SpaG-DiT: Enhancing Spatial Grounding for Diffusion Transformers

Zongliang Wu[1,2,3,†], Benlei Cui[2,†], Longtao Huang[2], Xiaoqian Xia[2], Yupeng Cao[2], Hui Xue[2], Tuo Chen[4], Haiwen Hong[2,‡], Xin Yuan[1,‡]

[1]Westlake University [2]Yuvion Team, Alibaba Group [3]Zhejiang University
[4]Southeast University

† Equal contribution. ‡ Corresponding authors.
hz_wzl@outlook.com

Abstract

Despite remarkable progress in text-to-image generation, current Diffusion Transformers (DiTs) inherently struggle with learning intricate visual content, such as accurate multilingual scene text rendering and novel visual concepts. This limitation stems from the inherent contradiction between vague semantic guidance provided by text and the objects of accurate reconstruction during diffusion training. To address this bottleneck, we propose **SpaG-DiT**, a novel spatial grounding training framework. SpaG-DiT directly injects explicit spatial priors into the DiTs before the attention module via Contextual Diagonal Position Encoding (CDPE), which dynamically modulates spatial coordinates without destroying the inherent sequential structure of the linguistic input. To evaluate the grounding ability, we implement our method on tasks that demand strong spatial grounding ability, including challenging text rendering and new concept learning, and establish two new datasets, **SpaG-DiT-MultiLingual** and **SpaG-DiT-Creature**, respectively. Extensive experiments demonstrate that SpaG-DiT achieves superior training efficiency and significantly outperforms vanilla self-finetuning on the above two tasks and the widely adopted GenEval benchmark. These results validate the plug-and-play nature of SpaG-DiT and its potential to enhance DiT training processes. Code repository: `https://github.com/Alibaba-VELLDEPTH/SPaG-DiT/`.

1 Introduction

Text-to-image (T2I) generation has witnessed remarkable progress driven by large-scale Diffusion Transformer (DiT) models (Peebles and Xie, 2023; Chen et al., 2024b), such as Stable Diffusion 3 (Esser et al., 2024) and FLUX.1 (Black Forest Labs et al., 2025). By formulating image synthesis as an iterative denoising process, these billion-parameter models map high-dimensional semantic conditions to visual content. Beyond text-to-image synthesis, diffusion models have also continuously advanced in sampling efficiency, controllability, and editing (Yin et al., 2024; Wu et al., 2026; Cui et al., 2026a; Liu et al., 2025), alongside emerging efforts on generation-quality prediction (Cui et al., 2026b). Despite their photorealistic generation capabilities, reliable and diverse scene text rendering—particularly in a multilingual context—remains a formidable challenge. Currently, the majority of leading open-source models predominantly support English and Chinese text generation, and their complete, large-scale training strategies for text rendering are rarely fully disclosed.

The open-source community has made numerous explorations to bridge this gap (Lu et al., 2026; Liu et al., 2024; Cui et al., 2026c). However, most of these approaches largely rely on explicit glyph guidance and manually specified bounding boxes during inference. Consequently, they fall short of a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: (a) Multilingual text rendering by our proposed method (Hindi, Arabic, Korean, and Bangla). (b) Semantic ambiguity in textual descriptions of spatial layouts.

pure text-to-image paradigm, remaining incapable of generating or being trained on complex text elements relying solely on textual prompts.

We argue that this limitation is fundamentally rooted in the fact that text-rendered images often feature complex spatial layouts, where relying solely on high-level prompt descriptions introduces significant semantic ambiguity, as shown in Fig. 1(b). This ambiguity directly prevents diffusion models from effectively learning unknown glyph structures. More importantly, this bottleneck is not exclusive to text generation; any training scenario involving images with complex layouts faces this exact contradiction: the severe mismatch between pixel-level spatial losses and vague semantic descriptions. Specifically, during standard DiT training, the only explicit spatial constraint is the Mean Squared Error (MSE) within the diffusion loss, which merely enforces a global, pixel-level alignment. Such vague semantic guidance is inherently insufficient to support effective learning in text-to-image training. To establish text token-to-image patch correspondence under this weak supervision, introducing explicit spatial priors becomes highly desirable. We observe that in current DiT paradigms—even those utilizing advanced positional encodings like 2D Rotary Position Embedding (RoPE) (Su et al., 2024; Heo et al., 2024; Ge et al., 2025) or Multimodal Scalable RoPE (MSRoPE) (Wu et al., 2025) for visual tokens—the positional encodings for text tokens inherently lack any spatial or geometric awareness. This absence presents an ideal, lightweight entry point for spatial intervention. Compared to computationally heavy injection mechanisms, such as employing explicit glyph representations, embedding spatial priors directly into the text positional encodings is a far more proper approach, because it leverages explicit spatial priors strictly as a training-time intervention: during inference it requires no spatial coordinates, thereby guaranteeing a pure text-to-image process without excessively disrupting the native inference capabilities of the base model. While such designs maintain the sequential regularity of the linguistic input and force the DiT to resolve text token-to-image patch alignment implicitly through attention, they still lack context-aware geometric anchoring. Directly learning this precise local alignment from an image rich in semantic content is extremely difficult, as it is severely entangled with, and interfered by, surrounding visual contents and complex textual contexts.

