UM-Text: A Unified Multimodal Model for Image Understanding and Visual Text Editing

Lichen Ma[1*], Xiaolong Fu[1*], GaojingZhou[1], Zipeng Guo[1,2], Ting Zhu[1], Yichun Liu[1], Yu Shi[1], Jason Li[1], Junshi Huang[1†]

[1]JD.COM
[2]Sun Yat-sen University
{malichen2020, fxlcumt, junshi.huang}@gmail.com

Abstract

With the rapid advancement of image generation, visual text editing using natural language instructions has received increasing attention. The main challenge of this task is to fully understand the instruction and reference image, and thus generate visual text that is style-consistent with the image. Previous methods often involve complex steps of specifying the text content and attributes, such as font size, color, and layout, without considering the stylistic consistency with the reference image. To address this, we propose UM-Text, a **u**nified **m**ultimodal model for context understanding and visual text editing by natural language instructions. Specifically, we introduce a Visual Language Model (VLM) to process the instruction and reference image, so that the text content and layout can be elaborately designed according to the context information. To generate an accurate and harmonious visual text image, we further propose the UM-Encoder to combine the embeddings of various condition information, where the combination is automatically configured by VLM according to the input instruction. During training, we propose a regional consistency loss to offer more effective supervision for glyph generation on both latent and RGB space, and design a tailored three-stage training strategy to further enhance model performance. In addition, we contribute the UM-DATA-200K, a large-scale visual text image dataset on diverse scenes for model training. Extensive qualitative and quantitative results on multiple public benchmarks demonstrate that our method achieves state-of-the-art performance.

1 Introduction

Visual text editing and generation play a crucial role in various applications, such as poster design, scene text editing, and the novel task of cross-language image translation. The main challenge of these tasks lies in manual design of text layout, attributes (*e.g.*, font type, size, color), language (*e.g.*, English, Chinese), and visual context (*e.g.*, poster, product image), which are cumbersome and error-prone. In this paper, we propose a method that enables users to perform visual text editing via natural language instructions. Given an input image and editing command, our model automatically generates compelling text content, appropriate layout, and visually harmonious text images with implicit text attributes.

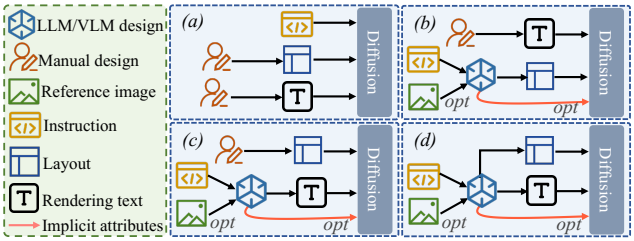

Figure 1: Illustration of traditional framework of visual text generation and three additional generation patterns of our method. Please note that the text content, layout and implicit attributes can be adaptively generated by VLM according to instruction.

Recently, with the rapid advancement of text-to-image (T2I), diffusion models have enabled the creation of highly realistic images with only instructions. For example, Stable Diffusion 3(Esser et al. 2024), FLUX.1(Labs 2024a) and FLUX.1 Kontext(Labs et al. 2025a) has gradually improved its capabilities in general image generation and visual text rendering. However, these commonly used T2I models are still deficient in generating complex characters such as handwriting or art text. Many researchers inject enhanced text generation capabilities into pre-trained diffusion models using various approaches (Tuo et al. 2023; Tuo, Geng, and Bo 2024; Chen et al. 2023b, 2024), but manual interactions are still required for text content and layout.

In visual text generation, text layout plays an important role in generating an appropriate result. Some text generation methods, such as TextDiffuser2(Chen et al. 2024), UniGlyph(Wang et al. 2025a), and GlyphDraw2(Ma et al. 2025), use the task-specific large language model (LLM) to predefine text positions. DesignDiffusion(Wang et al. 2025c) even achieves the layout and visual text image in an end-to-end manner. However, there approaches are still infeasible for visual text editing task, which requires the coordinates of the target text in image. Moreover, the potential of text editing task should be further explored to generate more harmonious and aesthetic visual text images.

In this paper, we propose a holistic framework that inte-

*These authors contributed equally.
†Corresponding Author.

Figure 2: Some results produced by our UM-Text, presenting its powerful effects on tasks such as image editing, image translation, and poster design. Please note that the bounding boxes of text are adaptively generated by UM-Text model.

grates multimodal understanding into the process of visual text generation and editing for implicit learning of text layout and attributes. As illustrated in Fig.1, our framework can support four different text generation and editing patterns, where the reference image and instruction information is extracted as context embeddings for visual text generation/editing. Furthermore, we introduce UM-Encoder, a module for multiple condition aggregation that incorporates T5 embeddings, character-level visual embeddings, and context embeddings. To improve the accuracy of text glyphs, the regional consistency loss in both latent and RGB space are proposed for better glyph supervision. We also contribute the UM-DATA-200K dataset containing 200k diverse image pairs with/without visual text for the pre-training of VLM.

