

ViType: High-Fidelity Visual Text Rendering via Glyph-Aware Multimodal Diffusion

Lishuai Gao¹, Jun-Yan He¹, Yingsen Zeng¹, Yujie Zhong¹, Xiaopeng Sun¹, Jie Hu^{1*}, Xiaoming Wei¹

¹Meituan

Abstract

Current text-to-image models face challenges in visual text rendering: text encoders like CLIP and T5 lack glyph-level understanding and often struggle to distinguish between the specific words to be rendered and their intended semantic meaning within prompts. In addition, inconsistencies between the base model and its plugins further compromise the quality of synthesized images. In this paper, we enhance the existing text-to-image method by addressing the following aspects: (1) Text-Glyph Alignment, a Visual Question Answering (VQA) manner to enable glyph understanding for the text encoder. This involves establishing an explicit alignment between the representations of the glyphs and their detailed attribute descriptions, which boosts the model’s ability to capture fine-grained visual features of the text. (2) Accurate and harmony visual text rendering: integrating pre-aligned glyph-visual embeddings with semantic text tokens through the Multimodal Diffusion Transformer (MMDiT) synchronously, ensuring coherent feature alignment and enhancing both the robustness and fidelity of visual text rendering. (3) Image Aesthetic Refinement: leveraging a multisource data training strategy that incorporates diverse, high-quality image-text pairs from various domains, exposing the model to extensive linguistic and visual diversity while maintaining superior aesthetic quality throughout training. Our experiments demonstrate that the proposed approach significantly outperforms the existing state-of-the-art method.

Introduction

The rapid development of e-commerce has accelerated the demand for advanced intelligent design solutions. In this context, text-to-image (T2I) generation (Ho, Jain, and Abbeel 2020; Song et al. 2021; Dhariwal and Nichol 2021; Nichol and Dhariwal 2021; Saharia et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Chang et al. 2023), which enables the automatic creation of high-quality visual content from textual product descriptions, has become a fundamental technique that significantly enhances user engagement and streamlines digital marketing workflows. Notably, the complexity of accurately rendering textual characters within images poses significant challenges for intelligent design systems in text-to-image generation. In particular, en-

sureing that the generated visual content faithfully represents the intended characters requires precise modeling of various scripts and glyph structures. Moreover, supporting multilingual generation while fostering diversity and creativity further complicates the task due to linguistic forms and cultural aesthetic variations. Consequently, there are three major objectives in text-to-image intelligent design: 1) achieving high accuracy in character-level text generation; 2) enabling diverse and creative image synthesis; and 3) ensuring a high degree of harmony and integration between textual and visual elements.

Despite previous research (Ma et al. 2023a; Yang et al. 2023a; Tuo et al. 2023a; Chen et al. 2023a) on visual rendering, they have mainly focused on multi-plugin modes combined with basic models and are unsuitable for high-quality T2I intelligent design. In particular, three major issues remain: (1) lack of glyph-level understanding: current text encoders (Raffel et al. 2020) fail to capture multilingual character visual shapes and glyph morphology; (2) ambiguity: the text encoder often struggles to distinguish between the exact textual content to be rendered from the prompt and its intended semantic meaning; and (3) Inconsistency in visual rendering-plugin or adapter-based approaches (Ma et al. 2023a; Yang et al. 2023a; Tuo et al. 2023a; Chen et al. 2023a) suffer from distributional discrepancies between plugin data and base model training data, undermining overall rendering consistency. Furthermore, enhancing visual character symbol generation jointly with the base model has not been thoroughly investigated in T2I scenarios. To address these limitations, we propose a *Visual-Glyph Integrated Typeface Yielding Precision Enhancement Pipeline, Vi-Type*. This practical T2I visual rendering pipeline improves performance in four key aspects:

1. *To empower the text encoder with the visual glyph pattern*, we introduce a dedicated text-glyph alignment phase by utilizing a native multilingual text encoder (Bai et al. 2025). In this stage, textual annotations are leveraged to semantically align each character symbol with its corresponding in the embedding space. This explicit alignment not only enables the model to capture subtle variations and intricate structural features of glyphs but also establishes a robust foundation for context-aware text generation in downstream T2I tasks.
2. *To learn visual text rendering accurately for the text-to-*

*Corresponding author and project lead.



Figure 1: The images synthesized by the proposed ViType method demonstrate high accuracy, diversity, and visual appeal in text rendering. Each segment showcases crisp, well-integrated Chinese typography that harmonizes with varied backgrounds—from pastoral scenes to dynamic urban art—highlighting both the versatility and aesthetic refinement of ViType’s innovative approach.

image model, we develop a joint training framework to enhance the accuracy of text-to-image synthesis in typographic generation. The proposed methodology integrates pre-aligned glyph-visual embeddings with semantic text tokens synchronously through the MMDiT (Labs 2024; Labs et al. 2025) architecture, establishing coordinated feature spaces that facilitate precise generation of character morphology. This holistic approach facilitates seamless integration of multilevel representations across modalities, ensuring coherent feature alignment and enhancing both the robustness and fidelity of visual text rendering with the T2I pipeline. Moreover, the incorporation of glyph visual embeddings combined with semantic information effectively resolves ambiguities present in traditional text encoders, thereby significantly improving the accuracy of character generation.

3. To ensure high aesthetic and character accuracy, we employ a comprehensive Supervised Fine-Tuning(SFT) approach focused on image aesthetics. By training on diverse, high-quality image-text pairs from heterogeneous domains, our method enhances the model’s ability to capture semantic nuances and stylistic variations, resulting in visually appealing outputs that faithfully reflect intended content.
4. To enable strong instruction-following capabilities, we curate an extensive dataset with meticulous, structured annotation. This process involves collecting large volumes of data and providing detailed labels for textual properties—such as glyph attributes and layout—as well as background context. Such fine-grained supervision en-

ures robust learning of text rendering and visual-textual integration within complex scenes.