Therefore, we advocate for introducing explicit spatial priors during the diffusion training process to resolve this contradiction. By injecting this spatial prior information, the diffusion process can focus

entirely on learning the structures of new objects, without being hindered by semantic ambiguity. To this end, we propose **SpaG-DiT** (Fig. 1), a spatial grounding training framework that explicitly guides token-to-patch alignment in diffusion transformers. Instead of leaving spatial correspondence to be implicitly learned through diffusion loss, SpaG-DiT injects explicit geometric priors directly into the model's representation learning. The core of our training framework is Contextual Diagonal Position Encoding (CDPE), a modified positional mechanism for DiT self-attention. CDPE dynamically modulates token spatial coordinates using contextual grounding priors, providing strong, layout-aware geometric cues during denoising while strictly preserving the sequential structure of linguistic inputs. This design resolves the tension between spatial anchoring and sequential regularity, enabling the model to bind complex semantics to precise image regions without architectural modification.

To evaluate our training framework on learning new visual concepts that the base model has not mastered, we choose two challenging scenarios that remain unsolved by the base model and demand strong spatial grounding ability, as follows. First, we consider multilingual scene text rendering, where accurate generation requires both strict spatial layout control and precise glyph composition. Existing approaches often rely on extremely large-scale training data for each language, making them impractical for long-tail scripts such as Arabic, Korean, or Hindi. By explicitly modeling spatial grounding, SpaG-DiT enables diffusion models to reliably bind complex glyph structures to image coordinates, substantially reducing the reliance on massive datasets. Second, we evaluate new visual concept generation, where models must rapidly learn novel visual attributes and compose them within complex spatial layouts. This setting requires the model to efficiently bind newly learned semantic features to spatial representations, a process that strongly depends on spatial grounding capability. Both settings directly probe the training framework's hypothesis: explicit spatial grounding can reduce the complexity of learning spatial alignment. To support these trainings and evaluations, we further construct two datasets: SpaG-DiT-MultiLingual, a multilingual scene text rendering dataset covering five underrepresented languages, and SpaG-DiT-Creature, a dataset designed to evaluate new visual concept learning.

In summary, our contributions are as follows:

- i) A spatial grounding training framework for DiTs. We introduce **SpaG-DiT**, which injects explicit geometric priors into the denoising process via Contextual Diagonal Position Encoding (CDPE). CDPE dynamically modulates spatial coordinates, enabling precise token-to-patch alignment while preserving the sequential structure of linguistic inputs.
- ii) Targeted training and evaluation datasets. To explicitly validate the learning capabilities of the SpaG-DiT framework, we propose **SpaG-DiT-MultiLingual** and **SpaG-DiT-Creature** datasets. SpaG-DiT-MultiLingual is a new dataset and benchmark covering five underrepresented languages with standardized evaluation protocols, providing training and evaluation data for languages beyond English and Chinese. SpaG-DiT-Creature is to validate the learning ability of new visual concepts.
- iii) Data-efficient spatial alignment with strong empirical gains. SpaG-DiT significantly outperforms vanilla self-finetuning on both of the aforementioned datasets, requiring about 50k samples to achieve robust multilingual scene text rendering. Moreover, for a fair comparison, we also evaluate our method on the widely adopted GenEval (Ghosh et al., 2023) benchmark, where it similarly surpasses vanilla self-finetuning. This empirically proves the fundamental value of explicit spatial anchoring in diffusion processes.

2 Related Works

2.1 Multimodal Text-to-Image Generation Models

Recent advances in text-to-image generation have been significantly driven by the integration of vision-language models (VLMs) and large-scale training data. Modern approaches such as Qwen-Image (Wu et al., 2025) and Hunyuan-Image-3 (Cao et al., 2025) leverage billion-scale datasets combined with Reinforcement Learning from Human Feedback (RLHF) to achieve remarkable improvements in visual text generation capabilities. These methods share a similar architecture with unified understanding and generation models (Pan et al., 2025; Ma et al., 2025; Xiao et al., 2025; Tong et al., 2025), employing powerful VLMs with strong text comprehension abilities that enable them to understand complex textual instructions and generate high-quality images with accurate text rendering.

The integration of RLHF has proven particularly effective for Chinese and English text generation, where these models demonstrate exceptional performance due to the abundance of training data in these languages. For instance, Qwen-Image utilizes over a billion image-text pairs with RLHF fine-tuning, resulting in near-perfect character accuracy for English and Chinese scripts. Similarly, Hunyuan-Image-3 (Cao et al., 2025) employs a multi-stage training pipeline with billions of samples, incorporating human preference learning to refine text rendering quality. The VLMs in these systems exhibit strong multilingual text understanding abilities, accurately interpreting user prompts across various languages.

Despite their powerful VLM-based text comprehension, these models face significant limitations when generating text in languages beyond Latin and Chinese. The performance degradation for other language scripts—including Arabic, Japanese, Korean, and South Asian languages—remains substantial. This disparity stems primarily from the severe data imbalance in training corpora, where non-Latin and non-Chinese text constitutes less than 5% of billion-scale datasets. Consequently, while the VLM encoder can understand instructions in multiple languages, the generation quality for less-represented scripts remains poor, with frequent character distortions, incorrect glyph shapes, and spelling errors.

