ATTRITEXTBENCH: FINE-GRAINED ATTRIBUTE EVALUATION FOR UNIFIED VISUAL TEXT GENERATION AND EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT



Figure 1: LeX-Bench focuses on sparse generation with a single attribute per sample, while OCRGenBench focuses on content editing. AttriTextBench evaluates dense multi-attribute generation and diverse attribute editing.

Recent state-of-the-art models have advanced rapidly in both text-to-image generation and instruction-based editing. Despite this progress, improving visual text synthesis remains a challenging and evolving frontier for image generation models. However, existing benchmarks primarily focus on content accuracy or rely on holistic whole-image evaluation. The former overlooks attribute control, while the latter cannot localize failures at individual text targets. We therefore introduce **AttriTextBench**, a fine-grained benchmark for attribute-level evaluation of visual text generation and editing. ***To unify both abilities***, AttriTextBench covers generation tasks and editing tasks such as text addition and modification across multiple rendering scenes. ***To cover text attributes***, we explicitly evaluate multiple attributes, such as font, size and weight. Specifically, we curate editing instructions from real user traffic and derive generation prompts from real images. Each target records its text content and applicable attributes. ***To support reliable evaluation***, we decompose each user prompt into atomic operations. For each operation, we resolve a target bounding box in which OCR measures content and a visual judge agent assesses each applicable attribute with a binary verdict. Meanwhile, masked background comparison evaluates background preservation for editing tasks. AttriTextBench contains 721 generation prompts with 10,803 text segments and 2,854 editing instructions with 3,449 operations. Experiments on 14 models show that AttriTextBench exposes fine-grained weaknesses in visual text control and provides actionable guidance for future model development.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Visual text benchmarks compared by capability, taxonomy and evaluation granularity. Atomic indicates that complex instructions are decomposed into atomic evaluation units. Region indicates the availability of precise, manually annotated bounding boxes for target text regions. ✓ denotes support, while ✗ denotes no support.

Benchmark	***Capability***		***Diagnosability***				
			Taxonomy			***Evaluation***	
	Gen.	Edit	Scene	Task	Attribute	Atomic	Region
OneIG-Bench Chang et al. (2026)	✓	✗	✗	1	✗	✗	✗
GlyphControl Yang et al. (2023)	✓	✗	✗	1	✗	✗	✗
CVTG-2K Du et al. (2025)	✓	✗	✗	1	3	✗	✗
LeX-Bench Zhao et al. (2025)	✓	✗	✗	1	3	✓	✗
TextAtlasEval Wang et al. (2025)	✓	✗	4	1	✗	✗	✗
AnyText-benchmark Tuo et al. (2024)	✓	✗	✗	1	✗	✗	✓
TextSculpt-Bench Lin et al. (2026)	✗	✓	2	4	✗	✗	✗
SkyReels-Text Yu et al. (2026)	✗	✓	✗	1	1	✗	✓
VTPBench Shu et al. (2025)	✓	✓	2	6	✗	✗	✗
TextEditBench Gui et al. (2025)	✗	✓	14	6	3	✗	✓
OCRGenBench Zhang et al. (2025)	✓	✓	5	13	✗	✗	✓
VDE-Bench Yi et al. (2026)	✗	✓	9	4	✗	✓	✓
AttriTextBench (Ours)	✓	✓	8	4	5	✓	✓

1 INTRODUCTION

Recent state-of-the-art image models have advanced rapidly in both text-to-image generation and instruction-based editing (Tuo et al., 2024; Shu et al., 2025). Despite this progress, improving visual text synthesis remains a challenging and evolving frontier for image generation models. Visual text carries names, instructions, prices and interface states that determine whether an image is usable. Even one local rendering error can therefore invalidate an otherwise plausible result. Accurate visual text synthesis is also essential in practical workflows that create an image from a prompt and refine it through natural language.

Leading diffusion and autoregressive models can already render short strings reliably, yet their ability to follow fine-grained instructions for text content and attributes remains limited. Early benchmarks evaluate content accuracy for short strings (Yang et al., 2023; Tuo et al., 2024). Later work extends this setting to longer and denser text (Wang et al., 2025; Chang et al., 2026), multiple text regions (Du et al., 2025; Zhao et al., 2025), and instruction-based editing (Lin et al., 2026; Gui et al., 2025; Yi et al., 2026). Recent benchmarks also cover both generation and editing (Shu et al., 2025; Zhang et al., 2025). However, current evaluation still has three limitations.