In summary, our contributions are as follows.

- We propose an innovative framework named UM-Text, which combines a unified multimodal understanding and image editing model. With a three-stage training strategy and region-based losses, UM-Text allows flexible visual text generation and editing by simple natural language instructions.
- We introduce UM-Encoder, a novel module for multiple condition aggregation that integrates text embedding, character-level visual embeddings, and multimodal embeddings. With this module, the implicit attributes and layout of text are adaptively generated for visual text generation and editing.
- We contribute a dataset called UM-DATA-200K with manual annotation for visual text generation and editing. Extensive experiments demonstrate the effectiveness of our dataset and framework.

2 Related Work

Image Generation and Understanding Diffusion models have become the primary method for high-quality image synthesis, offering powerful capabilities in terms of photo-realism, fidelity, and diversity. From DDPM (Ho, Jain, and Abbeel 2020) and DDIM (Song, Meng, and Ermon 2020) to Latent Diffusion Models (LDM) (Podell et al. 2023; Rombach et al. 2022; Tian et al. 2024), these models improve generation efficiency and scalability by operating directly within the latent embedding space, enabling image synthesis with higher resolution at lower computational costs. With the introduction of architectures such as DiT (Peebles and Xie 2023; Esser et al. 2024) and FLUX (Labs 2024a), diffusion models have made significant advances in generalization and image quality, laying a solid foundation for unified handling of multimodal conditions and becoming an important architecture in modern image generation (Labs et al. 2025b; Mou et al. 2024; Zhang, Rao, and Agrawala 2023). Despite these advancements, diffusion models still face significant challenges in understanding textual and visual information, highlighting the need to introduce VLM.

VLM (Bai et al. 2025; Team et al. 2025; Zhu et al. 2025; Open AI. 2024; Team et al. 2023) have made significant progress in vision-language understanding tasks. Models like Gemini (Team et al. 2024), Janus-Pro (Chen et al. 2025b), Mogao (Liao et al. 2025), BAGEL (Zhang et al. 2025b) and Nexus-Gen (Zhang et al. 2025a) further unify understanding and generation. Recent works such as MetaQueries (Pan et al. 2025), BLIP3o (Chen et al. 2025a), UniWorld-V1 (Lin et al. 2025), OmniGen2 (Wu et al. 2025), and Step1X-Edit (Liu et al. 2025) integrate VLMs into image generation via multimodal conditioning, exploring control mechanisms and latent-level fusion with diffusion models. However, these methods still face limitations in text rendering: most of them only support English and struggle with editing fine-grained textual regions.

Visual Text Generation and Editing In recent years, there have been substantial developments in the tasks of T2I