In summary, ViType offers a state-of-the-art solution for high-quality visual text rendering in the T2I pipeline. By introducing a dedicated pre-alignment phase for glyph property encoding, a visual text learning framework to ensure cross-modal consistency, and a comprehensive multisource data strategy for aesthetic quality and accuracy of generated images, ViType effectively addresses longstanding challenges in this domain. Extensive experiments demonstrate that our approach surpasses existing methods across multiple benchmarks. Furthermore, ViType is readily applicable to diverse intelligent design scenarios within e-commerce and digital content creation, paving the way for more creative and accurate automated visual generation systems.¹.

Related Work

Text-to-Image Synthesis. The development of text-to-image diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021; Dhariwal and Nichol 2021; Nichol and Dhariwal 2021; Saharia et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Chang et al. 2023) showcases their powerful capabilities in the field of image generation. The denoising diffusion probabilistic model first demonstrated the potential of image generation in 2020, and subsequent research, including works by Ramesh et al (Ramesh et al. 2022; Rombach et al. 2022), further validated the possibility of using text prompts for high-quality image syn-

¹We will release our system and codebase to foster further research and practical adoption in the community

thesis. GLIDE (Nichol et al. 2022) highlights the advantages of classifier-free guidance in high-resolution generation, while latent diffusion models successfully place the diffusion process in latent space, significantly reducing computational costs. Stable Diffusion and its enhanced version, SDXL (Podell et al. 2023), improve text-to-image generation performance by training on larger datasets. Unlike conventional U-Net architectures, Stable Diffusion3 (Esser et al. 2024) employs the MMDiT (Peebles and Xie 2023) architecture, integrating multiple text embedding technologies to acquire richer semantic information, thus augmenting image generation capabilities. Diffusion models have made significant impacts across various domains, including digital arts, gaming, and advertising, although the readability of the generated text remains an area for improvement. As diffusion model technology continues to advance, particularly with the widespread adoption of transformer-based denoisers, these models demonstrate remarkable scalability and efficacy in the wide-ranging applications of high-quality visual content generation.

Visual Text Generation. Text generation and rendering are classic tasks. Recent studies on visual text generation based on diffusion models focus on optimizing text encoding and spatial control mechanisms to enhance the fidelity and controllability of rendered text. To achieve precise spatial control of text, some methods introduce explicit glyph and layout conditions. TextDiffuser (Chen et al. 2023b), JoyType (Li et al. 2024), and ControlText (Jiang et al. 2025) handle this by segmenting text areas or applying text layout masks, whereas UDiffText (Zhao and Lian 2024) and EasyText (Lu et al. 2025) strengthen region-level alignment through optimized attention mechanisms. However, these methods of forcibly controlling text positions can impair the fusion effect between text and background. In terms of text encoding, glyph/character-aware encoders are incorporated as plugins into fundamental diffusion models. GlyphDraw (Ma et al. 2023b), GlyphControl (Yang et al. 2023b), and AnyText (Tuo et al. 2023b) embed glyph or ocr features into conditional inputs, while FluxText (Lan et al. 2025) combines using ocr features and Glyph-ByT5’s (Liu et al. 2024b) characteristics. Nonetheless, semantic bias persists, and performance is suboptimal when dealing with multi-scale and multilingual (particularly ideographic characters like Chinese) scenarios, where glyph encoders often require more complex alignment processes (such as Glyph-ByT5). The complex alignment process and text remain challenges, prompting further integration of glyph, position, and semantic signals to achieve higher-quality text generation and rendering.

Methodology

Preliminary. The task of visual text rendering is to generate an image whose theme aligns with the text prompt and renders the text prompt onto the image. Here, the text prompt can be decomposed into multiple text lines, denoted as $\mathbf{U} = \{U_1, U_2, \dots, U_n\}$, where each U_i represents the i -th text line. To achieve high-fidelity and reliable editing results, our approach utilizes a pretrained FLUX-like T2I model in which the text encoder (CLIP & T5) is replaced by the

Qwen2.5-7B (Team 2024b) Large Language Model(LLM) as our base model (see Supplementary for more details). The loss function is defined as follows:

$$\mathcal{L}_d = \mathbb{E}_{z_0, z_a, c_{te}, t, \epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_\theta(z_t, c_{mul}, t)\|_2^2]. \quad (1)$$

The \mathcal{L}_d represents the rectified flow loss (RF loss for brevity), z_t represents the latent variable at time step t , while c_{mul} corresponds to the multimodal conditional embeddings produced by the text-glyph pre-aligned model. ϵ_θ denotes a MMDiT denoiser that is utilized to estimate the noise added to the noisy latent image z_t with the objective \mathcal{L}_d .