First, existing benchmarks lack diversity. Most evaluate either generation or editing, while even benchmarks covering both abilities lack comprehensive scene and task coverage (Shu et al., 2025; Zhang et al., 2025). Diverse scenes are essential for comprehensive evaluation because they expose models to different text layouts, visual contexts and rendering requirements. ***Second, existing benchmarks primarily assess content accuracy.*** Real-world text carries multiple attribute requirements, but current benchmarks usually emphasize whether the requested characters appear. Their attribute coverage remains limited to restricted subsets (Zhao et al., 2025), unverified text attributes (Du et al., 2025), or font-centered editing (Yu et al., 2026). This focus overlooks fine-grained attribute control. ***Third, existing benchmarks rely on holistic whole-image evaluation.*** Generation evaluation commonly combines whole-image OCR aggregates with *FID* or *CLIP similarity* (Tuo et al., 2024; Chang et al., 2026), while editing evaluation uses holistic multimodal ratings (Shu et al., 2025; Gui et al., 2025). These scores cannot identify which requirement failed at an individual text target. Moreover, the lack of edit-region bounding boxes makes local evaluation and background-preservation assessment difficult (Yi et al., 2026).

To address these limitations, we introduce **AttriTextBench**, an **Attri**bute-level **Bench**mark for visual **Text** generation and editing. We structure the benchmark around three design choices. ***We first provide broad coverage of both abilities.*** AttriTextBench spans generation and editing tasks such

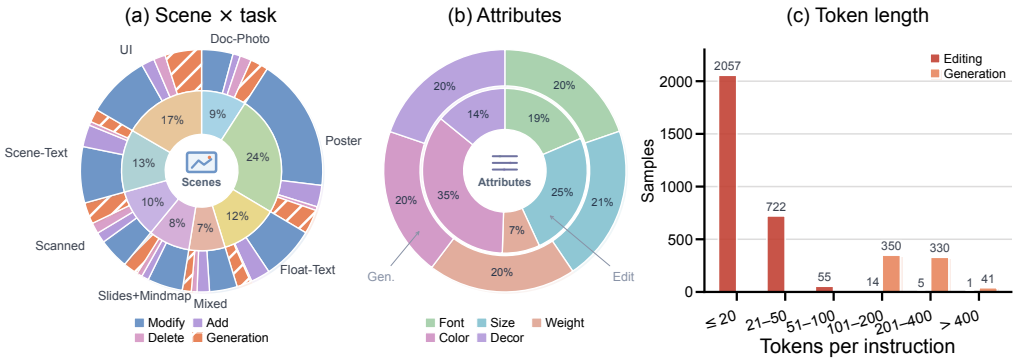

Figure 2: AttriTextBench covers diverse scenes, tasks and text attributes. Panels a, b and c show the taxonomy, five attributes and Qwen-Image2 token lengths.

as text addition and modification across multiple rendering scenes. ***We next evaluate multiple text attributes.*** These include font, size and weight. We curate editing instructions from real user traffic and derive generation prompts from real images. Each target records its text content and applicable attributes. ***We finally design an atomic evaluation protocol.*** We decompose each user prompt into atomic operations and resolve a target bounding box for every operation. Within each box, OCR measures content and a visual judge agent assesses each applicable attribute with a binary verdict. Meanwhile, masked background comparison evaluates background preservation for editing tasks.

AttriTextBench contains 721 generation prompts spanning 10,803 text segments and 2,854 editing instructions spanning 3,449 operations. We evaluate 14 current models on AttriTextBench. The results reveal a sharp gap between content and attribute control. We measure character-level correctness with *Accuracy* (AR) and sample-level attribute control with the *Attribute score* ($Attr$). The best generation *Accuracy* reaches 92.7%, whereas the best generation *Attribute score* reaches only 30.1%. Performance declines for text-dense images and multi-operation edits, while model rankings for editing correctness differ from those for background preservation. These findings show that AttriTextBench effectively exposes weaknesses in current generation models and provides guidance for improving fine-grained visual text generation.

The main contributions of this work are summarized below.