2.2 Visual Text Rendering in Diffusion Models

Generating and editing visual text in images remains challenging for diffusion-based models due to diverse fonts, complex backgrounds, and variable layouts. Recent methods have primarily focused on improving text generation by leveraging glyph conditions (Ma et al., 2023; Yang et al., 2023), segmentation masks (Chen et al., 2023; Tuo et al., 2024b), or character-aware text encoders (Zhao and Lian, 2024; Ma et al., 2024a), which provide additional visual priors and semantic information related to text rendering.

Large text encoders (e.g., T5 (Raffel et al., 2020)) and layout prediction (Chen et al., 2023, 2024a) further enhanced spelling accuracy and complex text placement, while OCR supervision and pre-rendered glyphs (Tuo et al., 2024b; Ma et al., 2024a) improved character fidelity. Many approaches (Tuo et al., 2024b,a; Ma et al., 2024b; Chen et al., 2023, 2024a; Zhang et al., 2024; Ma et al., 2023, 2024a) remain UNet-based, where the locality of convolution limits their ability to model global context and adapt flexibly to layout changes. These methods often rely on heavy auxiliary networks (e.g., ControlNet (Zhang et al., 2023)) and large labeled datasets. More recent specialist methods such as EasyText (Lu et al., 2026) and Glyph-byT5-v2 (Liu et al., 2024) are fundamentally designed around external reference inputs—glyph or layout reference images, or a specialized ByT5 glyph encoder—and therefore target a controllable, reference-guided generation setting rather than the pure text-to-image paradigm considered in this work.

2.3 Spatial Grounding and Layout Control in Diffusion Models

Beyond visual text rendering, precise spatial grounding is a fundamental challenge in general generation. Previous layout-to-image methods typically rely on adapter-based auxiliary networks (e.g., GLIGEN (Li et al., 2023), InstanceDiffusion (Wang et al., 2024)) or training-free cross-attention manipulation (e.g., BoxDiff (Xie et al., 2023)). However, these paradigms are predominantly designed for traditional UNet architectures. Extending them to modern Diffusion Transformers (DiTs) introduces significant computational overhead, architectural incompatibility, and often struggles to maintain strict boundaries for intricate structures like complex glyphs. Moreover, all such methods require auxiliary spatial or glyph inputs at inference time, whereas SpaG-DiT operates as a pure text-to-image model, which constitutes a practical deployment advantage independent of metric comparisons.

In contrast, our SpaG-DiT avoids appending heavy auxiliary modules or heuristically altering attention maps. Instead, it integrates spatial constraints directly into the DiT's representation learning by modulating positional encodings via CDPE. This provides explicit spatial anchoring from the beginning of attention computation, achieving superior spatial-semantic binding without disrupting the inherent sequential structures of the language tokens.

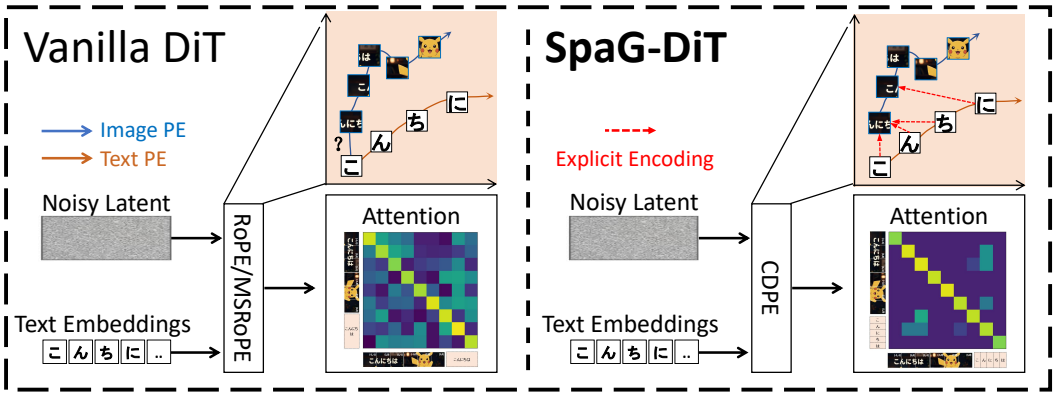


Figure 2: Comparison of vanilla DiT with our SpaG-DiT training framework.

2.4 Position Encoding in Diffusion Transformers

Rotary Position Embedding (RoPE) (Su et al., 2024; Heo et al., 2024; Ge et al., 2025) has become the standard for encoding sequential information in vision models, offering advantages in extrapolation and computational efficiency. Recent diffusion transformer models have extended RoPE to handle 2D spatial structures (Fu et al., 2024). Qwen-Image (Wu et al., 2025) proposes Multimodal Scalable RoPE (MSRoPE) that applies separate frequency encodings to different spatial dimensions, enabling the model to process images at various resolutions. However, MSRoPE treats all text positions uniformly with fixed diagonal encoding patterns, lacking the ability to adapt based on spatial differences. To the best of our knowledge, no existing positional encoding has been specifically designed to enhance the spatial grounding learning capability of DiTs.