Text-Glyph Alignment

Typically, Multi-modal Large Language Models (MLLMs) include a visual foundation model (VFM), a modality connector, and an LLM. Firstly, we render the corresponding glyph-line images $\mathbf{X} = \{X_i, X_2, \dots, X_n\}$ based on the text lines \mathbf{U} . Then, following the prevalent multi-scale adaptive cropping strategy, the glyph-line image $X_i \in \mathbb{R}^{H \times W \times 3}$ is processed by a VFM to extract visual tokens $v_i \in \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times C}$, where p is the patch size, set to 14 by default. After individually processing each line U_i , we obtain a sequence of visual tokens $\mathbf{V} = \{v_1, v_2, \dots, v_{n_v}\}$. Here, n_v denotes the number of visual tokens. For prompt input, the text is embedded as $\mathbf{T} = \{t_1, t_2, \dots, t_{n_t}\}$, where n_t is the number of text tokens. The modality connector acts as a projector that maps the visual tokens into language space. Finally, the visual tokens \mathbf{V} and text tokens \mathbf{T} are concatenated together to be fed into the LLM to generate a response. Specifically, the LLM process can be described as:

$$\mathbf{H}\mathbf{V}^{k+1}, \mathbf{H}\mathbf{T}^{k+1} = \text{Layer}_{\text{LLM}}((\mathbf{H}\mathbf{V}^k, \mathbf{H}\mathbf{T}^k)), \quad (2)$$

where $k \in \{1, \dots, K\}$, K is the number of layers in the LLM, and $\mathbf{H}\mathbf{V}^{k+1}, \mathbf{H}\mathbf{T}^{k+1}$ refer to the output hidden states of the k -th layer of LLM. Notably, we omit the superscript for $k = 0$, because these inputs consist of the text token \mathbf{T} and the visual token \mathbf{V} aforementioned. In this work, we implement our Text-Glyph alignment framework with the widely-used Qwen2.5-7B architecture (Bai et al. 2025), a transformer-based model comprising 32 stacked layers.

VQA-based Text Parsing. Following existing works (Wang et al. 2025; Wei et al. 2024; Liu et al. 2024a; Hu et al. 2024b,a), we introduce the text parsing task to align images and text at the semantic level implicitly. The outputs of the last layer of LLM are utilized to predict these answers \mathbf{T}_a , and the optimization loss is formulated as follows:

$$\mathcal{L}_{vqa} = \mathbf{T}_a \log p(\mathbf{H}\mathbf{T}_a^k | \mathbf{V}, \mathbf{T}). \quad (3)$$

Visual Text Learning

Multimodal Condition. Instead of utilizing the only text token as the condition for the MMDiT (Labs 2024; Labs et al. 2025) model, we employ a multimodal condition generated by the text-glyph aligned model for our image generation tasks, which induces the visual glyph shape pattern for better visual text rendering. Specifically, the multimodal condition $c_{mul} = [\mathbf{H}\mathbf{T}^K, \mathbf{H}\mathbf{V}^K]$.

Representation Alignment. To effectively supervise the Glyph Encoder’s training with glyph-specific information,

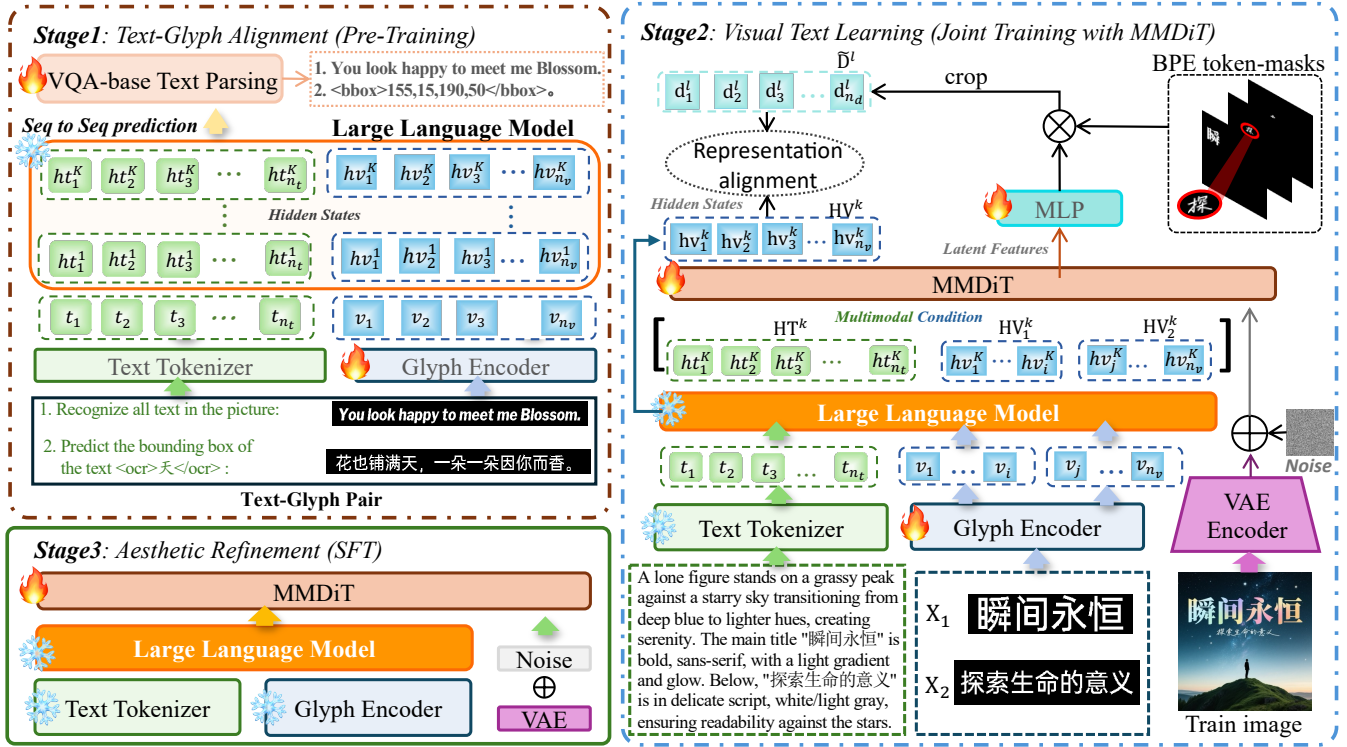


Figure 2: The proposed ViType method follows a three-stage training framework: 1) Text-Glyph Aligned Stage: This stage incorporates visual glyph representations created by the glyph encoder into the LLM, enabling the establishment of cross-modal correspondence between textual semantics and glyph patterns. 2) Visual Text Learning Stage: Co-training the glyph encoder with MMDiT enhances the precision of the model in rendering structurally accurate glyph-based visual text; 3) Aesthetic Refinement Stage: Supervised fine-tuning with curated high-quality datasets elevates the typographic aesthetics and visual harmony of generated outputs.