- We introduce AttriTextBench, a unified fine-grained benchmark for visual text generation and editing across eight real-world scenes and four tasks. It contains 721 generation prompts with 10,803 text segments and 2,854 real-user editing instructions with 3,449 atomic operations.
- We curate real-world data with annotations for text content and five style attributes, enabling diagnosis beyond content accuracy.
- We develop a region-grounded atomic protocol that evaluates content and applicable attributes for each operation. The protocol also measures background preservation for editing.
- We evaluate 14 current models and find a substantial gap between content accuracy and attribute control, with further degradation on text-dense images and multi-operation edits.

2 Related Work

2.1 Visual Text Generation and Editing Models

Current image models broadly fall into two categories. Specialized Generation Models integrate image synthesis and instruction-based editing, while Unified Und. & Gen. Models, namely unified understanding and generation models, additionally incorporate multimodal understanding. Qwen-Image, LongCat-Image, FLUX.1 Kontext, Seedream 4.0, JoyAI-Image and Boogu-Image are Specialized Generation Models that integrate image generation and image editing (Wu et al., 2025; Team et al., 2025; Labs et al., 2025; Seedream et al., 2025; Song et al., 2026; Chen et al., 2026). In parallel, Unified Und. & Gen. Models connect visual understanding with generation and editing. BAGEL, Show-o2, OmniGen2, BLIP3-o and InternVL-U represent this direction (Deng et al., 2025; Xie et al., 2026; Wu et al., 2026; Chen et al., 2025a; Tian et al., 2026). SenseNova-U1 further develops these

unified multimodal capabilities (Diao et al., 2026). Closed-source models such as Qwen-Image3 and GPT-Image-2 demonstrate strong general capabilities in image generation and image editing. We evaluate 14 current systems that support both interfaces required by AttriTextBench.

2.2 Benchmarks for Visual Text Generation and Editing

Most existing visual text benchmarks focus primarily on either text generation or text editing. Generation benchmarks test multilingual, multi-region or dense text through AnyText-benchmark, CVTG-2K, LeX-Bench, OneIG-Bench and TextAtlasEval (Tuo et al., 2024; Du et al., 2025; Zhao et al., 2025; Chang et al., 2026; Wang et al., 2025). Editing benchmarks instead cover localized replacement, font transfer and document editing through TextSculpt-Bench, SkyReels-Text, TextEditBench and VDE-Bench (Lin et al., 2026; Yu et al., 2026; Gui et al., 2025; Yi et al., 2026). VTPBench and OCRGenBench extend evaluation to both abilities (Shu et al., 2025; Zhang et al., 2025). However, they do not jointly provide a shared scene, task and attribute taxonomy, operation-by-attribute verdicts, and region-grounded scoring. Table 1 provides a structured comparison of these differences. AttriTextBench combines all three properties so that one fine-grained protocol can diagnose generation and editing instead of comparing them through separate aggregate scores.

3 AttriTextBench

AttriTextBench evaluates visual text *generation* and *editing* across eight scenes and five text attributes, as summarized in Fig. 2. We first derive generation prompts with explicit content and attribute requirements from real images in Sec. 3.1. We then construct the editing split from real image-instruction pairs and decompose each instruction into localized operations in Sec. 3.2. Both splits represent each text target as an *operation* that records its task, target content, attribute requirements and input region when available. This unified format supports the atomic evaluation protocol in Sec. 3.3, which resolves target regions and checks content, attributes and deletion where applicable. The protocol aggregates these local results into sample-level metrics and separately measures background preservation for editing.

3.1 Generation Data Construction

Visual text generation requires text prompts that specify the scene, text content and desired attributes. We construct these prompts from real-world images by extracting their textual content and describing the visual context. We further incorporate explicit attribute requirements to obtain generation prompts covering eight scenes and five text attributes.

An OCR system extracts visible text, bounding boxes and orientation from each image. We retain images that contain at least five segments and ten characters, remove illegible or branded content, and use a vision-language model to verify every scene assignment. The same model writes visual and textual descriptions, quotes the visible text, and corrects OCR noise using the image as evidence.

We assign one to three attributes to one or two segments, while every segment keeps content supervision. Scene quotas balance prompts with one, two, three or at least four attribute requests. We match font requests to the text language and restrict decoration to horizontal text in document-like scenes. We define size requests relative to another segment. The final generation split contains 721 prompts and 10,803 segments, of which 989 segments carry 1,897 attribute requests.

3.2 Editing Data Construction

Real-world image-instruction pairs provide realistic editing requests but vary substantially in image quality and text visibility. We therefore first screen for legible text and rank the retained images by aesthetic score to obtain suitable samples for annotation.

