This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# DCText: Scheduled Attention Masking for Visual Text Generation via Divide-and-Conquer Strategy

Jaewoo Song[1,2] Jooyoung Choi[1] Kanghyun Baek[3] Sangyub Lee[3]
Daemin Park[1] Sungroh Yoon[1,3,4,*]
[1]Department of Electrical and Computer Engineering, Seoul National University
[2]Global Technology Research, Samsung Electronics
[3]IPAI, [4]AIIS, ASRI, INMC, ISRC, Seoul National University
{woo.song, jy_choi, qor6271, nickyub, eoalsqkr12, sryoon}@snu.ac.kr




Figure 1. Given a global prompt and target regions (red boxes), DCText decomposes the target text (highlighted in red) and assigns it to regions, enabling accurate and coherent visual text generation, which the base Flux [13] model struggles to handle reliably. The prompts below each image are abbreviated from the original global prompts (full prompts in Appendix E.3).

## Abstract

*Despite recent text-to-image models achieving high-fidelity text rendering, they still struggle with long or multiple texts due to diluted global attention. We propose DCText, a training-free visual text generation method that adopts a divide-and-conquer strategy, leveraging the reliable short-text generation of Multi-Modal Diffusion Transformers. Our method first decomposes a prompt by extracting and dividing the target text, then assigns each to a designated region. To accurately render each segment within their regions while preserving overall image coherence, we introduce two attention masks—Text-Focus and Context-Expansion—applied sequentially during denoising. Additionally, Localized Noise Initialization further improves text accuracy and region alignment without increasing computational cost. Extensive experiments on single- and multi-sentence benchmarks show that DCText achieves the best text accuracy without compromising image quality while also delivering the lowest generation latency.*

*Correspondence to: Sungroh Yoon (sryoon@snu.ac.kr)

## 1. Introduction

Visual text generation has consistently been a challenging task within the text-to-image (T2I) [9, 13, 19, 23, 25] domain. Despite the impressive image quality of early T2I models, they still struggle to generate accurate and natural-looking visual text. While many methods [4, 14, 15, 26, 33] attempt to overcome this limitation by adding auxiliary modules or fine-tuning, they are often constrained by the capacity of the base model, resulting in poor quality.

Recently, Multi-Modal Diffusion Transformer (MM-DiT) models such as Stable Diffusion 3 series [9] and Flux [13], equipped with powerful text encoders [21], have significantly improved text rendering capabilities inherently. Building on this, several approaches [7, 11] directly leverage them as backbones in a training-free manner to improve visual text generation. However, these methods still introduce inference-time computational overhead and lack sufficient text accuracy. In particular, the latter issue primarily stems from the global attention mechanism: as the amount of text to be rendered increases, the full attention structure in MM-DiT—where all text and image tokens mutually attend to one another—dilutes attention to individ-

ual text tokens, thereby causing omissions, typos, and misplacements that degrade the overall text fidelity.

In this paper, we introduce DCText, which addresses the limitations of the full attention by applying scheduled attention masking to ensure accurate text generation at specified positions. Our method builds on the observation that Flux can reliably produce relatively short, single pieces of text with high fidelity. Leveraging this, we adopt a divide-and-conquer strategy: a global prompt containing multiple or lengthy rendering texts is decomposed into partial prompts based on its text content. Each partial prompt is then responsible for generating its assigned text within a designated region, while the overall image remains consistent with the global prompt. This divide-and-conquer generation is implicitly carried out within a single denoising process, enabled solely by attention masking. However, this approach requires careful mask design, as the split text segments belong to both partial and global prompts, increasing the risk of duplicated text generations, and the image regions generated by these separated prompts often struggle to blend seamlessly with the surrounding background.

To address these challenges, we design two attention masks. *Text-Focus Attention Mask*: concentrates attention into each designated region, ensuring that the generated text is accurate and confined to the target region without duplication. *Context-Expansion Attention Mask*: allows bidirectional interactions between each region and its surrounding background, enabling smooth and coherent transitions across boundaries. By sequentially applying these masks during the denoising process, our method balances accurate text fidelity and visual coherence in the final output. In addition, we propose a *Localized Noise Initialization* approach, which refines the initial noise to provides spatial guidance for the text to be rendered in each region. This approach is not only computationally efficient but also improves region–text alignment and enhances the text accuracy.

We evaluate DCText on diverse datasets containing both single and multiple rendering sentences. Our method achieves the highest text accuracy among other tuning-free approaches, while also delivering the best image quality. Remarkably, these results are obtained with the fewest denoising steps and the lowest latency, and DCText further provides flexible controllability over text placement.

## 2. Related Work

### 2.1. Visual Text Generation

With advances in text-to-image models, various methods have been proposed to generate images containing text. Some approaches train character-level text encoders [14–16, 29, 30, 35], enabling models to recognize and generate glyph structures at the character level. Others incorporate external modules into T2I models [4, 5, 17, 26–28, 33, 34] to handle glyph-level information, which not only improves text rendering accuracy but also enables spatial control over text layout through masks or region-based guidance. However, these methods are built upon U-Net–based Stable Diffusion models [24] and often require additional training, exhibiting limitations in both visual fidelity and text accuracy.