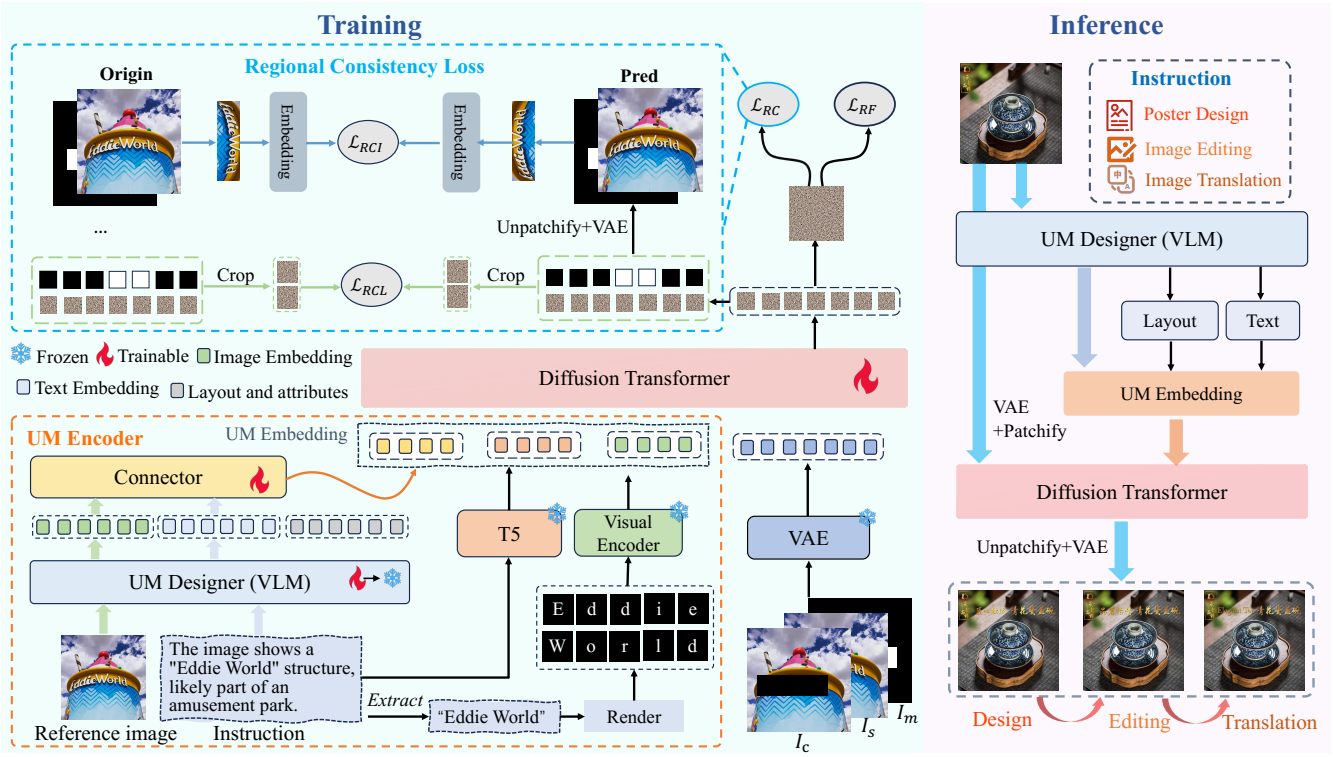

Figure 3: The framework of UM-Text for multi-lingual visual text generation and editing. The UM-Encoder integrates multiple modality embeddings as the condition of visual text generation. The mask in input and loss function is transformed from the predicted layout of UM-Designer. Please note our single model supports diverse downstream applications based on the instructions.

generation and image editing with visual text rendering. The goal of visual text generation and editing models is to produce accurate text images where the visual elements and text layout are harmoniously integrated. Text embeddings and various loss functions are employed to help the model generate more precise text.

DrawText(Liu et al. 2022) demonstrates that character-aware models consistently outperform their character-blind counterparts across various text rendering tasks. GlyphByT5(Liu et al. 2024a,b) and FLUX-Text(Lan et al. 2025) introduce a method utilizing box-level contrastive learning to align text features extracted from the language model with those derived from the visual encoder. In DiffUTE(Chen et al. 2023a) and GlyphDraw(Ma et al. 2023), glyph images are directly incorporated into the text embeddings. AnyText(Tuo et al. 2023), AnyText2(Tuo, Geng, and Bo 2024), and GlyphDraw2(Ma et al. 2025) render a glyph line containing multiple characters into an image, encode glyph information using a pretrained OCR recognition model, and inject it into the text embedding.

Unfortunately, it's challenging to represent multiple characters with a single token, and there's a lack of embedding for image content. To address these issues, we propose UM-Encoder for text embedding injection that integrates T5 embeddings, character-level visual embeddings, and VLM embeddings. This approach enables the model to generate more accurate text while achieving better stylistic consistency with the reference image.

Several T2I methods employ large language models (LLMs) for layout prediction. TextDiffuser has adopted a Layout Transformer that autoregressively outputs bounding boxes for keywords in an encoder-decoder manner. GlyphDraw2 and TextDiffuser2 further leverage LLMs to generate layouts. However, these methods simply learn layout information from the text modality and cannot be directly applied to visual text editing tasks. To overcome this limitation, we propose UM-Designer, a VLM that can simultaneously generate layouts and text related to the reference image.

3 Methodology

In this section, we present the details of UM-Text. We begin to introduce the construction process of UM-DATA-200K in Sec.3.1, which is a large-scale synthetic dataset designed to pretrain the UM-Designer with capabilities in layout planning and text content generation. In Sec.3.2, we present the framework of our UM-Text for visual text generation and editing tasks. Subsequently, Sec.3.3 introduces the UM-Encoder, which integrates various conditions into unified embeddings. In Sec.3.4, we propose a region-wise consistency loss to ensure that the generated text is semantically accurate and stylistically consistent with the reference image. Finally, Sec.3.5 outlines our training strategy.