2.5 Multilingual Scene Text Datasets

Existing text-to-image generation benchmarks such as GenEval (Ghosh et al., 2023), DPG-Bench (Hu et al., 2024), and OneIG-Bench (Chang et al., 2025) do not evaluate text rendering accuracy, and AnyWord (Tuo et al., 2024a) and GlyphDraw-3M (Ma et al., 2024a) focus primarily on English and Chinese text, providing limited training and evaluation for multilingual scenarios.

3 Method

The SpaG-DiT framework (Fig. 2) reformulates positional encoding in diffusion transformers so that explicit spatial priors can be injected without breaking the linguistic sequential structure. We first revisit two natural baselines, namely Offset Position Encoding (OPE) and Image Position Encoding (IPE), and then introduce our Contextual Diagonal Position Encoding (CDPE) that resolves the spatial–sequential trade-off. Throughout, $g_{h,w}(i, \mathbf{P}(i))$ denotes the 2D positional index assigned to the i-th text token given the spatial prior $\mathbf{P}(i) = (x_i^*, y_i^*)$. For a scalar z, $[z]$ denotes truncation toward zero, obtained by discarding its fractional part; $\lfloor z \rfloor$ denotes rounding down to the greatest integer not exceeding z.

Spatial priors use image-centered patch-grid coordinates, with the horizontal axis pointing right and the vertical axis pointing down. For a grid of $H \times W$ patches, the image boundaries span $[-W/2, W/2] \times [-H/2, H/2]$, and the geometric center is $(0, 0)$. Patch centers have coordinates $(w - (W+1)/2,\ h - (H+1)/2)$ for $w = 1, \ldots, W$ and $h = 1, \ldots, H$. The spatial priors are distinct from the final positional indices $g_{h,w}$.

Offset Position Encoding. Given a sequence of L text tokens with image-centered spatial priors $\mathbf{P} = \{(x_i^*, y_i^*)\}_{i=1}^{L}$, OPE perturbs the standard diagonal index of each token by a scaled, quantized spatial offset:

$$g_{h,w}^{\mathrm{OPE}}(i, \mathbf{P}(i)) = \Big(t_{\text{base}} + i + \big[\lambda_o^y\, y_i^*\big],\ t_{\text{base}} + i + \big[\lambda_o^x\, x_i^*\big]\Big), \tag{1}$$

where t_{base} denotes the base index offset of the text tokens within the multimodal sequence, and $\lambda_o^x, \lambda_o^y > 0$ convert patch-coordinate displacements into positional-index offsets. Larger λ yields larger spatial offsets for a fixed patch-grid displacement. OPE retains the sequential term $t_{\text{base}} + i$ as its reference diagonal.

Image Position Encoding. Conversely, Image Position Encoding (IPE) projects tokens directly onto the image-centered axes:

$$g_{h,w}^{\text{IPE}}(i, \mathbf{P}(i)) = \Big(c + \big[\lambda_{\text{img}}^y \, y_i^*\big], \; c + \big[\lambda_{\text{img}}^x \, x_i^*\big]\Big), \tag{2}$$

where $c = 0$ sets the image-center origin, and $\lambda_{\text{img}}^x, \lambda_{\text{img}}^y$ are scaling hyperparameters defined analogously to λ_o^x, λ_o^y. Tokens with zero spatial prior are assigned (c, c); other tokens receive the corresponding spatial offsets. Unlike OPE, IPE does not explicitly retain the token's sequential index i in its positional encoding.

Contextual Diagonal Position Encoding. CDPE combines a shifted sequential diagonal with image-centered spatial offsets. We define the diagonal shift as:

$$m = -t_{\text{base}} + \Big\lfloor \frac{L}{2} \Big\rfloor. \tag{3}$$

The priors are already expressed in image-centered patch coordinates, so we set $\hat{x}_i = x_i^*$ and $\hat{y}_i = y_i^*$ without further centering. CDPE composes the sequential index, the diagonal shift, and a quantized spatial perturbation:

$$\begin{aligned} g_{h,w}^{\text{CDPE}}(i, \mathbf{P}(i)) = \Big(&t_{\text{base}} + i + m + \big[\lambda_c^y \, \hat{y}_i\big], \\ &t_{\text{base}} + i + m + \big[\lambda_c^x \, \hat{x}_i\big]\Big), \end{aligned} \tag{4}$$

where $\lambda_c^x, \lambda_c^y > 0$ convert patch-coordinate displacements into positional-index offsets. Since $t_{\text{base}} + m = \lfloor L/2 \rfloor$, the reference index for token i is $(i + \lfloor L/2 \rfloor, \; i + \lfloor L/2 \rfloor)$. The spatial coordinates satisfy $\hat{x}_i \in [-W/2, W/2]$ and $\hat{y}_i \in [-H/2, H/2]$, so their quantized contributions are bounded for a fixed grid and fixed scaling factors. Truncation toward zero preserves the sign symmetry of the spatial offsets: $[-z] = -[z]$. At the image center, the spatial perturbation vanishes, leaving the shifted sequential diagonal. Spatial priors $\mathbf{P}$ are used exclusively during training. When they are absent, we use $(\hat{x}_i, \hat{y}_i) = (0, 0)$, which removes the spatial contribution in Eq. (4) and leaves the reference indices $(i + \lfloor L/2 \rfloor, \; i + \lfloor L/2 \rfloor)$. At inference, SpaG-DiT relies solely on the text prompt, without requiring user-provided bounding boxes or external layout planning.