To overcome these challenges, recent research has focused on training-free approaches that leverage the pretrained capabilities of Flux to enhance text generation at inference time. AMO Sampler [11] improves text rendering accuracy by introducing a stochastic sampler along with a mask based on cross attention, but lacks explicit control over text positioning. TextCrafter [7] addresses the multi-text rendering task by separately denoising layout regions where text is generated and re-weighting attention maps to mitigate text blurring. But it often struggles to achieve seamless integration between text regions and the background, resulting in unnatural visual transitions. Furthermore, both methods still suffer from limited text accuracy and inference-time computational overhead.

### 2.2. Attention Control

Attention maps have been widely explored for controlling generation in diffusion models. Prompt-to-Prompt [10] reveals that cross-attention maps reflect spatial alignment between prompt and image, which can be exploited for prompt-based image editing by replacing or re-weighting these maps. Additional techniques compute loss signals from attention maps and apply latent updates to better align outputs with textual prompts [1, 2, 22]. Other approaches extend this idea to incorporate layout conditioning, enabling models to reflect both prompt semantics and spatial constraints [6, 32]. However, these methods rely solely on cross-attention mechanism, which limits their flexibility in capturing fine-grained spatial relationships. In addition, the latent update procedures introduce extra memory and computational overhead during inference.

More recently, MM-DiT-based models have adopted joint attention mechanisms, where text and image tokens are concatenated and jointly attended. Some studies apply spatial masks directly to joint attention maps to enforce layout constraints without latent updates. For instance, Regional-Prompting [3] enables compositional generation by performing both manipulated and unmanipulated attention evaluations, which doubles number of function evaluations (NFE) per denoising step. In contrast, DreamRenderer [36] replicates image tokens for each instance during attention, leading to substantial overhead within each step, instead of doubling NFE. Above all, these approaches mainly focus on generic objects while overlooking text, which is more difficult to generate due to strict requirements on glyph accuracy and character sequencing—especially when rendering large amounts of text.

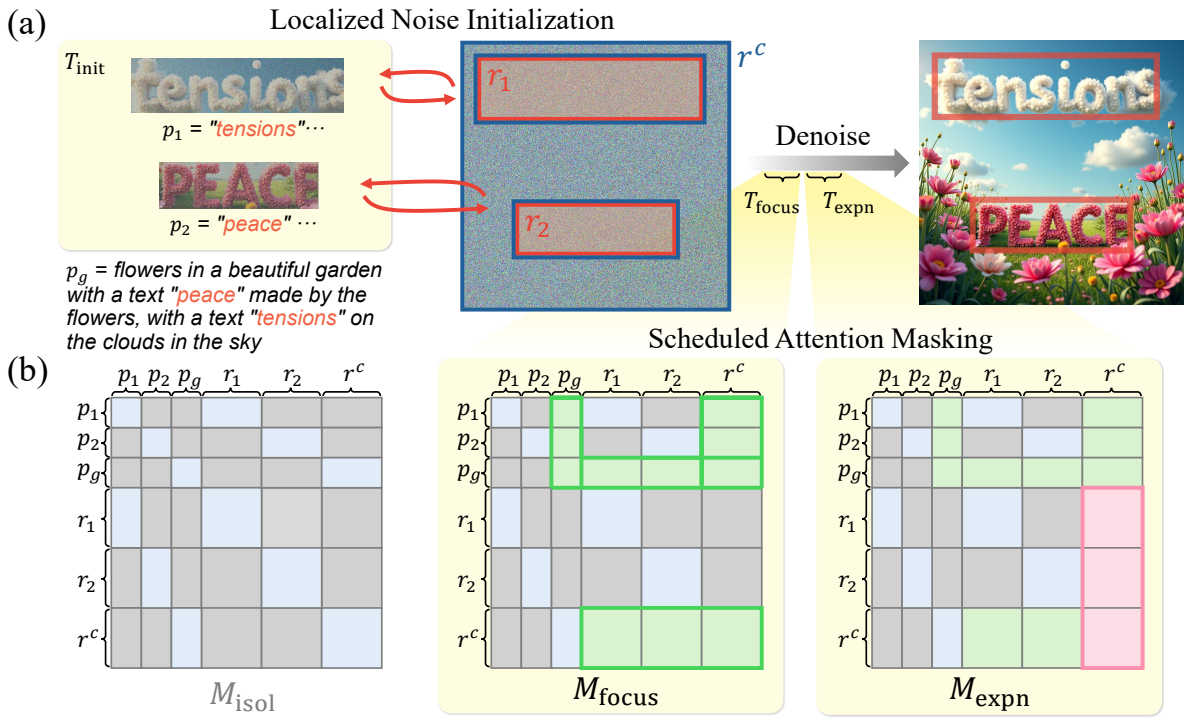



Figure 2. **Overview of DCText.** (a) Textual regions $(r_1, r_2)$ are extracted from the initial noise and independently denoised with their textual prompts $(p_1, p_2)$ for $T_{\text{init}}$ steps. The refined patches are then blended back to form the localized initial noise. (b) It is then sequentially denoised using two attention masks, $M_{\text{focus}}$ and $M_{\text{expn}}$. Both masks build on $M_{\text{isol}}$, which restricts attention to text and image tokens within each region (blue). $M_{\text{focus}}$ introduces four attention areas (green) to allow controlled information flow between textual regions and the background, supporting accurate text rendering. $M_{\text{expn}}$ further enables region-to-background attention (pink) to promote smooth transitions. In all three masks, colored areas denote allowed attention (1), while gray areas indicate masked attention (0).