we select the latent features \mathbf{D}^l from the l -th layer of MMDiT to supervise visual features \mathbf{V}^{k+1} from the k -th layer of the LLM. We avoid using the final reconstructed image or the initial noise, because, as mentioned in (Yu et al. 2024), these extremes may either lack sufficient detail or be too noisy for effective supervision. To explicitly facilitate spatial-wise visual-latent alignment at the LLM level, we conduct a fine-grained alignment task with token granularity by leveraging the BPE token masks $\{M_1, M_2, \dots, M_{n_m}\}$ as referenced from previous works (Guan et al. 2025). We further obtain $\tilde{\mathbf{D}}^l = M_i \circ \text{BI}(\mathbf{D}^l)$. The symbol \circ denotes the Hadamard product, $\text{BI}(\cdot)$ refers to the bilinear interpolation operation to match the feature resolution of M_i . Here, we employ a representation alignment loss. Specifically, we calculate the cosine similarity between these two sets of features and use it to form our loss function, as:

$$\mathcal{L}_{\text{align}} = - \sum \cos(\mathbf{V}^k, \tilde{\mathbf{D}}^l), \quad (4)$$

where k is set to 4 in this work.

Overall Objective During training, we jointly update the parameters of the Glyph Encoder and the MMDiT. The total loss function is a weighted combination of the rectified flow feature alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_d + \alpha * \mathcal{L}_{\text{align}}. \quad (5)$$

where α is a hyperparameter that controls the relative importance of representation alignment and is set to 0.6.

Aesthetic Refinement

In this stage, we collected high-quality data related to text generation scenarios and performed fine-grained processing, which will be introduced in the subsequent sections. Specifically, our Glyph Encoder inherits the weights from Stage 2. Representation alignment is discarded in this phase. All parameters of MMDiT are updated for SFT. Through these three phases of training, the proposed model is capable of generating images with high-quality backgrounds and highly accurate text.

Data Collection and Curation

We have gathered a substantial collection of original images enriched with diverse textual content sourced from professional design websites and social networks. The effort culminated in the creation of a comprehensive visual text rendering dataset. This dataset encompasses a wide range of domains, including creative posters, marketing posters, film and television posters, book covers, user interface data from applications, comic illustrations, and images capturing real-life scenarios. After data cleaning and detailed captioning, we develop a private dataset focused specifically on tex-



Figure 3: This figure presents an example of integrated image and text caption analysis, focusing on the structural organization of textual elements within a poster design.

tual elements. This dataset comprises over 20 million high-quality image-text pairs, each thoroughly annotated with detailed captions. Furthermore, we have strategically focused on collecting text data embedded with Chinese contextual information to enhance performance within Chinese linguistic domains.

Data Cleaning and Filtering. To construct high-quality training data, we filter our collected images based on three essential criteria: image quality, textual characteristics, and aesthetic appeal. Firstly, we evaluate image quality by eliminating entries that exhibit low resolution, lack clarity, suffer from unbalanced aspect ratios, or display motion blur. Next, with regard to textual characteristics, we employ both traditional optical character recognition (OCR) models and visual-language models (VLMs) to extract text from images, discarding those with empty text fields or watermarked content. Additionally, images with excessively dense text or text that is particularly small are excluded from the dataset. Lastly, we evaluate the aesthetic quality of the images using VLMs. We specifically dissect image aesthetics into three components: visual elements, textual content, and the integration of text and imagery—referred to as layout. Each of these components is individually scored by the VLMs. To tailor the evaluation to various types of images, we apply different weights to these scores, ensuring a nuanced and comprehensive assessment of quality across diverse categories. This weighted scoring process allows us to meticulously refine our collection, ultimately extracting high-quality data suitable for training purposes.

Caption for Text Property and Background. Text rendering data necessitates special consideration of text glyphs. To address this, we have developed a two-stage automated captioning pipeline, which includes glyph-centered and global layout descriptions. The first stage aims to ensure that captions focus on various text attributes. We accomplish this by incorporating earlier OCR results into the prompt, thus generating captions centered on text structure. This facilitates the extraction and output of detailed typographic features for each phrase or sentence within the image. These features in-

clude layout aspects (e.g. position, relationship to visual elements, alignment, etc.), text functions (e.g. main titles, slogans, dates, etc.), font characteristics (e.g. font type, shape, unique transformation features, etc.), as well as descriptions of color texture and stylistic details. The second stage involves infusing the structured textual information from the first stage into the input prompt for VLMs. This empowers the VLMs to focus on background context and the integration of text with imagery, without needing to concentrate on the text’s physical form. Ultimately, this results in the generation of captions that richly describe both the text attributes and the background, as exemplified in Fig. 3. Specific prompts and fields used in this process are detailed in the supplementary materials.

Experiment

Implementation Details

Stage 1. In practical implementation, Glyph Encoder selects DINOv2-L (Oquab et al. 2023) as the visual foundation model, and Qwen2.5 (Qwen et al. 2025), a 7B large language model as the language decoder. The batch size is set to 32 per GPU, and the training is conducted on 8 H800 GPUs for 10 hours.

