model. In *International Conference on Document Analysis and Recognition*, pages 20–37. Springer, 2023. 2

[12] DC Dowson and BV666017 Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. 6, 7

[13] Farzan Farnia and Asuman Ozdaglar. Do GANs always have Nash equilibria? In *International Conference on Machine Learning*, pages 3029–3039. PMLR, 2020. 2

[14] Ji Gan, Weiqiang Wang, Jiaxu Leng, and Xinbo Gao. HiGAN+: Handwriting Imitation GAN with Disentangled Representations. *ACM Transactions on Graphics (TOG)*, 42(1): 1–17, 2022. 2

[15] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 2

[16] Aniket Gurav, Narayanan C Krishnan, and Sukalpa Chanda. Word-Diffusion: Diffusion-Based Handwritten Text Word Image Generation. In *International Conference on Pattern Recognition*, pages 53–72. Springer, 2025. 2

[17] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 5, 6

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[19] Sana Khamekhem Jemni, Sourour Ammar, Mohamed Ali Souibgui, Yousri Kessentini, and Abbas Cheddad. ST-KeyS: Self-supervised Transformer for Keyword Spotting in Historical Handwritten Documents. *Pattern Recognition*, page 112036, 2025. 1

[20] Lei Kang, Pau Riba, Yaxing Wang, Marçal Rusinol, Alicia Fornés, and Mauricio Villegas. GANwriting: Content-Conditioned Generation of Styled Handwritten Word Images. In *European Conference on Computer Vision*, pages 273–289. Springer, 2020. 2

[21] Lei Kang, Pau Riba, Marcal Rusinol, Alicia Fornés, and Mauricio Villegas. Content and style aware generation of text-line images for handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[22] Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. Pay attention to what you read: non-recurrent handwritten text-line recognition. *PR*, 129:108766, 2022. 1

[23] Alex WC Lee, Jonathan Chung, and Marco Lee. GNHK: A Dataset for English Handwriting in the Wild. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 399–412. Springer, 2021. 5, 6

[24] Zhouchen Lin and Liang Wan. Style-preserving english handwriting synthesis. *Pattern Recognition*, 40(7):2097–2109, 2007. 2

[25] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional Visual Generation with Composable Diffusion Models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 2

[26] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. 3

[27] U-V Marti and Horst Bunke. The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5: 39–46, 2002. 5, 6

[28] Alexander Mattick, Martin Mayr, Mathias Seuret, Andreas Maier, and Vincent Christlein. Smartpatch: Improving Handwritten Word Imitation with Patch Discriminators. In *International Conference on Document Analysis and Recognition*, pages 268–283. Springer, 2021. 2

[29] Martin Mayr, Marcel Dreier, Florian Kordon, Mathias Seuret, Jochen Zöllner, Fei Wu, Andreas Maier, and Vincent Christlein. Zero-Shot Paragraph-level Handwriting Imitation with Latent Diffusion Models. *International Journal of Computer Vision*, pages 1–22, 2025. 2

[30] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, pages 3481–3490. PMLR, 2018. 2

[31] Konstantina Nikolaidou, Mathias Seuret, Hamam Mokayed, and Marcus Liwicki. A Survey of Historical Document Image Datasets. *International Journal on Document Analysis and Recognition (IJDAR)*, 25(4):305–338, 2022. 1

[32] Konstantina Nikolaidou, George Retsinas, Vincent Christlein, Mathias Seuret, Giorgos Sfikas, Elisa Barney Smith, Hamam Mokayed, and Marcus Liwicki. WordStylist: Styled Verbatim Handwritten Text Generation with Latent Diffusion Models. In *International Conference on Document Analysis and Recognition*, pages 384–401. Springer, 2023. 2

[33] Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. DiffusionPen: Towards Controlling the Style of Handwritten Text Generation. In *European Conference on Computer Vision*, pages 417–434. Springer, 2024. 2, 5, 6, 7

[34] Konstantina Nikolaidou, George Retsinas, Giorgos Sfikas, and Marcus Liwicki. Rethinking HTG Evaluation: Bridging Generation and Recognition. In *European Conference on Computer Vision*, pages 179–195. Springer, 2024. 6, 7

[35] Vittorio Pippi, Silvia Cascianelli, and Rita Cucchiara. Handwritten Text Generation from Visual Archetypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22458–22467, 2023. 2

[36] Vittorio Pippi, Fabio Quattrini, Silvia Cascianelli, and Rita Cucchiara. HWD: A Novel Evaluation Score for Styled Handwritten Text Generation. In *BMVC*, 2023. 6, 7

[37] George Retsinas, Georgios Louloudis, Nikolaos Stamatopoulos, and Basilis Gatos. Keyword Spotting in Handwritten Documents Using Projections of Oriented Gradients. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 411–416. IEEE, 2016. 1

[38] George Retsinas, Georgios Louloudis, Nikolaos Stamatopoulos, and Basilis Gatos. Efficient Learning-Free Keyword Spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1587–1600, 2018. 1

[39] George Retsinas, Giorgos Sfikas, Christophoros Nikou, and Petros Maragos. From Seq2Seq Recognition to Handwritten Word Embeddings. In *BMVC*, 2021. 1

[40] George Retsinas, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. Best practices for a handwritten text recognition system. In *International Workshop on Document Analysis Systems*, pages 247–259. Springer, 2022. 6, 7

[41] George Retsinas, Konstantina Nikolaidou, and Giorgos Sfikas. Enhancing CRNN HTR Architectures with Transformer Blocks. In *International Conference on Document Analysis and Recognition*, pages 425–440. Springer, 2024. 1

[42] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating Oversaturation and Artifacts of High Guidance Scales in Diffusion Models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 5, 6

[43] Sharon Fogel and Hadar Averbuch-Elor and Sarel Cohen and Shai Mazor and Roee Litman. ScrabbleGAN: Semi-Supervised Varying Length Handwritten Text Generation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4323–4332, 2020. 2

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2, 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 3

[46] Achint Oommen Thomas, Amalia Rusu, and Venu Govindaraju. Synthetic handwritten captchas. *Pattern Recognition*, 42(12):3365–3373, 2009. 2

[47] Bram Vanherle, Vittorio Pippi, Silvia Cascianelli, Nick Michiels, Frank Van Reeth, and Rita Cucchiara. VATr++: Choose Your Words Wisely for Handwritten Text Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):934–948, 2025. 2

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2

[49] Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, and Cong Yao. Conditional Text Image Generation with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14235–14245, 2023. 2