## 3. Method

Our goal is to accurately and naturally generate texts from a global prompt, which may contain long or multiple sentences. To achieve this, we leverage bounding box conditions: given a set of regions requiring visual text segments (textual regions) and a global prompt, we extract the target texts embedded in the prompt and divide them to match the number of regions, generating each within its assigned region. Building on the attention mechanism of the Multi-Modal Diffusion Transformer, we control the information flow between textual and non-textual content (both across prompts and regions) using carefully designed attention masks. This approach enables region-centric text generation while ensuring smooth transitions in the background, keeping the overall image coherent with the global prompt. Our method, DCText, improves text accuracy by applying these attention masks at inference time, with minimal overhead compared to the base text-to-image model. It requires no training or fine-tuning and avoids heavy computations such as gradient calculations during inference [1, 22].

In the following sections, we first review the attention mechanism in MM-DiT (Sec. 3.1), then describe how this attention is regulated by two novel attention masks and their design formulation (Sec. 3.2), next present the strategy for obtaining the initial noise (Sec. 3.3), and finally outline the overall DCText pipeline (Sec. 3.4).

### 3.1. Preliminary: Attention in MM-DiT

Recent T2I models [8, 13] employ MM-DiT blocks, which operate on a unified token sequence formed by concatenating text and image tokens. Each block computes attention over this combined sequence as follows:

$$\text{Attn}(Q, K, V) = AV, \quad A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \quad (1)$$

where $Q = \text{concat}(Q_{\text{text}}, Q_{\text{img}})$, with $Q_{\text{text}}$ and $Q_{\text{img}}$ representing the queries derived from the text prompt tokens and image patch tokens, respectively; key $(K)$ and value $(V)$ are constructed in the same manner as $Q$.

During attention computation, it is possible to regulate the information flow between tokens by modifying or masking the attention map $A$. In particular, applying an attention mask $M$ enables selective interaction by suppressing irrele-

vant connections while preserving meaningful ones:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \log(M)\right). \tag{2}$$

Here, the binary values in $M$ (1 for allowed connections, 0 for masked) are transformed via the $\log$ function into 0 and $-\infty$, respectively, to be added before the softmax operation.

### 3.2. Scheduled Attention Masking

We first decompose a global prompt, which contains multiple or lengthy target texts to be rendered, into a set of partial prompts (textual prompts) based on the number of textual regions $n$. For example, given a global prompt such as *'flowers in a beautiful garden with a text "peace" made by the flowers, with a text "tensions" on the clouds in the sky'* and two regions, we construct textual prompts by separating *peace* and *tensions*, and assign each to its region.

These $n$ textual prompts are encoded in the same manner as the global prompt, resulting in $\{p_i\}_{i=1}^{n}$. We extend the attention inputs $Q$, $K$, and $V$ in Eq. (1) to incorporate both the textual prompts and the global prompt as follows:

$$Q_{\text{text}} = \text{concat}(\{Q_{p_i}\}_{i=1}^{n}, Q_{p_g}), \tag{3}$$

where $Q_{p_i}$ and $Q_{p_g}$ are queries from the $i$-th textual prompt and the global prompt, respectively. In parallel, we organize the image-side query $Q_{\text{img}}$ not by extending it, but by decomposing it based on the textual regions. Specifically, image patches are divided into region-specific parts $\{r_i\}_{i=1}^{n}$ based on the given bounding boxes, and a background region $r^c = \left(\bigcup_{i=1}^{n} r_i\right)^c$, *i.e.* the complement of all regions:

$$Q_{\text{img}} = \text{concat}\left(\{Q_{r_i}\}_{i=1}^{n},\ Q_{r^c}\right). \tag{4}$$

$K$ and $V$ are constructed analogously to $Q$.

We thus compute attention over the token sequence consisting of prompts $\{p_i\}_{i=1}^{n+1}$ and regions $\{r_i\}_{i=1}^{n+1}$, where $p_{n+1} = p_g$ and $r_{n+1} = r^c$. Note that the last prompt $p_g$ serves as a global description that encompasses the textual prompts, while the last region $r^c$ corresponds solely to the background region, excluding all textual regions.

In this setup, we propose two types of attention masks to guide the denoising process for accurate and coherent text generation. The text-focus attention mask $M_{\text{focus}}$ biases the overall attention toward textual regions, enabling accurate text rendering at the correct regions while allowing the background to remain contextually natural. The context-expansion attention mask $M_{\text{expn}}$, on the other hand, further allows each textual region to attend to the background region, promoting smoother transitions and coherence at the boundaries between the region and the background.