Stage 2 & Stage 3. The learning rate for the Diffusion module is configured to $3e-4$, utilizing the AdamW optimizer. Initially, a linear warmup is applied, followed by cosine decay to $1e-5$. The batch size is set to 16 per GPU. Training is executed on 16 H800 GPUs over a duration of 30 hours.

Evaluation

The evaluation set is divided into five distinct parts, each serving to assess the model’s performance.

AnyText-Benchmark (Tuo et al. 2023b) contains one thousand English images and Chinese images from LAION (Schuhmann et al. 2021) and Wukong (Gu et al. 2022) respectively.

ICDAR13 (Karatzas et al. 2013) serves as the benchmark for assessing the detection of near-horizontal text, and it consists of 233 images for testing purposes.

MARIO-Eval (Chen et al. 2023b) serves as a comprehensive tool for evaluating text rendering quality collected from the subset of MARIO-10M test set and other sources.

Complex-Benchmark (Ma et al. 2025) comprises 200 bilingual Chinese and English prompts. In the Chinese prompts, the characters to be rendered are randomly composed with intricate strokes and structures, whereas the English prompts feature longer words with consecutive letter repetitions.

Poster-Benchmark (Ma et al. 2025) includes 240 prompts that describe the generation of posters. This benchmark aims to assess the layout accuracy, robustness, and overall aesthetic quality of automatically generated posters.

Evaluation Metrics. For these evaluation sets, we utilized four evaluation metrics to assess the accuracy and quality of poster generation: (1) **Accuracy (Acc)** calculates the proportion of correctly generated characters in the rendered text compared to the total number of characters that need to be rendered. (2) **Normalized Edit Distance (NED)**, the calculation method remains consistent with AnyText. (3) **Clip-**

| Benchmark | Model | Chinese | | | | English | | | |
|-------------------|----------------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| | | Acc | NED | ClipScore | HPSv2 | Acc | NED | ClipScore | HPSv2 |
| AnyText-Benchmark | SD3† | - | - | - | - | 0.3261 | - | 0.4517 | 0.2215 |
| | Kolors† | 0.0665 | - | 0.4011 | 0.2654 | 0.0243 | - | 0.4854 | 0.2512 |
| | FLUX.1-schnell† | - | - | - | - | 0.3884 | - | 0.4914 | 0.2541 |
| | ControlNet† | 0.7598 | 0.8254 | 0.3749 | 0.2347 | 0.7098 | 0.8467 | 0.4558 | 0.2245 |
| | ControlNet w/ canny† | 0.7804 | 0.8365 | 0.3752 | 0.2384 | 0.7954 | 0.8745 | 0.4599 | 0.2287 |
| | TextDiffuser† | 0.0605 | 0.1262 | - | - | 0.5921 | 0.7951 | - | - |
| | AnyText-v1.1† | 0.7661 | 0.8423 | 0.3968 | 0.2272 | 0.7108 | 0.8564 | 0.4721 | 0.2121 |
| | AnyText-v2 | 0.7993 | 0.8457 | 0.3991 | 0.2563 | 0.8159 | 0.9057 | 0.4715 | 0.2652 |
| | UDiffText† | - | - | - | - | 0.6435 | 0.8284 | 0.4645 | 0.2214 |
| | Glyph-ByT5-v1† | 0.7227 | 0.7799 | 0.4005 | 0.2601 | 0.7307 | 0.8353 | 0.4802 | 0.2511 |
| | Glyph-ByT5-v2†† | 0.7998 | 0.8441 | 0.3858 | 0.2625 | 0.7659 | 0.8842 | 0.4558 | 0.2315 |
| | GlyphDraw1.1† | 0.7892 | 0.8476 | 0.3921 | 0.2555 | 0.7369 | 0.8921 | 0.4616 | 0.2350 |
| | GlyphDraw2† | 0.8266 | 0.8543 | 0.3986 | 0.2589 | 0.8627 | 0.9278 | 0.4796 | 0.2451 |
| | PosterMaker | 0.9015 | 0.9004 | 0.4123 | 0.2726 | 0.9148 | 0.9376 | 0.4759 | 0.2557 |
| | ViType | 0.9215 | 0.9014 | 0.4365 | 0.2896 | 0.9326 | 0.9632 | 0.4995 | 0.2653 |
| ICDAR13 | UDiffText† | - | - | - | - | 0.5840 | 0.7221 | 0.4521 | 0.2101 |
| | GlyphDraw2† | - | - | - | - | 0.6901 | 0.7629 | 0.4657 | 0.2345 |
| | ViType | - | - | - | - | 0.7559 | 0.8362 | 0.4985 | 0.2802 |
| MARIO-Eval | TextDiffuser† | - | - | - | - | 0.5609 | - | - | - |
| | GlyphDraw2† | - | - | - | - | 0.7672 | 0.9330 | 0.4765 | 0.2464 |
| | ViType | - | - | - | - | 0.8852 | 0.9753 | 0.5245 | 0.2871 |
| Complex-Benchmark | SD3† | - | - | - | - | 0.2515 | - | 0.4391 | 0.2492 |
| | Kolors† | 0.0198 | - | 0.3878 | 0.2546 | 0.0033 | - | 0.4254 | 0.2546 |
| | FLUX.1-schnell† | - | - | - | - | 0.2969 | - | 0.4298 | 0.2544 |
| | ControlNet† | 0.6943 | 0.8745 | 0.3589 | 0.2364 | 0.2254 | 0.4025 | 0.4214 | 0.2385 |
| | ControlNet w/ canny† | 0.7546 | 0.8812 | 0.3512 | 0.2386 | 0.4215 | 0.4532 | 0.4311 | 0.2298 |
| | AnyText-v1.1† | 0.5749 | 0.8560 | 0.3633 | 0.2434 | 0.0342 | 0.3755 | 0.4104 | 0.2312 |
| | Glyph-ByT5† | 0.7895 | 0.8263 | 0.3711 | 0.2455 | 0.4834 | 0.7034 | 0.4256 | 0.2412 |
| | Glyph-ByT5-v2†† | 0.8221 | 0.8645 | 0.3554 | 0.2331 | 0.5477 | 0.6345 | 0.4089 | 0.2116 |
| | LLMs+ControlNet† | 0.5812 | 0.8012 | 0.3687 | 0.2365 | 0.1856 | 0.5841 | 0.4215 | 0.2356 |
| | TextDiffuser-2† | - | - | - | - | 0.0999 | 0.4428 | 0.3985 | 0.2285 |
| | LLMs+AnyText-v1.1† | 0.4850 | 0.7888 | 0.3697 | 0.2534 | 0.0455 | 0.4680 | 0.4038 | 0.2380 |
| | GlyphDraw1.1† | 0.6215 | 0.8479 | 0.3756 | 0.2427 | 0.2264 | 0.6273 | 0.4362 | 0.2415 |
| | GlyphDraw2† | 0.6691 | 0.7975 | 0.3754 | 0.2498 | 0.4158 | 0.6294 | 0.4312 | 0.2488 |
| | ViType | 0.9325 | 0.9187 | 0.4026 | 0.2988 | 0.8351 | 0.7469 | 0.4849 | 0.2758 |
| Poster-Benchmark | SD3† | - | - | - | - | 0.2310 | - | 0.4128 | 0.2337 |
| | Kolors† | 0.0426 | - | 0.4110 | 0.2510 | 0.0020 | - | 0.4120 | 0.2421 |
| | FLUX.1-schnell† | - | - | - | - | 0.3744 | - | 0.4215 | 0.2541 |
| | ControlNet† | 0.7878 | 0.8453 | 0.3844 | 0.2298 | 0.3421 | 0.7514 | 0.3902 | 0.2125 |
| | ControlNet w/ canny† | 0.7911 | 0.8541 | 0.3801 | 0.2225 | 0.5012 | 0.8014 | 0.3955 | 0.2106 |
| | TextDiffuser-2† | - | - | - | - | 0.1046 | 0.3623 | 0.3914 | 0.2110 |
| | LLMs+AnyText-v1.1† | 0.7421 | 0.8894 | 0.3956 | 0.2362 | 0.2604 | 0.7120 | 0.4093 | 0.2289 |
| | Glyph-ByT5† | 0.8248 | 0.9040 | 0.4012 | 0.2366 | 0.7341 | 0.8411 | 0.4101 | 0.2354 |
| | Glyph-ByT5-v2†† | 0.8512 | 0.9133 | 0.4278 | 0.2389 | 0.7554 | 0.7641 | 0.4012 | 0.2113 |
| | GlyphDraw1.1† | 0.8215 | 0.9590 | 0.3908 | 0.2378 | 0.3999 | 0.7667 | 0.3984 | 0.2297 |
| | GlyphDraw2† | 0.8263 | 0.9585 | 0.3987 | 0.2314 | 0.7590 | 0.8759 | 0.4114 | 0.2301 |
| | ViType | 0.9286 | 0.9831 | 0.4641 | 0.2863 | 0.9025 | 0.9354 | 0.4772 | 0.2713 |