Text layouts and visual contexts also vary across scenes, creating distinct challenges for editing models. We therefore classify the retained samples into eight scene categories. This categorization enables evaluation across scenes to reveal model weaknesses and guide targeted improvement.

Finally, free-form instructions may combine multiple edits whose targets must be identified from the image. We jointly process each source image and its instruction with a vision-language model to de-

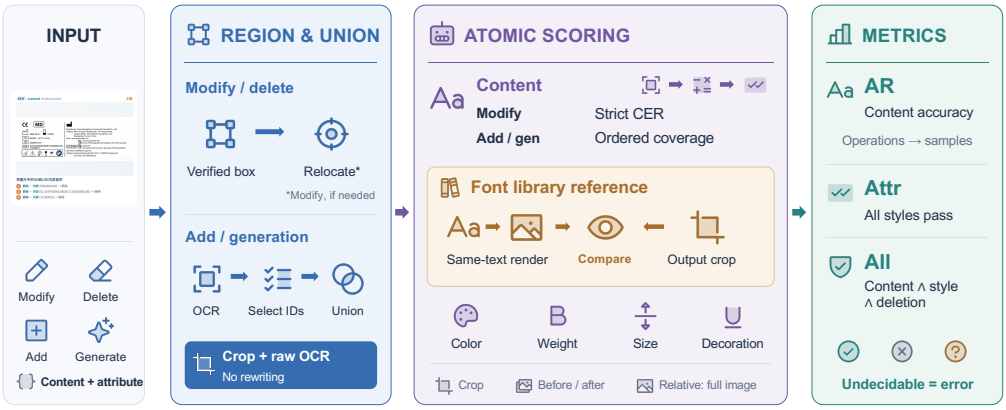

Figure 3: The atomic protocol scores region-level content, attributes and deletion before aggregating AR, $Attr$ and All. Background preservation is reported separately.

compose the request into atomic add, delete or modify operations. For delete and modify operations, the model also localizes the existing target text with bounding boxes in the source image. Each operation records its task and separates target content from attribute requirements. Automated checks reject inconsistent labels and unsupported layout requests, after which human reviewers verify the operations and bounding boxes against the image and instruction. The final editing split contains 2,854 instructions and 3,449 operations.

3.3 Atomic Evaluation Protocol

A single instruction-following score cannot reveal which requirement failed in a complex request. For example, rendering "hello world" in Comic Sans MS with black color and an underline involves separate font, color and decoration requirements. Our protocol in Fig. 3 therefore checks every atomic requirement independently, using OCR for text content and a visual judge agent for each attribute. We aggregate these checks into *Accuracy*, denoted AR, the *Attribute score*, denoted $Attr$, and All. AR measures character-level correctness, $Attr$ measures the proportion of samples whose attribute checks all pass, and All measures the proportion of samples that satisfy every content, deletion and attribute requirement.

Target localization. The visual judge agent struggles to discern fine-grained attributes in the full image, so every check starts from the complete target region. Enlarged or rewritten text can extend beyond its annotated box, so we refine the box using the input-output pair. Multi-line targets can also span several OCR lines, so a vision-language model selectively merges neighboring lines and regions. The agent then judges locally observable attributes on the resulting crop, while relational attributes such as relative size use the full image for context.

Content accuracy. We apply Unicode compatibility folding to the OCR text of each target region and remove whitespace while preserving case and punctuation. Content-modification operations use strict character error rate. Add and generation operations instead use ordered target-character coverage, because their recovered regions may include neighboring text. Let y_o and $\hat{y}_o$ denote the target and selected OCR text, D_o their Levenshtein distance, and m_o the target characters matched in order. We compute

$$CER_o^{\text{strict}} = \min\left(1, \frac{D_o}{|y_o|}\right) \qquad CER_o^{\text{cover}} = 1 - \frac{m_o}{|y_o|} \tag{1}$$

and define the operation-level *Accuracy* as $AR_o = 1 - CER_o$ under the applicable rule. A missing region yields an empty prediction and zero accuracy.

Attribute verdicts. Each attribute check returns an independent binary verdict and ignores differences unrelated to the requested attribute. Glyph shapes differ across characters, so we render the target content in the requested font from our font library. The agent then compares this same-content reference directly with the output crop. For unavailable fonts, the agent checks coarse properties from the textual description, such as serif, sans serif, handwriting and italic form. The agent reads

color and decoration from the target crop. For weight and size, it compares the text before and after a modification, while add and generation operations use the output directly.