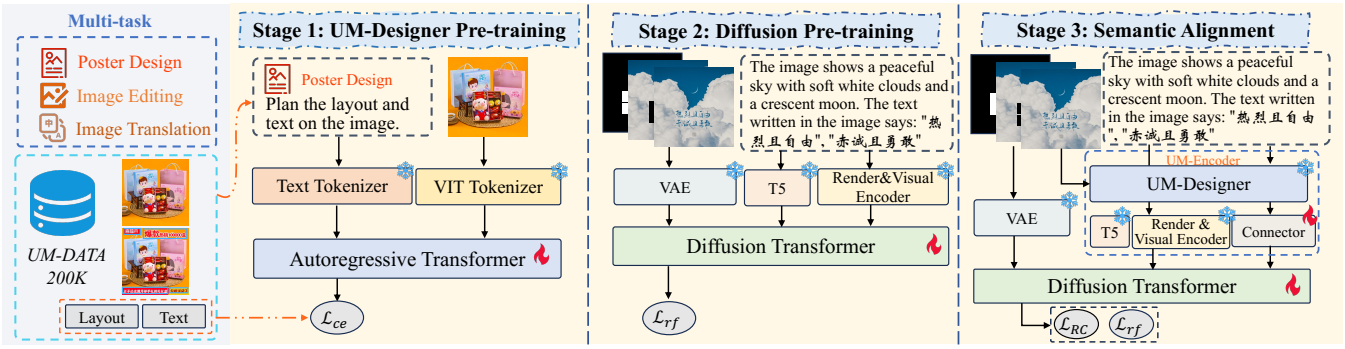

Figure 4: The illustration of three training stages for UM-Text optimization.

3.1 Dataset Construction

Recently, many layout planning and text content generation datasets are limited in the scale or quality of collected data. To address this gap, this paper endeavors to assemble a large-scale, high-quality dataset particularly tailored for layout planning, visual text generation, and editing tasks. Generally, we crawled 40 million product posters from online e-commerce platforms. To construct our dataset, we developed an advanced data pipeline including image aesthetics filtering, object segmentation, OCR, image erasure, and manual annotation.

Specifically, we used the PPOCRv4(Cui et al. 2025) to extract text content and bounding boxes from images and employed Aesthetic Predictor V2.5 to rate the images. We utilized OCR results and aesthetic scores to filter five million images with detailed text layouts and contents. To achieve higher-quality images, we further applied SAM2(Ravi et al. 2024) to segment the main product, filtered out inconsistent text layouts, and used FLUX-Fill (Labs 2024b) to generate clean images based on the text layout. Ultimately, we selected 200k images, including various styles of main product and poster images.

3.2 The Framework of UM-Text

As illustrated in Fig. 3, the main components of UM-Text include the UM-Encoder, the Diffusion Transformer, and training losses for optimization. Generally, UM-Designer is implemented as a VLM to capture the semantic information of instruction and reference image for the prediction of text content, layout, and implicit attributes. In addition to the instruction embedding from T5 and visual embedding of rendering text images, these predicted results are adaptively selected as additional conditions for downstream tasks according to the instruction, all of which constitute the conditions of diffusion model, named UM-Embedding c_e.

In the flow-matching-based diffusion model, we use VAE encoder to extract the latent representations of input image I_s, binary mask image I_m from UM-Designer or manually designed layout, and condition image $I_c = I_s \odot I_m$, resulting in z_0, z_m and z_c, respectively. Subsequently, the diffusion algorithm progressively adds random noise to z_0 at each time step t, resulting in a series of noisy latent variables z_t. Flow-based diffusion models employ a neural network V_θ to predict the velocity field at each time step, with the objective of matching the model's velocity field to the ideal velocity field that transports the data distribution along the diffusion process. This is achieved by minimizing the flow matching loss:

$$\mathcal{L}_{RF} = \mathbb{E}_{z_t, z_m, z_c, c_e, t \sim \mathcal{N}(0,1)}[||V^*(z_t, t) - V_\theta(z_t, z_m, z_c, c_e, t)||_2^2]. \quad (1)$$

where $\mathcal{L}_{RF}$ denotes the calibrated flow matching loss, and $V^*(z_t, t)$ is the target velocity field derived from the diffusion process.

3.3 UM-Encoder

Currently, many ControlNet-like approaches(Ma et al. 2024; Wang et al. 2025b; Zhao and Lian 2023; Chen et al. 2023a) typically inject glyph image and text conditions into the model. However, the glyph image condition is highly susceptible to the pre-defined text attributes, which often harm its robustness. Some methods replace the text embedding with line-level OCR embedding as text condition. However, these approaches have some limitations: (1) The visual embedding from OCR model only encodes the visual text information, missing the detailed description of generated image. (2) Line-level visual embedding is insufficient for the representation of character stroke information. (3) The layout and attributes of visual text are designed without considering the context information of reference image.