4 Dataset and Benchmark

4.1 SpaG-DiT-MultiLingual Dataset

Our SpaG-DiT-MultiLingual dataset focuses on five languages: Hindi, Arabic, Japanese, Korean, and Bengali. We source images from MLT2019 (Nayef et al., 2019) and additional internet images. To ensure quality, we re-run OCR detection using EasyOCR (JaidedAI, 2024) and remove samples with large discrepancies from manual annotations or excessively small text regions. A no-reference image quality metric further filters out low-quality images, resulting in a final corpus of 50k high-quality images with reliable OCR labels. For training, we generate descriptive captions using Gemini 2.5-Flash (Comanici et al., 2025), providing semantic context alongside OCR annotations. The dataset is split about 95:5 per language into training and testing subsets.

For the evaluation process, we compute OCR accuracy metrics by comparing the recognized text from generated images against ground truth annotations. The OCR accuracy includes Sentence accuracy (Sen.ACC) and Normalized Edit Distance (NED). In addition to text accuracy, we evaluate image generation quality using standard image synthesis metrics, including Frechet Inception Distance (FID) (Heusel et al., 2017) and perceptual metrics LPIPS (Zhang et al., 2018).

4.2 SpaG-DiT-Creature Dataset

Our SpaG-DiT-Creature dataset is designed to evaluate spatial binding for novel concept generation. We construct the dataset through a multi-stage synthetic pipeline. First, we use Gemini 3

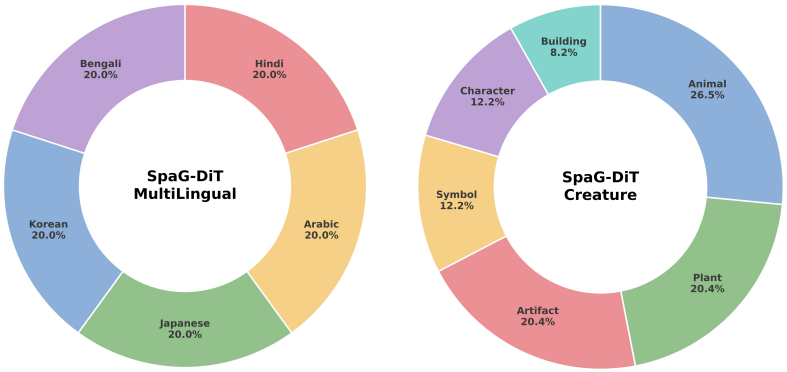


Figure 3: Composition of the proposed datasets.

Flash (Comanici et al., 2025) to generate 200 novel creature concepts along with detailed descriptions of their anatomical attributes. For each concept, a reference image is generated using Gemini 2.5 Flash Image (Comanici et al., 2025), providing a canonical visual prior of the creature. To introduce compositional diversity, Gemini 3 Flash (Comanici et al., 2025) further produces 40–50 prompts per creature, varying spatial arrangements of anatomical components.

Using these prompts and reference images, we synthesize text-image pairs through image-to-image generation, producing diverse spatial configurations while preserving creature identity. We then apply automated filtering to remove samples with structural artifacts, cluttered backgrounds, or low aesthetic quality, resulting in approximately 5,000 high-quality samples.

The dataset is split by creature concept to prevent identity leakage. From the test portion, we curate a representative evaluation set of 145 samples covering diverse spatial compositions. **Evaluation metric.** The percentage scores reported on SpaG-DiT-Creature (Table 3) denote a simple accuracy percentage over this 145-sample test set, similar to GenEval (Ghosh et al., 2023). They are computed via human evaluation: annotators independently judge whether each generated image correctly depicts the target creature concept (correct / incorrect), and we report the agreement rate (all correct) as the accuracy percentage.

4.3 General Text-to-Image Evaluation

For the general text-to-image generation task, our model is trained on the MS COCO 2017 dataset (Lin et al., 2014). To enhance the quality of the training data, the image prompts were re-captioned using Gemini 2.5-Flash (Comanici et al., 2025), and we incorporated the annotated bounding boxes to serve as spatial priors. To systematically assess the model's capabilities, we tested it on GenEval (Ghosh et al., 2023), a widely adopted objective benchmark specifically designed for evaluating text-to-image spatial alignment and object composition (e.g., rendering single objects and understanding positional relations). All models were evaluated at a resolution of 512×512 without employing any best-of-N selection strategy. **More details, along with additional quantitative and qualitative results, are provided in the Supplementary Material (SM).**

5 Experiments

5.1 Experimental Setup

Implementation Details. We train our model using the AdamW optimizer with a learning rate of 5×10^{-5}, batch size of 2, and LoRA rank of 256. Images are processed with a shortest side length of 512 using an aspect ratio boxing strategy. All experiments are conducted on NVIDIA H100 GPUs. For text rendering baseline comparisons, we evaluate the state-of-the-art FLUX.1 Kontext dev (Black Forest Labs et al., 2025) (abbreviated as FLUX.1 Kon. dev), Qwen-Image (Wu et al., 2025), AnyText2 (Tuo et al., 2024a), and a vanilla self-finetuned version (Qwen-Image-SFT) trained

Table 1: Evaluation results on SpaG-DiT-MultiLingual. **Bold** indicates the best result.