Preservation and harmony. An edit should change only what the instruction requests. A modify operation therefore passes its attribute check only when every requested attribute is correct and every unrequested attribute stays unchanged. When an operation explicitly permits a font change, we drop size and weight from this preserved set, since a new font resets their natural scale. A pure-content addition requests no attribute, so the agent instead judges whether the new text harmonizes with the surrounding design. A composite addition scores only its requested attributes.

Deletion. A successful deletion must leave no trace of the original text. Delete operations therefore bypass CER and apply OCR to the original verified region. After whitespace removal, a deletion succeeds when valid regional OCR shares no character with the old text, while missing OCR evidence or old text makes it undecidable. We compute

$$Del = \frac{N_{\text{delete-correct}}}{N_{\text{delete}}} \tag{2}$$

and feed the same verdict into sample-level content correctness and All.

Aggregation. Attribute and deletion checks return correct, wrong or undecidable verdicts, and we count undecidable checks as errors while also tracking them separately. For AR, we average operation accuracy within each eligible sample and then across samples, excluding delete operations. Let $\mathcal{S}_{\text{content}}$ contain eligible samples, including undecidable cases, and let $\mathcal{O}_s$ contain their content operations. We compute

$$AR = \frac{1}{|\mathcal{S}_{\text{content}}|} \sum_{s \in \mathcal{S}_{\text{content}}} \frac{1}{|\mathcal{O}_s|} \sum_{o \in \mathcal{O}_s} AR_o. \tag{3}$$

$Attr$ conjoins the final attribute results of all add and modify operations in a sample. Samples without attribute checks, such as deletion-only edits, stay outside its denominator. All further requires every content, deletion and attribute request to pass. With N_{attr} eligible attribute samples and N evaluated samples, we compute

$$Attr = \frac{N_{\text{attr-correct}}}{N_{\text{attr}}} \qquad All = \frac{N_{\text{all-correct}}}{N}. \tag{4}$$

This conjunctive scoring ensures that a sample earns credit only when its complete request succeeds.

Background preservation. Editing should leave everything outside the target regions intact. We therefore mask all resolved operation regions and report $PSNR$ over the remaining pixels of the aligned input and output. Generation lacks pixel correspondence with an input and therefore receives no background score.

4 Experiments

We first describe the evaluation setting and then analyze overall performance, attribute-level behavior, coordination failures and qualitative examples.

We randomly sample 196 editing examples for attribute evaluation. Human annotators and three candidate Qwen visual judge agents assess the same requirements, assigning 1 if satisfied and 0 otherwise. Fig. 5a reports Pearson correlations between each agent's binary judgments and the human labels, with Qwen3.8-Max reaching $r = 0.581$. These results indicate a degree of positive correlation between automated evaluation and human judgments.

4.1 Experimental Setup

We evaluate current systems with both required interfaces. The evaluation includes 14 models that support text-to-image generation and instruction-based editing. The model set contains BAGEL, Boogu-Image, FLUX.2-Klein-9B, GLM-Image, GPT-Image-2, Illume+, InternVL-U, Janus-4o, JoyAI-Image, LongCat-Image, OmniGen2, Qwen-Image2, Qwen-Image3 and SenseNova-U1. Every model receives the same 721 generation prompts and 2,854 editing instructions. Upstream API issues and safety restrictions prevented GPT-Image-2 and Qwen-Image3 from returning a small

Table 2: Generation and editing results across scenes. AR measures content accuracy, $Attr$ measures attribute control and All requires full correctness. Green and orange bold values mark the best closed and open models. Upstream API issues and safety restrictions caused missing outputs for GPT-Image-2, Qwen-Image3 and Nano Banana. GPT-Image-2 lacked 4 generation and 89 editing outputs, while Qwen-Image3 lacked 8 and 21, respectively. We excluded Nano Banana from the statistics due to substantial missing outputs.