To address these issues, we propose the UM-Encoder for comprehensive condition representation on instruction, reference image, and character-level glyph image. Specifically, we use a pretrained VLM, known as UM-Designer, to capture the semantic information of instruction and reference image. The side information, including text content, layout, and attribute embeddings, for visual text generation/editing task is predicted by UM-Designer. To obtain fine-grained glyph information of text, we render the text content into glyph images in character-level, and use an OCR model to extract the visual embeddings of glyph images. Meanwhile, the output tokens of instruction and reference image are used as implicit attribute embeddings. We claim that those token embeddings are effective enough for implicit representation of text attributes due to the well-designed pre-training task of UM-Designer. More details can be found in Sec 3.5. Fi-

Figure 5: Qualitative comparison of UM-Text and state-of-the-art models in visual text editing task.

nally, the character-level visual embeddings, attribute embeddings, and instruction embeddings from T5 are aligned and concatenated into UM-Embedding as the condition embeddings of diffusion model.

3.4 Regional Consistency Loss

Previous text generation methods often face challenges in generating correct strokes for complex characters, due to the lack of detailed supervision in nuanced glyph shapes. To address this problem, we propose a Regional Consistency Loss (RC Loss) to constraint the structural consistency of visual text in various spaces. Specifically, RC Loss receives the mask image I_m from either the UM-Designer prediction result or manual annotation to localize the target regions, and calculate the L_2 distance between prediction result and ground-truth within target regions.

In our implementation, we design two types of RC Loss to constraint the structural consistency in both latent and RGB spaces. In the latent space, we calculate the RC Loss in the velocity field, which is analogous to the re-weighting strategy of flow matching loss. After that, we use VAE decoder to obtain the predicted image and use Canny edge detector to extract the edge maps of predicted image and input image. Therefore, the RC Loss can be simply calculated on the localized regions of both edge images. Formally, the RC Loss on latent and RGB spaces, denoted $\mathcal{L}_{RCL}$ and $\mathcal{L}_{RCI}$ respectively, can be formulated as:

$$\mathcal{L}_{RC} = \left\| \mathcal{C}(\hat{I} \odot I_m) - \mathcal{C}(I_s \odot I_m) \right\|_2^2 + \lambda \mathbb{E}\left[\| V^*(z_t, t) \odot z_m - V_\theta(z_t, z_m, z_c, c_e, t) \odot z_m \|_2^2 \right] \quad (2)$$

where $\hat{I}$ is the predicted image, and $\mathcal{C}(\cdot)$ denotes the Canny edge operator. The overall training loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{RF} + \beta \mathcal{L}_{RC}. \quad (3)$$

and λ, β are the hyper-parameters for balancing different losses. This dual-space regional consistency loss effectively preserves stroke integrity in complex character generation while maintaining stability in the editing process. Notably, the RC Loss in latent space mitigates the "dilution effect" commonly observed in mask-based editing, where the gradients outside the mask dominate the optimization direction.

3.5 Training Strategy

Based on the model architecture, we propose a progressive three-stage training strategy to learn a text editing model with context-aware designing capabilities. The training process is illustrated in Fig.4, including the pre-training of UM-Designer, pre-training of diffusion model, and semantic aligment between UM-Designer and diffusion model. The details are specified as follows.

Stage 1: UM-Designer Pre-training. In this stage, we initialize our UM-Designer by the weights of Qwen2.5-VL and continue the training process on UM-DATA-200K dataset. Generally, this dataset contains various tasks, including layout planning, text content generation, text detection and recognition. These tasks simulate the process of visual text generation and editing, and thus enhance the capability of UM-Designer for image-text understanding.

Stage 2: Diffusion Pre-training. In this stage, we initialize the text generation model with FLUX-Fill and train all parameters on public benchmark. This process enhances the foundational text generation capabilities for subsequent learning stage.

Stage 3: Semantic Alignment. In this stage, we train the connector of UM-Encoder and diffusion model to establish the connection between condition representation and application. By further introducing VLM embedding, our UM-Embedding complements the vision-language understanding in large-scale text generation task, thereby enhances the glyph consistency and aesthetics of generated image.

Overall, with the structural vision-language guidance, detailed visual condition, and a powerful training strategy on unified framework, our UM-Text significantly improves the

Methods	Task	English				Chinese			
		Sen.ACC ↑	NED ↑	FID ↓	LPIPS ↓	Sen.ACC ↑	NED ↑	FID ↓	LPIPS ↓
GlyphControl	T2I	0.5262	0.7529	43.10	-	0.0454	0.1017	49.51	-
AnyText		0.7239	0.8760	33.54	-	0.6923	0.8396	31.58	-
AnyText-2		0.8096	0.9184	33.32	-	0.7130	0.8516	27.94	-
FLUX-Fill	Editing	0.3093	0.4698	33.87	0.1582	0.0292	0.0625	29.93	0.1207
AnyText		0.6843	0.8588	21.59	0.1106	0.6476	0.8210	20.01	0.0943
AnyText-2		0.7915	0.9100	29.76	0.1734	0.7022	0.8420	26.52	0.1444
FLUX-Text		0.8175	0.9193	12.35	0.0674	0.7213	0.8555	12.41	0.0487
UM-Text		**0.8553**	**0.9395**	**10.15**	**0.0656**	**0.7988**	**0.8866**	**10.50**	**0.0481**

Table 1: Comparison on AnyText-benchmark dataset.

capability of model to generate high-fidelity and harmonious images in text editing task.