To motivate the construction of these two masks, we first introduce the region-isolation attention mask $M_{\text{isol}}$, which enforces strict separation across all regions. While not used in our method, $M_{\text{isol}}$ serves as a conceptual foundation from which our proposed masks are derived. In the following, we detail the formulation of $M_{\text{isol}}$ and then construct $M_{\text{focus}}$ and $M_{\text{expn}}$ by gradually relaxing the strict constraints imposed by $M_{\text{isol}}$. These masks are illustrated in Fig. 2b.

**Algorithm 1** DCText Pipeline

**Input**:
Prompts $\{p_i\}_{i=1}^{n+1}$ (with $p_{n+1}$: global prompt), Regions $\{r_i\}_{i=1}^{n+1}$ (with $r_{n+1}$: background), Denoising steps $T$, $T_{\text{init}}$, $T_{\text{focus}}$, $T_{\text{expn}}$, Blending weight $\alpha$
**Output**: Final image $x_0$

Make attention masks $M_{\text{focus}}$, $M_{\text{expn}}$
Sample $z_T \sim \mathcal{N}(0, I)$
▷ Localized Noise Initialization
**for** $i = 1$ to $n$ **do**
  $z^{(i)} \leftarrow$ Extract($z_T$, $r_i$)
  **for** $t$ for $T_{\text{init}}$ steps **do**
    $z^{(i)} \leftarrow$ Denoise($z^{(i)}, p_i$)
  **end for**
  $z_T \leftarrow$ Reinsert($z_T$, $r_i$, $z^{(i)}, \alpha$)
**end for**
$z \leftarrow z_T$
▷ Text-Focus Denoising
**for** $t$ for $T_{\text{focus}}$ steps **do**
  $z \leftarrow$ DenoiseWithMask($z$, $\{p_i\}_{i=1}^{n+1}$, $M_{\text{focus}}$)
**end for**
▷ Context-Expansion Denoising
**for** $t$ for $T_{\text{expn}}$ steps **do**
  $z \leftarrow$ DenoiseWithMask($z$, $\{p_i\}_{i=1}^{n+1}$, $M_{\text{expn}}$)
**end for**
▷ Global Denoising
**for** $t$ for $T - (T_{\text{init}} + T_{\text{focus}} + T_{\text{expn}})$ steps **do**
  $z \leftarrow$ Denoise($z$, $p_{n+1}$)
**end for**
**return** Decode($z$)

**Region-Isolation Attention Mask.** Our divide-and-conquer approach begins by partitioning the task, enabling each textual prompt to render its target text independently within its assigned region. This is enabled by constructing an isolated attention mask that implicitly regulates inter-region interactions, ensuring that each region focuses solely on its target text. Specifically, we enforce that only the text and image tokens associated with a particular region (i.e., $\{p_i, r_i\}$) can attend to one another, both across and within modalities. Let $m_{p_i} \in \{0,1\}^{L_T}$ and $m_{r_i} \in \{0,1\}^{L_I}$ be binary masks that activate only the token indices of the $i$-th textual prompt and its corresponding region, respectively. Here, $L_T$ is the total number of text tokens across all $n{+}1$ prompts, and $L_I$ is the number of image patch tokens (i.e.,




Figure 3. **Qualitative comparison on single sentence.** Prompts, including the sentence to be rendered (highlighted in red), are shown below each column. All comparisons are generated from the same initial noise.

$h \times w$). By concatenating $m_{p_i}$ and $m_{r_i}$, we construct a joint mask $m_i \in \{0,1\}^{L_T+L_I}$ that activates all text and image tokens associated with the $i$-th region, while masking out the rest. We then obtain the region-isolation attention mask $M_{\text{isol}}$ by summing the outer products of these masks:

$$M_{\text{isol}} = \sum_{i=1}^{n+1} m_i \cdot m_i^\top \in \{0,1\}^{(L_T+L_I)\times(L_T+L_I)}, \quad (5)$$

where $\cdot$ denotes an outer product, resulting in a square mask.

However, while applying this mask allows each region to render its target text correctly, the resulting image fails to follow the global prompt. This occurs because $M_{\text{isol}}$ eliminates all information flow between textual regions and the background, causing each region to be generated independently according to its own prompt. As a result, the final image may appear unnatural, as if multiple disjoint images were stitched together. Moreover, the background region, guided solely by the global prompt, may re-render text that is already rendered within the individual textual regions (left image in Fig. 6). In other words, this approach only divides the task but fails to conquer it.

**Text-Focus Attention Mask.** To address the limitations of $M_{\text{isol}}$, DCText expands the attendable regions in the mask (*i.e.* areas with value 1), allowing directional information flow between the textual regions, the background region, the textual prompts, and the global prompt.