Table 1: Evaluation Results on five benchmarks.

Score measures how well the generated image aligns with the textual prompt or description provided. (4)**HPSv2** (Wu et al. 2023) whether the generated images align with human preferences and serve as an indicator to assess the quality of the images.

In our comparison study, we assessed a variety of methods, generally grouped into three categories. The first category encompasses large-scale text generation models recently made available as open-source by the industry, featuring font rendering capabilities. This includes StableDiffusion3 (SD3) (Esser et al. 2024), Kolors (Team 2024a), and the FLUX.1 series developed by Black Forest Labs. Notably, only SD3 and FLUX.1 provide support for English. Additionally, due to NED calculations typically being anchored on text box positioning, we have chosen to omit NED calculations for methods in this category. The second category covers open-source approaches to font rendering in text gen-

eration, such as the TextDiffuser series (Chen et al. 2023b, 2024a), AnyText series (Tuo et al. 2023b; Tuo, Geng, and Bo 2024), UDiffText, Glyph-ByT5 series (Liu et al. 2024b,c), and PosterMaker (Gao et al. 2025). The third category compares experiments utilizing the basic ControlNet (Yang et al. 2023b), featuring two distinct types of conditional inputs: the first involves earlier methods that directly utilize rendered fixed-fonts as a condition (Ma et al. 2023b; Yang et al. 2023b; Tuo et al. 2023b), while the second relies on the canny edge of the original font from training images as a condition.

Experimental Results

In the following section, we provide a comprehensive analysis of both quantitative and qualitative results, comparing our method with state-of-the-art approaches in the fields of text rendering and poster generation. To guarantee a fair

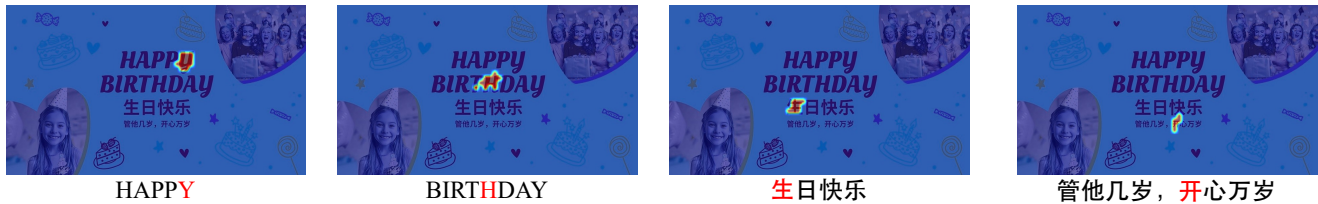


Figure 4: Visualization of the cross attention maps within our Diffusion model. We show the heat maps between all pixels and the selected characters from the text lines.

evaluation, we adopted certain assessment settings inspired by (Liu et al. 2024c): all methods utilized a sampling step of 50, along with a fixed random seed set to 100. We record all the comparative experiments in Table 1, where † denotes results from evaluations within (Liu et al. 2024c), †† denotes the absence of open-sourced pre-trained weights, which we subsequently reproduced on our dataset.