4 Experiments

4.1 Implementation Details

In stage 1 of training process, we initialize the weights of UM-Designer with Qwen2.5-VL-3B and train it on the UMDATA-200K dataset for 10 epochs using 16 Tesla A100 GPUs. In stage 2-3, we initialize the weights of the diffusion model with FLUX-Fill and train the model on AnyWord-3M dataset for 25 epochs. After that, we introduce the UM-Designer of stage 1 into our framework, and train the connector and diffusion model on AnyWord-3M dataset for 5 epochs using 16 Tesla A100 GPUs. The resolution of image is 512×512, and the resolution of rendering image for single character is 80×80. The strength coefficients λ and β are set to 5 and 2 by grid-search strategy, respectively.

4.2 Dataset and Evaluation Metrics

We use UMT-DATA-200K to train the UM-Designer model for layout and text design, and train UM-Text for visual text generation on AnyWord-3M (Tuo et al. 2023), which combines Wukong (Gu et al. 2022), LAION (Schuhmann et al. 2021), and OCR-specific datasets (3M images). To ensure a fair comparison, UMT-DATA-200K is **not used** for visual text generation training in our experiments.

We evaluate on several public benchmarks following prior work. AnyWord-Benchmark (Tuo et al. 2023) includes 1,000 English and 1,000 Chinese images. TextSeg (Xu et al. 2021) and LAION-OCR (Chen et al. 2023b) provide 1,024 and 9.1M real-world text images, respectively. ICDAR13 (Karatzas et al. 2013) contributes 233 test images for text detection evaluation. Following DREAMTEXT, we randomly select 100 images from the test sets of TextSeg, LAION-OCR, and ICDAR13 for evaluation.

AnyWord-Benchmark includes three evaluation metrics: Sentence Accuracy (Sen.ACC), Normalized Edit Distance (NED), and Frechet Inception Distance (FID) for distribution-level style similarity. We use Learned Perceptual Image Patch Similarity (LPIPS) to assess the consistency and realism of generated images, ensuring style consistency in edited regions while preserving non-target areas. All settings are consistent with FLUX-Text.

Following DREAMTEXT, we use an off-the-shelf scene text recognition (STR) model to identify the rendered text and then evaluate word-level correctness using sequence accuracy (SeqAcc) by comparing the STR result with the ground truth.

4.3 Experiment Result

Quantitative Results We comprehensively evaluate UM-Text and state-of-the-art methods using the AnyText-benchmark, UDiffText benchmark, and our self-constructed UMT-benchmark. As shown in Table 1, on the AnyText-benchmark, our method consistently outperforms competing approaches for both Chinese and English text across all metrics, including OCR accuracy (Sen.ACC, NED) and realism (FID, LPIPS). As shown in Table 2, our method outperforms previous approaches in SeqAcc and FID, although our LPIPS score is lower than DreamText's. This may be because our method produces colors and textures that better match the image style during text reconstruction.

Our UM-Designer model is capable of designing both layout and text, which motivates us to propose the UMT-benchmark to evaluate the performance of the entire pipeline. The UM-Designer model can also be integrated with other text editing models, such as AnyText and AnyText2, to generate product posters from clean product images. For fair comparison, our generative model, like previous state-of-the-art methods, is trained on the AnyWord-3M dataset without using any product-specific data. As shown in Table 3, we compare the Sen.ACC and NED metrics for both Chinese and English, and our method significantly outperforms previous approaches on both metrics.

Qualitative Results We conduct qualitative comparisons with state-of-the-art methods, including AnyText, AnyText-2, and FLUX-Text, on both English and Chinese multi-line text scenarios, as illustrated in Fig.5. Our method demonstrates superior performance in generating accurate, coherent, and visually harmonious text that blends seamlessly with the background, under complex conditions for both languages. In contrast, AnyText and AnyText-2 frequently produce results with blurred characters, duplicated text, or even incorrect glyphs. FLUX-Text generates text that is visually inconsistent with the background, suffers from color distortion, and also exhibits glyph errors, particularly in complex

Methods	Task	SeqAcc				FID	LPIPS
		ICDAR13(8ch)	ICDAR13	TextSeg	LAION-OCR		
AnyText	Recon	0.89	0.87	0.81	0.86	22.73	0.0651
UDiffText		0.94	0.91	0.93	0.90	15.79	0.0564
DreamText		0.95	0.94	0.96	0.93	12.13	**0.0328**
UM-Text		**0.99**	**0.98**	**0.97**	**0.96**	**6.57**	0.0479
AnyText	Editing	0.81	0.79	0.80	0.72	-	-
UDiffText		0.84	0.83	0.84	0.78	-	-
DreamText		0.87	0.89	0.91	0.88	-	-
UM-Text		**0.93**	**0.93**	**0.95**	**0.93**	-	-

Table 2: Comparison on the UDiffText benchmark dataset: The Recon task involves reconstructing text from the original image, while the Editing task focuses on modifying the text within the image.