Generation

Model	Doc-Photo			Poster			Float-Text			Mixed			Slides+Mindmap			Scanned			Scene-Text			UI			*Overall*		
	AR↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑
Closed-Source Models																											
Specialized Generation Model																											
GPT-Image-2	**93.9**	**38.4**	**27.9**	**91.1**	**28.7**	**10.4**	**97.7**	**32.8**	**22.4**	**93.3**	**34.9**	**16.3**	93.9	25.0	11.1	**86.9**	**33.3**	**14.7**	**91.8**	**29.0**	**17.7**	**94.3**	25.3	**13.2**	**92.7**	**30.1**	**16.0**
Qwen-Image3	87.9	31.4	14.0	89.4	18.3	6.1	94.3	22.4	11.9	90.1	27.9	14.0	**94.5**	**30.6**	**20.8**	82.4	23.5	8.8	90.1	16.1	6.5	91.6	**25.9**	10.3	89.8	24.4	11.0
Open-Source Models																											
Specialized Generation Model																											
JoyAI-Image (Song et al., 2026)	**87.9**	25.6	9.3	90.4	9.6	3.5	**94.1**	14.9	4.5	**87.1**	9.3	**2.3**	88.7	16.7	5.6	80.5	14.7	0.0	**90.4**	17.7	6.5	**90.3**	12.1	3.4	**88.7**	14.7	4.2
Boogu-Image (Chen et al., 2026)	87.5	**30.2**	**10.5**	**91.8**	13.0	**7.0**	93.8	**23.9**	7.5	85.6	9.3	**2.3**	**92.6**	15.3	6.9	**82.5**	13.7	**3.9**	88.5	19.4	**9.7**	88.3	**17.2**	2.9	**88.7**	17.8	**6.0**
Qwen-Image2 (Zhao et al., 2026)	74.3	26.7	4.7	77.8	**18.3**	2.6	73.0	22.4	1.5	62.4	11.6	0.0	77.6	**23.6**	5.6	68.8	**18.6**	2.0	73.3	19.4	1.6	73.3	**17.2**	3.4	73.2	**19.7**	2.9
LongCat-Image (Team et al., 2025)	75.1	10.5	2.3	84.3	10.4	1.7	91.0	16.4	4.5	80.0	9.3	**2.3**	77.8	12.5	4.2	67.3	6.9	0.0	87.8	12.9	1.6	81.6	10.9	2.3	80.2	11.0	2.2
FLUX.2-Klein-9B (Labs, 2025)	47.0	16.3	0.0	40.4	11.3	0.0	50.1	20.9	0.0	42.2	**14.0**	0.0	29.6	13.9	1.4	37.5	9.8	0.0	44.8	14.5	0.0	44.2	8.6	0.6	42.0	12.6	0.3
GLM-Image (ai Team)	45.8	10.5	0.0	58.7	15.7	0.0	55.0	10.4	0.0	50.1	9.3	0.0	36.0	5.6	0.0	20.1	3.9	0.0	51.4	9.7	1.6	44.2	8.0	0.6	44.4	9.2	0.3
Unified Und. & Gen. Model																											
SenseNova-U1 (Diao et al., 2026)	86.3	20.9	3.5	87.5	**18.3**	5.2	91.0	13.4	**10.4**	**87.1**	**14.0**	**2.3**	90.4	**23.6**	**12.5**	78.5	10.8	1.0	89.5	**21.0**	6.5	88.8	13.8	**6.3**	87.2	16.5	5.8
BAGEL (Deng et al., 2025)	30.3	7.0	0.0	34.0	5.2	0.0	33.0	6.0	0.0	29.1	4.7	0.0	21.2	6.9	0.0	20.4	2.9	0.0	36.3	9.7	0.0	27.8	5.2	0.0	28.7	5.7	0.0
Janus-4o (Chen et al., 2025b)	1.9	0.0	0.0	2.9	0.0	0.0	2.5	0.0	0.0	1.8	0.0	0.0	1.9	1.4	0.0	2.1	0.0	0.0	2.7	0.0	0.0	2.3	0.0	0.0	2.3	0.1	0.0
InternVL-U (Tian et al., 2026)	28.2	3.5	0.0	31.7	1.7	0.0	51.8	10.4	0.0	43.3	2.3	0.0	49.6	8.3	0.0	46.7	7.8	0.0	61.9	9.7	0.0	58.1	10.9	0.6	46.7	7.2	0.1
OmniGen2 (Wu et al., 2026)	19.2	3.5	0.0	15.8	0.0	0.0	18.8	0.0	0.0	18.2	2.3	0.0	9.3	1.4	0.0	10.2	1.0	0.0	16.8	1.6	0.0	18.1	1.1	0.0	15.8	1.2	0.0
Illume+ (Huang et al., 2025)	0.2	0.0	0.0	0.5	0.0	0.0	0.2	0.0	0.0	0.5	0.0	0.0	0.6	0.0	0.0	0.6	0.0	0.0	0.4	0.0	0.0	0.5	0.0	0.0	0.5	0.0	0.0