For the background region, we allow it to attend to the textual regions so that the background can connect naturally to the text areas and avoid redundant text generation within the background. The subscript arrow $r^c \to \{r_i\}$ indicates the direction of attention from query to key:

$$M_{r^c \to \{r_i\}} = m_{r^c} \cdot \mathbf{1}_{L_I}^\top \in \{0,1\}^{L_I \times L_I}. \quad (6)$$

We also allow the background region to be attended by the textual prompts, enabling each prompt to incorporate surrounding visual context. This helps the prompts generate more natural and contextually appropriate descriptions, leading to more accurate text rendering:

$$M_{\{p_i\} \to r^c} = \mathbf{1}_{L_T} \cdot m_{r^c}^\top \in \{0,1\}^{L_T \times L_I}. \quad (7)$$

For the global prompt, we apply a similar strategy: enabling it to attend to the textual regions and receive attention

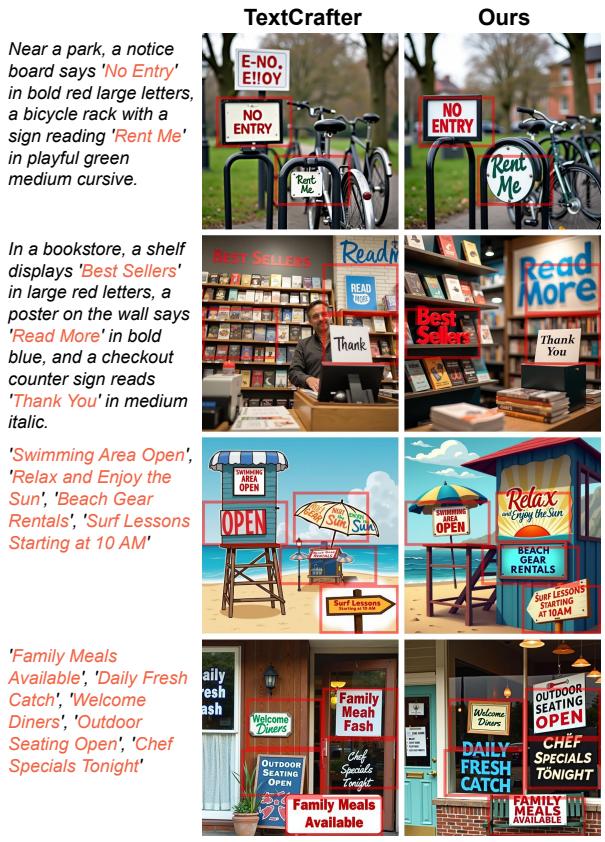



Figure 4. **Qualitative comparison on multiple sentences.** Comparison of generation results with varying numbers of sentences (2–5) in a single prompt. Sentences and corresponding regions are highlighted in red (only target texts are shown for the two prompts below; full prompts are in Appendix E.3). Our method consistently renders accurate text in the correct regions.

from the textual prompts. This promotes better coordination across the image and contributes to more natural and accurate text rendering (see Appendix B.1):

$$M_{p_g \to \{r_i\}} = m_{p_g} \cdot \mathbf{1}_{L_I}^\top \in \{0,1\}^{L_T \times L_I},$$
$$M_{\{p_i\} \to p_g} = \mathbf{1}_{L_T} \cdot m_{p_g}^\top \in \{0,1\}^{L_T \times L_T}. \tag{8}$$

Finally, the text-focus attention mask $M_{\text{focus}}$ is constructed by combining $M_{\text{isol}}$ with the four additional partial masks defined above:

$$M_{\text{focus}} = M_{\text{isol}} \vee \begin{bmatrix} M_{\{p_i\} \to p_g} & M_{\{p_i\} \to r^c} \vee M_{p_g \to \{r_i\}} \\ 0 & M_{r^c \to \{r_i\}} \end{bmatrix}. \tag{9}$$

**Context-Expansion Attention Mask.** While $M_{\text{focus}}$ enables accurate text generation within the textual regions and allows the background to naturally incorporate these regions, the textual regions themselves still attend solely to their own content. In other words, each region remains blind to the background, potentially leading to an interior that looks visually isolated from its surroundings (see Appendix Fig. S7). To address this, we allow textual regions to attend to the background after a few denoising steps with $M_{\text{focus}}$, once the target texts have been reasonably placed. This enables bidirectional information flow between the regions and the background:

$$M_{\{r_i\} \to r^c} = \mathbf{1}_{L_I} \cdot m_{r^c}^\top \in \{0,1\}^{L_I \times L_I},$$
$$M_{\text{expn}} = M_{\text{focus}} \vee \begin{bmatrix} 0 & 0 \\ 0 & M_{\{r_i\} \to r^c} \end{bmatrix}. \tag{10}$$

### 3.3. Localized Noise Initialization

We observe that generating text over an entire region-sized image is often easier than placing the same text within that region in the context of a full-sized image. Motivated by this, we propose a initialization strategy (Fig. 2a) that performs light denoising on region-specific noise patches, allowing each region to receive early guidance before global denoising begins. We first sample initial noise for the full image, $z_T \sim \mathcal{N}(0, I)$, where $z_T \in \mathbb{R}^{c \times h \times w}$. From $z_T$, we extract a set of localized noise patches $\{z_{T_i} \in \mathbb{R}^{c \times h_i \times w_i}\}_{i=1}^n$, each corresponding to a textual region $\{r_i\}_{i=1}^n$. Each patch $z_{T_i}$ is then independently denoised for a very small number of steps using its associated textual prompt, yielding a lightly refined latent $z_{T_i'}$, where $T' = T - T_{\text{init}}$. These refined patches are blended back into their original locations within $z_T$ using a weighting factor $\alpha$, resulting in the updated global latent $z_{T'}$. Specifically, for each region $r_i$:

$$z_{T'}[r_i] = \alpha \cdot z_{T_i} + (1 - \alpha) \cdot z_{T_i'}. \tag{11}$$

After this initialization for $T_{\text{init}}$ steps, we proceed with the remaining $T'$ denoising steps using the updated latent $z_{T'}$. This simple initialization approach improves region-text spatial alignment and text rendering accuracy while remaining computationally efficient, as the total region area $\sum_i h_i w_i$ is smaller than the full image area $hw$.