Methods	English		Chinese	
	Sen.ACC	NED	Sen.ACC	NED
Flux-Kontext	0.325	0.502	-	-
Step1X-Edit	0.358	0.524	-	-
OmniGen2	0.371	0.541	-	-
AnyText	0.518	0.643	0.557	0.706
AnyText-2	0.693	0.723	0.720	0.806
UM-Text	**0.790**	**0.866**	**0.956**	**0.981**

Table 3: Comparison on UMT-benchmark. Please note that all methods use the layout and text by UM-Designer.

Figure 6: Compare UM-Text and ChatGPT4o in multi-turn image editing using natural language instructions, specifically in poster design, image editing, and image translation.

Chinese text scenarios. Notably, our method maintains precise glyph integrity and strong background consistency, even in challenging multi-line text settings. Meanwhile, we also conducte a multi-turn task comparison with ChatGPT-4o, as shown in Fig 6. Our method maintains precise glyph integrity and consistency with the background, even in complex multi-line scenarios, whereas ChatGPT-4o often introduces unnecessary text modifications.

4.4 Ablation Study

We randomly sampled 100k images from the AnyWord-3M dataset, including 50k Chinese and 50k English images. To evaluate the contribution of each module in our method, we conducte ablation studies on the AnyText-benchmark by training for 10 epochs on this small-scale dataset, as shown in Table 4. We used FLUX-Fill as the baseline, which demonstrates a lack of Chinese text generation capability. In addition, we compared the effect of adding a character-level visual encoder, which significantly improved the text generation ability of the baseline. We verified that the VLM embedding further enhanced the accuracy of text generation. Finally, we evaluated the impact of $\mathcal{L}_{RCL}$ and $\mathcal{L}_{RCI}$, which obtained an improvement of 4.8% and 4.2% respectively.

Methods	English		Chinese	
	Sen.ACC	NED	Sen.ACC	NED
Baseline	0.309	0.469	0.029	0.062
+Visual	0.759	0.887	0.676	0.839
+VLM	0.782	0.901	0.698	0.848
+RCL Loss	0.799	0.915	0.725	0.856
+RCI Loss	**0.824**	**0.925**	**0.746**	**0.863**

Table 4: Ablation experiments of UM-Text conducted on a subset of the AnyWord-3M dataset.

5 Conclusion

In this paper, we introduce UM-Text, a novel unified multimodal method designed to accomplish complex visual text editing tasks via simple natural language instructions. We explore a three-stage joint training strategy that integrates VLM and diffusion models, and propose the UM-Designer module for layout and text planning. Furthermore, we present the UM-Encoder, which fuses VLM embeddings, character-level visual embeddings, and T5 embeddings to enhance the model's understanding of both scene images and text glyphs, thereby enabling accurate and style-consistent editing and generation of textual and visual content. To supervise fine-grained visual text glyph information, we propose regional consistency loss. In addition, we contribute UM-DATA-200K, a large-scale and diverse dataset of layouts and texts, as well as the UMT-benchmark for evaluating instruction-based visual text editing. Extensive qualitative and quantitative results demonstrate the superiority of our approach.

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-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, H.; Xu, Z.; Gu, Z.; Li, Y.; Meng, C.; Zhu, H.; Wang, W.; et al. 2023a. Diffute: Universal text editing diffusion model. *Advances in Neural Information Processing Systems*, 36: 63062–63074.

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. 2024. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, 386–402. Springer.

Chen, J.; Xu, Z.; Pan, X.; Hu, Y.; Qin, C.; Goldstein, T.; Huang, L.; Zhou, T.; Xie, S.; Savarese, S.; et al. 2025a. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*.

Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; and Ruan, C. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Cui, C.; Sun, T.; Lin, M.; Gao, T.; Zhang, Y.; Liu, J.; Wang, X.; Zhang, Z.; Zhou, C.; Liu, H.; Zhang, Y.; Lv, W.; Huang, K.; Zhang, Y.; Zhang, J.; Zhang, J.; Liu, Y.; Yu, D.; and Ma, Y. 2025. PaddleOCR 3.0 Technical Report. arXiv:2507.05595.

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*.

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.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

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. 2024a. FLUX. https://huggingface.co/black-forest-labs/FLUX.1-dev.