Ablation Experiments

All ablation studies are performed by first implementing Stage 2 Visual Text Learning, then proceeding with training the ViType model using our graphic design benchmarks.

| Visual Encoder Setting | Acc | | | |
|-------------------------------|---------------|---------------|---------------|---------------|
| | ≤20 chars | ≤20-50 chars | ≤50-100 chars | ≥100 chars |
| (Baseline) w/o Visual Encoder | 0.7892 | 0.7586 | 0.6857 | 0.6209 |
| SigLIP2-so400m | 0.8735 | 0.8541 | 0.8133 | 0.7012 |
| InternViT-300M | 0.9217 | 0.8657 | 0.8379 | 0.7187 |
| Qwen2.5-ViT | 0.9346 | 0.8848 | 0.8532 | 0.7653 |
| DINOv2-L | 0.9583 | 0.8894 | 0.8762 | 0.7964 |

Table 2: Ablation study on visual encoder and the selection of visual encoders reveal that visual encoders significantly enhance text accuracy, with DINOv2-L demonstrating the best performance.

Glyph Encoder Ablation and Selection. Initially, we use a basic LLM without the Glyph Encoder as our baseline. Subsequently, we validate the effectiveness of incorporating the Glyph Encoder. We examine the impact of choosing four different pre-trained visual encoders: SigLIP2-so400m (Tschannen et al. 2025), InternViT-300M (Chen et al. 2024b), and Qwen2.5-ViT (Bai et al. 2025), all of which have undergone visual-language space alignment. Additionally, we consider DINOv2-L (Oquab et al. 2023), which has not been pre-aligned. For DINOv2-L, we adopt the alignment strategy described in COMP (Chen et al. 2025) to integrate it into the language space. Detailed comparison results are presented in Table 2.

| Method | Acc | | | |
|------------------------------|---------------|---------------|---------------|---------------|
| | ≤20 chars | ≤20-50 chars | ≤50-100 chars | ≥100 chars |
| w/o representation alignment | 0.8735 | 0.8541 | 0.8133 | 0.7012 |
| w/ representation alignment | 0.9583 | 0.8894 | 0.8762 | 0.7964 |

Table 3: Effect of representation alignment.

Efficiency of representation alignment. We study the effect of representation alignment during Stage 2. As indicated in Table 3, representation alignment provides a notable improvement compared to the non-alignment setting.

| hidden layer k | Acc | | | |
|----------------|---------------|---------------|---------------|---------------|
| | ≤20 chars | ≤20-50 chars | ≤50-100 chars | ≥100 chars |
| 2 | 0.9395 | 0.8876 | 0.8753 | 0.7864 |
| 4 | 0.9583 | 0.8894 | 0.8762 | 0.7964 |
| 26 | 0.8735 | 0.8541 | 0.8133 | 0.7012 |
| 28 | 0.8412 | 0.8264 | 0.8076 | 0.7009 |

Table 4: The results of the ablation experiments regarding the selection of different hidden layers.

Selection hidden layer for LLM. Our LLM consists of 28 hidden layers, while MMDiT comprises 30 layers. According to (Yu et al. 2024), we understand that deep-layer image information in MMDiT offers more detail, leading us to select the 29th layer. Previous research by Marten (Wang et al. 2025) highlights the importance of visual and text information in both the first 4 and the last 8 hidden layers of the LLM. Therefore, we examine both shallow and deep layers of the LLM’s hidden layers. We will present the impact on accuracy when choosing different indices of hidden layers k in Table 4.

Qualitative Analysis

Representation Alignment. To gain a deeper understanding of how our Glyph Encoder excels at the visual text rendering task, we further visualize the cross-attention maps between glyph text prompts and rendered images, providing an example in Fig. 4. This visualization confirms that the diffusion model effectively utilizes the glyph-alignment prior encoded within our text encoder.

Aesthetic Refinement. We will provide detailed visual content in the Appendix section.

Conclusion

In conclusion, this work addresses key limitations in current text-to-image generation models by introducing a comprehensive framework that enhances visual text rendering. By aligning text and glyph representations in a VQA-inspired manner, the proposed method enables the encoder to capture fine-grained glyph-level features. The integration of pre-aligned glyph-visual embeddings with semantic tokens ensures accurate and harmonious visual text rendering through coherent feature alignment. Furthermore, the adoption of a multisource data training strategy exposes the model to diverse linguistic and visual contexts while preserving high aesthetic quality. Experimental results demonstrate that our approach achieves substantial improvements over existing state-of-the-art methods, advancing both the fidelity and robustness of synthesized images containing textual content.