### 3.4. DCText Pipeline

DCText divides the denoising process into four sequential stages: (1) Localized Noise Initialization for $T_{\text{init}}$ steps, (2) Text-Focus denoising using the attention mask $M_{\text{focus}}$ for $T_{\text{focus}}$ steps, (3) Context-Expansion denoising using the attention mask $M_{\text{expn}}$ for $T_{\text{expn}}$ steps, and (4) standard denoising without attention masking control, using only the global prompt for the remaining steps. The full pseudocode of the DCText pipeline is provided in Algorithm 1.

## 4. Experiments

### 4.1. Experiment Setting

**Benchmark.** We evaluate our method on four datasets: ChineseDrawText [17], DrawTextCreative [14], TMDBE-

| Method | Acc. | NED | CLIP | Qual. | Aesth. | Steps | Latency |
|---|---|---|---|---|---|---|---|
| Flux | 0.266 | 0.579 | 0.343 | 4.657 | 3.851 | 24 | 13.89 |
| AMO | 0.274 | 0.569 | 0.342 | 4.658 | 3.863 | 28 | 25.93 |
| TC | 0.329 | 0.722 | **0.350** | 4.704 | **3.911** | 30 | 36.89 |
| **Ours** | **0.387** | **0.751** | 0.349 | **4.737** | 3.904 | 24 | 16.60 |

Table 1. **Quantitative comparison on single sentence.** Results are averaged over three single-sentence datasets (ChineseDrawText [17], DrawTextCreative [14], TMDBEval500 [4]). Steps and latency indicate denoising steps and total generation time per image (seconds, measured on an L40 GPU).

| $n$ | Method | Acc. | NED | CLIP | Qual. | Aesth. | Latency |
|---|---|---|---|---|---|---|---|
| 5 | Flux | 0.366 | 0.608 | 0.336 | 4.591 | 3.165 | 13.91 |
| | AMO | 0.432 | 0.660 | 0.335 | 4.652 | 3.218 | 26.02 |
| | TC | 0.685 | 0.859 | **0.349** | 4.659 | 3.506 | 40.53 |
| | **Ours** | **0.693** | **0.860** | 0.343 | **4.697** | **3.569** | 19.26 |
| 4 | Flux | 0.389 | 0.628 | 0.338 | 4.707 | 3.421 | 13.93 |
| | AMO | 0.488 | 0.690 | 0.337 | 4.709 | 3.518 | 25.97 |
| | TC | 0.693 | 0.867 | 0.352 | 4.665 | 3.488 | 39.60 |
| | **Ours** | **0.760** | **0.892** | **0.353** | **4.745** | **3.659** | 18.14 |
| 3 | Flux | 0.508 | 0.715 | 0.345 | 4.662 | 3.377 | 13.88 |
| | AMO | 0.575 | 0.750 | 0.343 | 4.689 | 3.489 | 25.99 |
| | TC | 0.722 | 0.880 | **0.351** | 4.659 | 3.571 | 39.01 |
| | **Ours** | **0.768** | **0.906** | **0.351** | **4.735** | **3.709** | 16.96 |
| 2 | Flux | 0.608 | 0.809 | 0.344 | 4.675 | 3.471 | 13.88 |
| | AMO | 0.642 | 0.826 | 0.341 | 4.654 | 3.584 | 25.91 |
| | TC | 0.758 | 0.919 | 0.346 | 4.697 | 3.616 | 38.22 |
| | **Ours** | **0.792** | **0.923** | **0.347** | **4.791** | **3.726** | 15.66 |

Table 2. **Quantitative comparison on multiple sentences.** Comparison results on the CVTG-Style [7] dataset across different numbers of sentences ($n$) to be rendered.

val500 [4], and CVTG-Style [7]. The first three datasets mainly contain single-sentence prompts, while CVTG-Style includes prompts with 2–5 rendering sentences. These datasets cover diverse text rendering scenarios, from artistic typography to real-world scenes and stylized text.

**Metric.** We evaluate text rendering accuracy using sentence-level accuracy (Acc.) and Normalized Edit Distance (NED) [26], with PP-OCRv4 [18] for single-sentence prompts and GPT-4o [12] for multi-sentence prompts to handle sentence-level recognition. Image quality is assessed using Q-Align [31] for Quality and Aesthetic Scores, and CLIP Score [20] for prompt–image alignment. Latency per image generation is measured for efficiency, and a user study is conducted to assess preference.