Labs, B. F. 2024b. flux.1-fill. https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; et al. 2025a. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; et al. 2025b. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint 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*.

Liao, C.; Liu, L.; Wang, X.; Luo, Z.; Zhang, X.; Zhao, W.; Wu, J.; Li, L.; Tian, Z.; and Huang, W. 2025. Mogao: An omni foundation model for interleaved multi-modal generation. *arXiv preprint arXiv:2505.05472*.

Lin, B.; Li, Z.; Cheng, X.; Niu, Y.; Ye, Y.; He, X.; Yuan, S.; Yu, W.; Wang, S.; and Ge, Y. 2025. UniWorld: High-Resolution Semantic Encoders for Unified Visual Understanding and Generation.

Liu, R.; Garrette, D.; Saharia, C.; Chan, W.; Roberts, A.; Narang, S.; Blok, I.; Mical, R.; Norouzi, M.; and Constant, N. 2022. Character-aware models improve visual text rendering. *arXiv preprint arXiv:2212.10562*.

Liu, S.; Han, Y.; Xing, P.; Yin, F.; Wang, R.; Cheng, W.; Liao, J.; Wang, Y.; Fu, H.; Han, C.; Li, G.; Peng, Y.; Sun, Q.; Wu, J.; Cai, Y.; Ge, Z.; Ming, R.; Xia, L.; Zeng, X.; Zhu, Y.; Jiao, B.; Zhang, X.; Yu, G.; and Jiang, D. 2025. Step1X-Edit: A Practical Framework for General Image Editing. *arXiv preprint arXiv:2504.17761*.

Liu, Z.; Liang, W.; Liang, Z.; Luo, C.; Li, J.; Huang, G.; and Yuan, Y. 2024a. 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. 2024b. Glyph-byt5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *arXiv preprint arXiv:2406.10208*.

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. 2023. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870*.

Ma, L.; Yue, T.; Fu, P.; Zhong, Y.; Zhou, K.; Wei, X.; and Hu, J. 2024. CharGen: High Accurate Character-Level Visual Text Generation Model with MultiModal Encoder. *arXiv preprint arXiv:2412.17225*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Open AI. 2024. gpt40. Accessed: 2025-06-05.

Pan, X.; Shukla, S. N.; Singh, A.; Zhao, Z.; Mishra, S. K.; Wang, J.; Xu, Z.; Chen, J.; Li, K.; Juefei-Xu, F.; et al.

2025. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*.

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. *arXiv preprint arXiv:2307.01952*.

Ravi, N.; Gabeur, V.; Hu, Y. T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rdle, R.; Rolland, C.; and Gustafson, L. 2024. SAM 2: Segment Anything in Images and Videos.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37: 84839–84865.

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.-Y.; Geng, Y.; and Xie, X. 2023. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*.

Wang, Y.; Han, C.; Jin, Z.; Li, X.; Du, S.; Tao, W.; Yang, Y.; Yuan, C.; Lin, L.; et al. 2025a. UniGlyph: Unified Segmentation-Conditioned Diffusion for Precise Visual Text Synthesis. *arXiv preprint arXiv:2507.00992*.

Wang, Y.; Zhang, W.; Xu, H.; and Jin, C. 2025b. DreamText: High Fidelity Scene Text Synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28555–28563.

Wang, Z.; Bao, J.; Gu, S.; Chen, D.; Zhou, W.; and Li, H. 2025c. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20906–20915.

Wu, C.; Zheng, P.; Yan, R.; Xiao, S.; Luo, X.; Wang, Y.; Li, W.; Jiang, X.; Liu, Y.; Zhou, J.; Liu, Z.; Xia, Z.; Li, C.; Deng, H.; Wang, J.; Luo, K.; Zhang, B.; Lian, D.; Wang, X.; Wang, Z.; Huang, T.; and Liu, Z. 2025. OmniGen2: Exploration to Advanced Multimodal Generation. *arXiv preprint arXiv:2506.18871*.

Xu, X.; Zhang, Z.; Wang, Z.; Price, B.; Wang, Z.; and Shi, H. 2021. Rethinking text segmentation: A novel dataset and a text-specific refinement approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12045–12055.

Zhang, H.; Duan, Z.; Wang, X.; Zhao, Y.; Lu, W.; Di, Z.; Xu, Y.; Chen, Y.; and Zhang, Y. 2025a. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, X.; Guo, J.; Zhao, S.; Fu, M.; Duan, L.; Wang, G.-H.; Chen, Q.-G.; Xu, Z.; Luo, W.; and Zhang, K. 2025b. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*.

Zhao, Y.; and Lian, Z. 2023. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. *arXiv preprint arXiv:2312.04884*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *URL https://arxiv. org/abs/2504.10479*, 9.