Methods	Sen.ACC(↑)	NED(↑)	FID(↓)	LPIPS(↓)
FLUX.1 Kon. dev (Black Forest Labs et al., 2025)	0.0000	0.0242	95.21	0.6512
TextDiffuser-2 (Chen et al., 2024a)	0.0412	0.1220	108.45	0.6650
AnyText2 (Tuo et al., 2024a)	0.0521	0.1850	98.12	0.6313
Qwen-Image (Wu et al., 2025)	0.0482	0.2150	97.41	0.6201
Qwen-Image-SFT	0.1054	0.2841	89.62	0.6023
SpaG-DiT	**0.1738**	**0.4326**	**87.64**	**0.5937**



Figure 4: Qualitative comparison across multiple languages. Our method consistently generates accurate text with proper character shapes and ligatures across diverse writing systems, while baseline methods fail to produce readable text in underrepresented scripts.

on the exact same data as SpaG-DiT. To ensure a fair comparison, all models are evaluated under a text-to-image paradigm during inference. None of the models are provided with explicit bounding boxes or spatial coordinates during testing.

5.2 Multilingual Scene Text Rendering

Quantitative Results. The mixed-language scenario presents a significantly challenging setting where models must handle five distinct writing systems simultaneously. As shown in Table 1, while billion-scale models like FLUX.1 Kon. dev and Qwen-Image nearly fail (achieving near-zero Sen.ACC), vanilla fine-tuning (Qwen-Image-SFT) provides limited improvement (0.1054 Sen.ACC). In contrast, SpaG-DiT achieves **0.1738 Sen.ACC and 0.4326 NED**, representing a 1.6× and 1.5× gain over Qwen-Image-SFT, respectively. This demonstrates that explicit spatial anchoring via CDPE fundamentally unlocks the model's ability to render complex text layouts.

Qualitative Analysis. Qualitative comparisons (Fig. 4) reveal SpaG-DiT's superior capability in capturing script-specific features. While baseline models generate malformed glyphs or ignore right-to-left directionality (e.g., in Arabic), SpaG-DiT seamlessly binds complex structural properties (like Devanagari conjuncts or Korean syllabic blocks) to precise image coordinates, maintaining high fidelity across all languages.

Table 2: Evaluation results on the GenEval benchmark. Models were evaluated at 512×512 resolution without best-of-N selection.

Method	Single Object	Two Object	Counting	Position	Attribute Binding	Overall
Qwen-Image	98.75	**93.94**	**87.50**	77.00	67.00	84.71
Qwen-Image-SFT	98.75	90.91	**87.50**	77.00	**70.00**	84.52
SpaG-DiT	**100.00**	**93.94**	**87.50**	**78.00**	69.00	**85.95**

Table 3: Quantitative results on the SpaG-DiT-Creature dataset across different categories. All values are human-evaluation accuracy (%) over the 145-sample test set.

Method	Animal	Plant	Symbol	Artifact	Character	Building	Average
Qwen-Image-SFT	68.5%	69.2%	68.8%	70.1%	68.9%	68.5%	69.0%
SpaG-DiT	**75.5%**	**76.2%**	**75.8%**	**77.0%**	**75.9%**	**75.1%**	**75.9%**

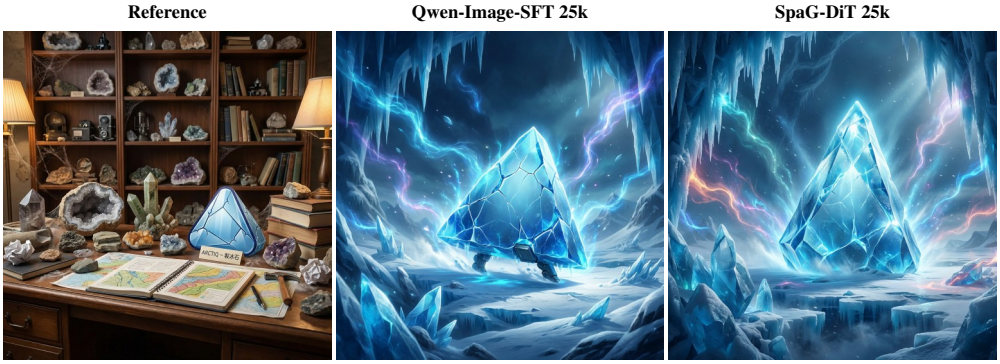


Figure 5: Qualitative comparison of vanilla SFT and SpaG-DiT on SpaG-DiT-Creature at 25K training steps. Additional examples across tasks and training checkpoints are provided in the Supplementary Material.

5.3 New Visual Concept Generation

To prove that SpaG-DiT is not merely an OCR-specific trick but a universal spatial grounding framework, we extend our evaluation to general object positioning and new concept learning.