**Baselines.** We compare our method against tuning-free visual text generation methods based on Flux.1-dev [13]: Flux, the base model; AMO Sampler (AMO) [11], a sampler that adaptively corrects sampling trajectories using attention weights; TextCrafter (TC) [7], explicitly isolating

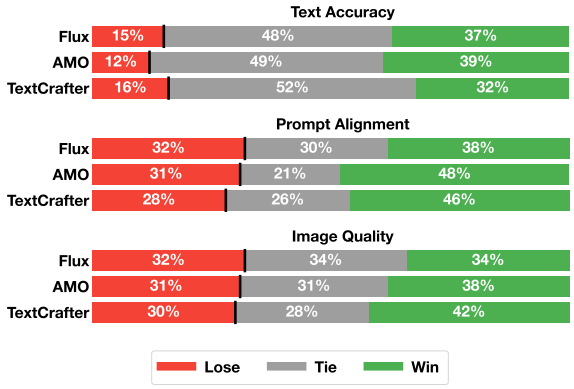



Figure 5. **Human Evaluation.** User preference on text accuracy, prompt–image alignment, and overall image quality. Green bars indicate cases where our method is preferred.

text regions for denoising and enhances attention for clearer rendering. All baselines use official implementations with default settings. Further comparisons with training-based methods and tuning-free methods based on Stable Diffusion 3.5 are provided in Appendix A.3 and C.2.

**Implementation Details.** We adopt Flux.1-dev [13] as our base model and generate all images at $1024 \times 1024$ resolution with 24 denoising steps. For each sentence in a prompt, we construct a textual prompt and region, using GPT-4o [12] for single-sentence cases and adopting CVTG-Style prompts while sharing regions from TextCrafter [7] for multi-sentence cases. We set $(T_{\text{init}}, T_{\text{focus}}, T_{\text{expn}})$ to (1,2,2) for single-sentence generation and (2,3,2) for multi-sentence generation; more details are in Appendix E.1.

## 4.2. Qualitative Results

Fig. 3 compares DCText with all tuning-free baselines on datasets containing single rendering sentences. In columns 1 and 4, baselines omit the sentence or produce only partial outputs, while our method generates the full text. In columns 2, 5, and 6, baselines generate incorrect or repeated text, whereas DCText produces the correct text as prompted. Notably, TextCrafter often produces text regions visually detached from the rest of the image (columns 3–5, *e.g.* with awkward white backgrounds), resulting in unnatural compositions. In contrast, DCText integrates text seamlessly into the scene, maintaining overall visual coherence.

Fig. 4 further compares our method with TextCrafter on multi-sentence generation. TextCrafter frequently fails to generate text within the designated areas or renders unintended text outside them (rows 1–3). As the number of regions increases (row 4), it often generates phrases in incorrect regions (*e.g. "OUTDOOR SEATING OPEN"*) or repeats the same phrase across different regions (*e.g. "Family*

*Meals Available"*). This stems from its global attention operating over the entire image, which introduces cross-region interference even when denoising is applied independently per region. In contrast, with Localized Noise Initialization and the Text-Focus Attention Mask, our method enforces region-focused attention control, accurately rendering the text in each correct region while preventing cross-region interference. At the same time, background region remains text-free and blend naturally with the surrounding scene. Additional samples are provided in Appendix A.1.

### 4.3. Quantitative Results

Tab. 1 presents the results on three single-sentence datasets. In terms of text accuracy, DCText achieves 45% and 30% improvements over Flux in sentence-level accuracy and NED, respectively, outperforming all baselines. Notably, our method maintains high aesthetic and quality scores, showing that these improvements do not compromise generative priors of the base model. The high CLIP score further suggests that the model continues to follow the global prompt well, even when conditioned on decomposed multiple prompts. Remarkably, these gains are achieved with the fewest denoising steps and the lowest latency, requiring only about 20% additional latency compared to Flux. Latency is measured over the entire pipeline, including Localized Noise Initialization (1 step) and 23 denoising steps.

Tab. 2 compares DCText against all baselines on the multi-sentence dataset across varying numbers of regions. As in Tab. 1, our method consistently achieves the highest text accuracy without compromising image quality across all region counts. These results demonstrate the effectiveness of our divide-and-conquer strategy, which decomposes text by region and employs scheduled attention masking to ensure accurate and coherent text generation within the overall image context. Although latency increases slightly as the number of regions grows, it remains lower than all baselines except Flux, underscoring efficiency of DCText. More comprehensive results are available in Appendix A.2.

To further assess user preferences, we conduct a human evaluation across three datasets, assessing text accuracy, alignment between the image and the global prompt, and overall image quality, following the same protocol as Tab. 1. A total of 30 evaluators perform pairwise comparisons between DCText and each baseline over 1,323 image pairs (see Appendix E.2). The results, presented in Fig. 5, show that our method is consistently favored across all criteria, with particularly significant improvements in text accuracy.

