UniGlyph: Unified Segmentation-Conditioned Diffusion for Precise Visual Text Synthesis

Supplementary Material

1. LLMs Layout Prediction Experiment.

Based on the complexity of spatial and style specifications, following [**?**], three task variants are defined to structure the text-to-image generation pipeline:

- **Input**: Caption describing the image (including text content to render) + image size. **Output**: Text strings paired with 4 absolute coordinates (polygon vertices) and free-form style attributes (font, color). **Characteristics**: Highest difficulty (pixel-level precision for irregular layouts) and maximum diversity in positioning and styling.
- **Input**: Caption describing the image (including text content to render) + image size. **Output**: Text strings paired with 2 absolute coordinates (axis-aligned rectangle: top-left and bottom-right) and free-form style attributes. **Characteristics**: Simplified spatial prediction (rectangular regions) with high style flexibility; suitable for rigid layouts.
- **Input**: Caption describing the image (including text content to render). **Output**: Text strings paired with 2 normalized coordinates (axis-aligned rectangle) and predefined style tokens. **Characteristics**: Easiest to implement (normalized coordinates + restricted styles) but sacrifices diversity in layout and design.

After evaluating the trade-offs between task complexity, generation consistency, and computational efficiency, the third variant was prioritized. This approach constrains the output space to normalized bounding coordinates and predefined style tokens, significantly reducing spatial ambiguity and aligning generated images more precisely with structured text prompts. The deterministic mapping between normalized coordinates (e.g., x_center,y_center, width, height) and layout semantics simplifies geometric reasoning while maintaining sufficient expressiveness for common axis-aligned scenarios.

For LLM implementation, comparative experiments were conducted across three candidate models: Qwen-2.5, Llama-3, and Baichuan-2. Qwen-2.5-7B demonstrated superior performance in structured output generation, particularly in token-to-coordinate alignment fidelity and style attribute grounding, as quantified by BLEU-4 (text accuracy) and IoU (layout consistency) metrics. A curated dataset of 1,000 high-precision annotated samples was used for fine-tuning, with prompt templates enforcing strict JSON schema compliance (e.g., "text": str, "bbox": [x0, y0, h, w], "font":`<font-cn-Heiti>`,"color":`<color-red>`). Training leveraged low-rank adaptation (LoRA) to preserve pretrained knowledge while adapting to coordinate regression and constrained style classification subtasks. This configuration achieved 89.3% exact-match accuracy on held-out test data, outperforming alternatives by 12% in cross-model benchmarking.

2. Data Processing for GlyphMM-3M and Poster-100K

To build a high-quality text-image dataset while filtering low-quality or irrelevant samples, a multi-stage processing pipeline is implemented. First, images are filtered based on their aspect ratios to retain only standard formats (1:1, 16:9, 9:16, 3:2, 2:3, 4:3, 3:4), with tolerances of ±5% applied to accommodate rounding errors. Next, bounding boxes containing text regions are analyzed: images are discarded if any text region occupies less than 2% or more than 95% of the total image area, as such extremes hinder learning meaningful image-text relationships. To eliminate subtitle-like content or screenshots, images with text regions near the edges (within 10% of the image boundary) are removed. Additionally, images with excessive text density—defined as having over 15 bounding boxes—are excluded, as they often represent structured documents (e.g., resumes, tables) lacking contextual interplay between visual and textual elements.

The remaining images are then scored by a ResNet-50-based aesthetic model pretrained on 50,000 human-annotated images rated for composition, contrast, and clarity. Images scoring below 0.4 (normalized to a 0–1 scale) are filtered out to preserve visually coherent samples. OCR validation is performed using PP-OCRv4[**?**], where images with an average text recognition confidence below 80% are discarded to ensure textual accuracy. For the final dataset, bilingual captions are generated: BLIP-2[**?**] (combining ViT-G, Q-Former, and OPT-2.7B) produces Chinese captions emphasizing contextual nuances, while CogVLM generates English captions with cross-modal attention to object relationships and scene dynamics. Redundant or nonsensical captions are further pruned using BERT-based classifiers.

The processed dataset includes images in standard formats, paired with metadata containing OCR-extracted text, bilingual captions, aesthetic scores, and resolution details. Validation metrics ensure diversity ($\geq$ 95% coverage of original resolution distribution), OCR-text alignment via human sampling, and caption relevance evaluated by CLIP-Score. This end-to-end pipeline prioritizes data integrity, visual-textual synergy, and robustness for multimodal learn-

ing tasks.

3. Model complexity and reusability

While our training uses a segmentation model and a layout generator, the runtime relies only on a single ControlNet, adding just 27% overhead (see Tab.,1), making it much simpler than prior multi-branch designs.

Table 1. Parameter breakdown for UniGlyph and baselines

Method	Backbone Params	Glyph-Module Count	Addl. Params (%)
GlyphControl	865 M	1	50.29
AnyText	859 M	3	66.66
GlyphDraw2	2.6 B	3	76.47
UniGlyph	11.9 B	1	27.74

4. Performance comparison of models trained solely on Anyword-3M

We compare our model, CharGen[1], and AnyText, all trained solely on Anyword-3M. The results show that our English performance mainly benefits from Anyword-3M. However, visual comparisons reveal that models trained only on Anyword-3M struggle to generate complex Chinese glyphs and multi-line layouts, likely due to the simplicity of the text in Anyword. Training with GlyphMM-3M and Poster-100K significantly improves these capabilities.

Table 2. Performance when trained *only* on Anyword-3M

Method	Chinese		English	
	Sen. Acc	NED	Sen. Acc	NED
AnyText-v1.1	0.6823	0.8423	0.6564	0.8685
CharGen	0.8096	0.9205	0.7499	0.8609
UniGlyph	**0.8102**	0.8783	**0.9014**	**0.9579**