As detailed in Table 3, SpaG-DiT demonstrates exceptional data efficiency in learning new creature characters. Compared to vanilla self-finetuning, which often struggles with identity leakage or layout collapse when placing new concepts, SpaG-DiT locks the new semantic tokens to the target regions.

Fig. 5 provides a qualitative comparison at 25K training steps: SpaG-DiT renders the new concept closer to the reference than vanilla SFT. Additional comparisons on multilingual text rendering and creature generation at 25K and 75K training steps are provided in the Supplementary Material.

5.4 General Image Generation

We further test our proposed method using GenEval. As reported in Table 2, vanilla SFT (Qwen-Image-SFT) does not improve the performance. However, tuning with SpaG-DiT on the same data, it achieves perfect scores in single object evaluation (100.00) and enhances spatial positioning accuracy (78.00). This confirms that CDPE acts as a stable geometric anchor, preventing the spatial drift typically observed during the denoising process of complex scenes.

5.5 Ablation Study and Learning Efficiency

We conduct ablation studies to validate each component of our framework. For experimental efficiency, ablation studies are performed solely on Korean text rendering.

Table 4: Ablation study on different modules. **Bold** indicates the best result.

Methods	OCR Pos.	MSRoPE	+IPE	+OPE	+CDPE	Sen.ACC(↑)	FID(↓)
Vanilla SFT	✗	✓	✗	✗	✗	0.4076	129.59
w/ IPE	✓	✗	✓	✗	✗	0.6461	125.12
w/ OPE	✓	✗	✗	✓	✗	0.6303	125.46
SpaG-DiT	✓	✗	✗	✗	✓	**0.6552**	**124.82**

Effectiveness of CDPE vs. Other PEs. As shown in Table 4, the vanilla SFT baseline (with standard MSRoPE) struggles (0.4076 Sen.ACC). While adding Image Position Encoding (+IPE) or Offset Position Encoding (+OPE) provides some improvements, our proposed CDPE (+CDPE) achieves the highest performance (0.6552 Sen.ACC) among the evaluated configurations.

Robustness to prior quality. CDPE applies spatial priors as bounded perturbations (Eqs. 1–4) rather than hard constraints, this design may help the model tolerate moderate noise in spatial priors. A quantitative analysis under noisier, automatically extracted priors is provided in the Supplementary Material.

Learning Efficiency. SpaG-DiT learns spatial features effectively from a limited training dataset. At matched training steps, Fig. 5 shows better preservation of the target concept than vanilla Qwen-Image-SFT. The Supplementary Material provides further qualitative comparisons on multilingual text rendering and creature generation at 25K and 75K training steps. These comparisons support more effective learning of the target visual structures under matched training settings.

6 Conclusion

In this paper, we have presented SpaG-DiT, a novel spatial grounding training framework that mitigates the spatial grounding bottleneck in Diffusion Transformers. By integrating CDPE, our method injects explicit spatial priors into position representations, requiring no spatial annotations at inference time and without disrupting linguistic sequential structures. Extensive evaluations across our newly established SpaG-DiT-MultiLingual and SpaG-DiT-Creature benchmarks, as well as GenEval (Ghosh et al., 2023), demonstrate the framework's promising spatial generalization capabilities. Furthermore, SpaG-DiT demonstrates exceptional data efficiency, requiring only a limited number of training samples to significantly outperform vanilla baselines, thereby proving the fundamental value of explicit spatial anchoring in diffusion models. Despite these advances, several limitations remain, including the rendering of intricate, low-resource scripts.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key R&D Program of China under Grant 2024YFF0505603; the Alibaba Group through Alibaba Research Intern Program; the National Natural Science Foundation of China under Grant 62271414; the Zhejiang Provincial Science Fund for Distinguished Young Scholar Project under Grant LR23F010001; the Zhejiang "Pioneer" and "Leading Goose" R&D Program under Grants 2024SDXHDX0006 and 2024C03182; the Key Project of Westlake Institute for Optoelectronics under Grant 2023GD007; and the Ningbo Science and Technology Bureau's "Science and Technology Yongjiang 2035" Key Technology Breakthrough Program under Grant 2024Z126.

Competing interests: The authors declare no competing interests.

References

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. OneIG-Bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025.

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

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Benlei Cui, Shaoxuan He, Bukun Huang, Zhizeng Ye, Yunyun Sun, Longtao Huang, Hui Xue, Yang Yang, Jingqun Tang, Zhou Zhao, et al. Tc-padé: Trajectory-consistent padé approximation for diffusion acceleration. *arXiv preprint arXiv:2603.02943*, 2026a.

Benlei Cui, Bukun Huang, Zhizeng Ye, Xuemei Dong, Tuo Chen, Hui Xue, Dingkang Yang, Longtao Huang, Jingqun Tang, and Haiwen Hong. Diffusion probe: Generated image result prediction using cnn probes. *arXiv preprint arXiv:2602.23783*, 2026b.

Benlei Cui, Fangao Zeng, Weitao Jiang, Yuwen Zhai, Haiwen Hong, Longtao Huang, Hui Xue, Wenxiang Shang, and Pipei Huang. simpleposter: a simple baseline for product poster generation. *arXiv preprint arXiv:2605.08784*, 2026c.

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

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. GeoWizard: Unleashing the diffusion priors for 3D geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.

Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. V2PE: Improving multimodal long-context capability of vision-language models with variable visual position encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21070–21084, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

JaidedAI. EasyOCR: Ready-to-use OCR with 80+ languages supported. `https://github.com/JaidedAI/EasyOCR`, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

Yi Liu, Hao Zhou, Benlei Cui, Wenxiang Shang, and Ran Lin. Erase diffusion: Empowering object removal through calibrating diffusion pathways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2418–2427, 2025.

Zeyu Liu, Weicong Liang, Yiming Zhao, Bohan Chen, Lin Liang, Lijuan Wang, Ji Li, and Yuhui Yuan. Glyph-byT5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *arXiv preprint arXiv:2406.10208*, 2024.

Runnan Lu, Yuxuan Zhang, Jiaming Liu, Haofan Wang, and Yiren Song. EasyText: Controllable diffusion transformer for multilingual text rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7565–7573, 2026.

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

Jian Ma, Yonglin Deng, Chen Chen, Nanyang Du, Haonan Lu, and Zhenyu Yang. GlyphDraw2: Automatic generation of complex glyph posters with diffusion models and large language models. *arXiv preprint arXiv:2407.02252*, 2024a.

Lichen Ma, Tiezhu Yue, Pei Fu, Yujie Zhong, Kai Zhou, Xiaoming Wei, and Jie Hu. CharGen: High accurate character-level visual text generation model with multimodal encoder. *arXiv preprint arXiv:2412.17225*, 2024b.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. JanusFlow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7739–7751, 2025.

Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. ICDAR 2019 robust reading challenge on multi-lingual scene text detection and recognition–RRC-MLT-2019. In *International Conference on Document Analysis and Recognition*, pages 1582–1587. IEEE, 2019.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with MetaQueries. *arXiv preprint arXiv:2504.06256*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. MetaMorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17001–17012, 2025.

Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. AnyText2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024a.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. AnyText: Multilingual visual text generation and editing. In *International Conference on Learning Representations*, 2024b.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. InstanceDiffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6057–6068, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Zongliang Wu, Siming Zheng, Peng-Tao Jiang, and Xin Yuan. Realism control one-step diffusion for real-world image super resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10906–10914, 2026.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2025.

Jinheng Xie, Yueqi Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Linlin Shen. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. GlyphControl: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36: 38441–38462, 2023.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024.

Lingjun Zhang, Xinyuan Chen, Yaohui Wang, Yue Lu, and Yu Qiao. Brush Your Text: Synthesize any scene text on images via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7175–7183, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

Yiming Zhao and Zhouhui Lian. UDiffText: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *European Conference on Computer Vision*, pages 217–233. Springer, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explicitly enumerate the main contributions.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Conclusion section explicitly discusses limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results requiring formal proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and the Supplementary Material report key training and inference details. Dataset construction pipelines are described in Section 4, and we will release code, models, and data after publication.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We plan to release code, models, and the datasets after publication, together with detailed instructions.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and the Supplementary Material report the optimizer (AdamW), learning rate (5×10^{-5}), batch size, LoRA rank, image preprocessing, and hardware.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following the common practice in large-scale text-to-image diffusion benchmarks (e.g., GenEval, AnyText), error bars are not reported because each training run of a billion-parameter DiT is computationally expensive, making multi-seed evaluation infeasible within our compute budget. The relative gains of SpaG-DiT over Qwen-Image-SFT are consistent and substantial across all metrics (Sen.ACC, NED, FID, LPIPS) and across two independent datasets, which provides strong evidence that the improvements are not due to random variation.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 5.1 reports experiments settings. Additional details are provided in the Supplementary Material.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and confirm that this research conforms to it in every respect.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positively, SpaG-DiT improves text-to-image generation for under-represented scripts (Hindi, Arabic, Japanese, Korean, Bengali), which can help bridge the digital-content gap for the long tail of world languages in design, education, and accessibility tools. Negatively, like any high-quality generative image model, the technology can be misused to produce misleading multilingual visual content (e.g., forged signage or documents). Mitigation strategies, including responsible release, anonymized data filtering, and recommended use of safety classifiers downstream, will accompany the eventual release of the models and datasets.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: For the SpaG-DiT-MultiLingual dataset, all collected images are filtered through a multi-stage pipeline that removes low-quality samples and cross-validates OCR labels with EasyOCR; images sourced from the internet are reviewed and filtered for unsafe content prior to release. The SpaG-DiT-Creature dataset is fully synthetic (generated by Gemini) and contains no real persons or sensitive content. Released models will follow gated access guidelines comparable to those used by the underlying base model (Qwen-Image).

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All third-party assets used in this paper are properly credited with citations. Each asset is used in accordance with its respective license and terms of service.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New datasets and the SpaG-DiT training framework are introduced. Their construction protocols, statistics, and intended evaluation procedures are documented in Section 4 and Section 5. The datasets and code will be released with detailed usage instructions.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: Table 3 reports small-scale human-evaluation results. Unpaid evaluation.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: Some results are evaluated by human volunteers. IRB approval is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Section 4 describes the use of Gemini for caption generation and synthetic creature concepts, reference images, and prompts. LLM use is part of data construction, not solely writing or formatting.