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-v1 technical report. *arXiv preprint arXiv:2502.13923*.
- Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; Li, Y.; and Krishnan, D. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. *arXiv preprint, abs/2301.00704*.
- Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2023a. TextDiffuser: Diffusion Models as Text Painters. *arXiv preprint, abs/2305.10855*.
- Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2023b. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36: 9353–9387.
- Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2024a. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, 386–402. Springer.
- Chen, Y.; Meng, L.; Peng, W.; Wu, Z.; and Jiang, Y.-G. 2025. Comp: Continual multimodal pre-training for vision foundation models. *arXiv preprint arXiv:2503.18931*.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12): 220101.
- Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 8780–8794.
- Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Gao, Y.; Lin, Z.; Liu, C.; Zhou, M.; Ge, T.; Zheng, B.; and Xie, H. 2025. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8083–8093.
- Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35: 26418–26431.
- Guan, T.; Wang, Z.; Fu, P.; Guo, Z.; Shen, W.; Zhou, K.; Yue, T.; Duan, C.; Sun, H.; Jiang, Q.; et al. 2025. A token-level text image foundation model for document understanding. *arXiv preprint arXiv:2503.02304*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *NeurIPS*.
- Hu, A.; Xu, H.; Ye, J.; Yan, M.; Zhang, L.; Zhang, B.; Li, C.; Zhang, J.; Jin, Q.; Huang, F.; et al. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.
- Hu, A.; Xu, H.; Zhang, L.; Ye, J.; Yan, M.; Zhang, J.; Jin, Q.; Huang, F.; and Zhou, J. 2024b. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*.
- Jiang, B.; Yuan, Y.; Bai, X.; Hao, Z.; Yin, A.; Hu, Y.; Liao, W.; Ungar, L.; and Taylor, C. J. 2025. Controltext: Unlocking controllable fonts in multilingual text rendering without font annotations. *arXiv preprint arXiv:2502.10999*.
- Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, 1484–1493. IEEE.
- Labs, B. F. 2024. FLUX. <https://github.com/black-forest-labs/flux>.
- Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv:2506.15742*.
- Lan, R.; Bai, Y.; Duan, X.; Li, M.; Sun, L.; and Chu, X. 2025. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329*.
- Li, C.; Jiang, C.; Liu, X.; Zhao, J.; and Wang, G. 2024. Joytype: A robust design for multilingual visual text creation. *arXiv preprint arXiv:2409.17524*.
- Liu, Y.; Yang, B.; Liu, Q.; Li, Z.; Ma, Z.; Zhang, S.; and Bai, X. 2024a. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.
- Liu, Z.; Liang, W.; Liang, Z.; Luo, C.; Li, J.; Huang, G.; and Yuan, Y. 2024b. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *European Conference on Computer Vision*, 361–377. Springer.
- Liu, Z.; Liang, W.; Zhao, Y.; Chen, B.; Liang, L.; Wang, L.; Li, J.; and Yuan, Y. 2024c. Glyph-byt5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *arXiv preprint arXiv:2406.10208*.
- Lu, R.; Zhang, Y.; Liu, J.; Wang, H.; and Song, Y. 2025. EasyText: Controllable Diffusion Transformer for Multilingual Text Rendering. *arXiv preprint arXiv:2505.24417*.
- Ma, J.; Deng, Y.; Chen, C.; Du, N.; Lu, H.; and Yang, Z. 2025. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5955–5963.
- Ma, J.; Zhao, M.; Chen, C.; Wang, R.; Niu, D.; Lu, H.; and Lin, X. 2023a. GlyphDraw: Learning to Draw Chinese Characters in Image Synthesis Models Coherently. *arXiv preprint, abs/2303.17870*.
- Ma, J.; Zhao, M.; Chen, C.; Wang, R.; Niu, D.; Lu, H.; and Lin, X. 2023b. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870*.

- Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In Meila, M.; and Zhang, T., eds., *ICML*, volume 139, 8162–8171.
- Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*, volume 162, 16784–16804.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *CoRR*, abs/2307.01952.
- Qwen; ; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. *arXiv:2412.15115*.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*, 21: 140:1–140:67.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint*, abs/2204.06125.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 10684–10695.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, S. K. S.; Lopes, R. G.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.
- Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.
- Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.
- Team, K. 2024a. Kolos: Effective Training of Diffusion Model for Photorealistic Text-to-Image Synthesis. *arXiv preprint (2024)*. URL https://github.com/Kwai-Kolos/Kolos/blob/master/imgs/Kolos_paper.pdf.
- Team, Q. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- Tuo, Y.; Geng, Y.; and Bo, L. 2024. Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*.
- Tuo, Y.; Xiang, W.; He, J.; Geng, Y.; and Xie, X. 2023a. AnyText: Multilingual Visual Text Generation And Editing. *CoRR*, abs/2311.03054.
- Tuo, Y.; Xiang, W.; He, J.-Y.; Geng, Y.; and Xie, X. 2023b. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*.
- Wang, Z.; Guan, T.; Fu, P.; Duan, C.; Jiang, Q.; Guo, Z.; Guo, S.; Luo, J.; Shen, W.; and Yang, X. 2025. Marten: Visual question answering with mask generation for multimodal document understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14460–14471.
- Wei, H.; Kong, L.; Chen, J.; Zhao, L.; Ge, Z.; Yang, J.; Sun, J.; Han, C.; and Zhang, X. 2024. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, 408–424. Springer.
- Wu, X.; Hao, Y.; Sun, K.; Chen, Y.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.
- Yang, Y.; Gui, D.; Yuan, Y.; Ding, H.; Hu, H.; and Chen, K. 2023a. GlyphControl: Glyph Conditional Control for Visual Text Generation. *arXiv preprint*, abs/2305.18259.
- Yang, Y.; Gui, D.; Yuan, Y.; Liang, W.; Ding, H.; Hu, H.; and Chen, K. 2023b. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36: 44050–44066.
- Yu, S.; Kwak, S.; Jang, H.; Jeong, J.; Huang, J.; Shin, J.; and Xie, S. 2024. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*.
- Zhao, Y.; and Lian, Z. 2024. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *European conference on computer vision*, 217–233. Springer.