### 4.4. Ablation Study

Tab. 3 and Fig. 6 present ablation results, evaluating each component of DCText. To assess the effectiveness of our proposed masks, we compare the combination of $M_{\text{focus}}$ and $M_{\text{expn}}$ against using only $M_{\text{isol}}$ during the $T_{\text{focus}}$ + $T_{\text{expn}}$ steps. As shown in Fig. 6 (left) and the first row of Tab. 3, using $M_{\text{isol}}$ results in redundant text rendering and lower text accuracy (Acc., NED). This is due to the textual region and background independently attending to separate prompts—a textual prompt and a global prompt. In addition, the restricted information flow enforced by $M_{\text{isol}}$ creates a hard boundary around the region, preventing smooth transitions between the textual region and the background. In contrast, our two-stage masks ($M_{\text{focus}} + M_{\text{expn}}$) first focus attention on the textual region to prevent redundancy, and then gradually expand it to the background, enabling seamless transitions. This improves both text accuracy and overall image quality by achieving precise attention control between region and background. Finally, denoising from an latent obtained via Localized Noise Initialization (*i.e.* full DCText pipeline) provides region-specific guidance, achieving tighter alignment between rendered text and its region (right image in Fig. 6) and further boosting text accuracy (last row in Tab. 3). The sensitivity analysis of each component can be found in Appendix B.2-B.4.

| Method | Acc. | NED | CLIP | Qual. | Aesth. |
|---|---|---|---|---|---|
| $M_{\text{isol}}$ | 0.222 | 0.640 | 0.345 | 4.564 | 3.683 |
| $M_{\text{focus}}$ + $M_{\text{expn}}$ | 0.306 | 0.700 | 0.346 | 4.569 | 3.700 |
| $T_{\text{init}}$ + $M_{\text{focus}}$ + $M_{\text{expn}}$ | **0.387** | **0.751** | **0.349** | **4.737** | **3.904** |

Table 3. **Ablation study on DCText components.** $M_{\text{isol}}$ denotes denoising using only the isolation attention mask. $M_{\text{focus}}$ + $M_{\text{expn}}$ uses our attention masks during two-stage denoising. $T_{\text{init}}$ + $M_{\text{focus}}$ + $M_{\text{expn}}$ further incorporates Localized Noise Initialization, forming our full pipeline.

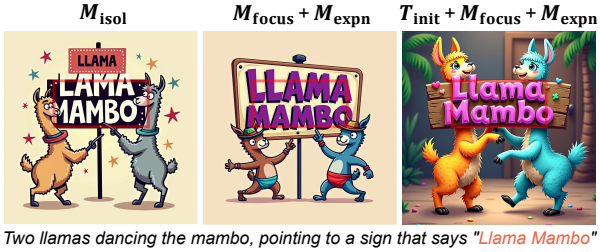



Figure 6. **Qualitative results of ablation study.** $M_{\text{isol}}$ causes redundant text and harsh boundary, while $M_{\text{focus}} + M_{\text{expn}}$ yields accurate text and smooth transition. Localized Noise Initialization further enhances text-region alignment

## 5. Conclusion

We present DCText, a training-free method that adopts a divide-and-conquer strategy for accurate and coherent visual text generation. By decomposing target texts and guiding generation with scheduled attention masking, DCText improves text accuracy with the lowest latency among all compared methods, without degrading image quality.

**Acknowledgement** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [No. 2022R1A3B1077720; 2022R1A5A7083908], Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2022-II220959; No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2025.

## References

[1] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 2, 3

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[3] Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024. 2

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023. 1, 2, 7

[5] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer, 2024. 2

[6] Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pages 432–448. Springer, 2024. 2

[7] Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025. 1, 2, 7

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[11] Xixi Hu, Keyang Xu, Bo Liu, Qiang Liu, and Hongliang Fei. Amo sampler: Enhancing text rendering with overshooting, 2025. 1, 2, 7

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[13] Black Forest Labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. 1, 3, 7

[14] Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. *arXiv preprint arXiv:2212.10562*, 2022. 1, 2, 6, 7

[15] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *European Conference on Computer Vision*, pages 361–377. Springer, 2024. 1

[16] Zeyu Liu, Weicong Liang, Yiming Zhao, Bohan Chen, Lin Liang, Lijuan Wang, Ji Li, and Yuhui Yuan. Glyph-byt5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *arXiv preprint arXiv:2406.10208*, 2024. 2

[17] Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870*, 2023. 2, 6, 7

[18] PaddlePaddle. Paddleocr. https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.7/doc/doc_ch/PP-OCRv4_introduction.md, 2023. 7

[19] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 1

[22] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023. 2, 3

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[26] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023. 1, 2, 7

[27] Tong Wang, Xiaochao Qu, and Ting Liu. Textmastero: Mastering high-quality scene text editing in diverse languages and styles. *arXiv preprint arXiv:2408.10623*, 2024.

[28] Yibin Wang, Weizhong Zhang, Changhai Zhou, and Cheng Jin. High fidelity scene text synthesis. *arXiv preprint arXiv:2405.14701*, 2024. 2

[29] Yibin Wang, Weizhong Zhang, Honghui Xu, and Cheng Jin. Dreamtext: High fidelity scene text synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28555–28563, 2025. 2

[30] Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20906–20915, 2025. 2

[31] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 7

[32] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 2

[33] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36:44050–44066, 2023. 1, 2

[34] Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. Textctrl: Diffusion-based scene text editing with prior guidance control. *Advances in Neural Information Processing Systems*, 37:138569–138594, 2024. 2

[35] Yiming Zhao and Zhouhui Lian. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *European conference on computer vision*, pages 217–233. Springer, 2024. 2

[36] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. *arXiv preprint arXiv:2503.12885*, 2025. 2