Editing

Model	Doc-Photo			Poster			Float-Text			Mixed			Slides+Mindmap			Scanned			Scene-Text			UI			*Overall*			
	AR↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*AR*↑	*Attr*↑	*All*↑	*PSNR*↑
Closed-Source Models																												
Specialized Generation Model																												
GPT-Image-2	**78.2**	49.7	**46.2**	82.4	**71.5**	**53.3**	80.5	**68.5**	**53.4**	79.8	**68.6**	**57.8**	82.0	**64.9**	**51.8**	**80.6**	**62.1**	**54.7**	83.6	**65.7**	52.4	**82.5**	**56.6**	**54.5**	81.7	**65.2**	**53.1**	19.2
Qwen-Image3	77.1	**53.5**	43.8	**82.8**	63.9	50.1	**88.5**	66.5	51.1	**80.5**	52.9	41.7	**85.8**	58.4	48.7	80.3	58.5	47.4	**85.7**	62.2	**52.7**	80.5	53.3	47.6	**83.1**	60.1	48.7	**27.3**
Open-Source Models																												
Specialized Generation Model																												
JoyAI-Image (Song et al., 2026)	67.1	31.4	23.8	**82.8**	53.1	37.9	84.6	54.0	38.1	73.0	39.8	27.5	**73.9**	40.6	27.2	**70.5**	41.5	34.4	84.3	55.0	40.2	63.4	29.1	24.0	77.1	45.8	33.0	18.3
Boogu-Image (Chen et al., 2026)	66.9	32.4	**34.2**	**82.8**	**56.2**	**43.3**	**85.0**	**55.5**	**39.5**	**76.9**	**44.0**	**32.6**	69.0	37.1	30.3	67.0	40.0	**38.9**	**86.1**	**57.4**	**46.8**	71.2	**40.4**	**41.0**	**78.1**	**48.8**	**39.9**	21.3
Qwen-Image2 (Zhao et al., 2026)	**67.7**	**42.7**	33.3	81.6	51.5	40.2	80.4	50.9	38.6	69.8	39.8	29.8	71.6	**41.1**	**33.8**	67.8	**42.6**	34.0	78.5	54.7	43.0	**73.0**	40.1	36.2	76.1	47.3	37.4	**27.5**
LongCat-Image (Team et al., 2025)	66.7	21.1	19.2	80.7	37.6	29.3	82.2	38.7	30.4	72.1	35.6	26.1	66.4	30.2	23.7	69.0	33.8	24.7	81.4	46.4	38.9	67.4	25.5	25.0	75.5	35.1	28.2	22.1
FLUX.2-Klein-9B (Labs, 2025)	56.8	26.5	27.5	64.7	43.5	29.4	56.0	33.2	15.9	53.8	27.2	17.0	49.0	17.8	18.0	61.1	31.8	30.0	52.2	33.5	19.9	64.1	32.4	28.3	58.8	33.9	24.3	19.1
GLM-Image (ai Team)	30.1	10.3	13.8	63.4	17.7	13.6	58.7	17.3	10.5	52.4	9.4	14.7	17.7	3.0	8.8	21.1	12.3	23.1	60.1	24.9	15.9	22.9	4.7	12.9	46.7	14.2	13.9	14.4
Unified Und. & Gen. Model																												
SenseNova-U1 (Diao et al., 2026)	31.2	11.9	22.9	70.8	37.0	26.6	68.5	34.7	24.1	53.4	19.9	21.1	62.9	26.7	25.4	50.2	19.5	30.4	63.5	33.5	27.1	40.9	15.7	21.2	59.0	28.0	25.1	14.0
BAGEL (Deng et al., 2025)	50.8	21.1	21.2	53.8	25.8	19.1	56.3	23.4	13.9	53.8	20.9	16.5	42.9	15.8	15.8	54.6	27.2	26.3	48.9	26.3	19.2	60.1	23.6	24.0	53.3	23.9	19.6	26.8
Janus-4o (Chen et al., 2025b)	0.6	2.2	19.6	3.9	0.9	2.2	1.2	1.2	1.7	0.7	0.5	11.9	1.0	0.0	10.1	5.1	4.1	19.4	1.5	1.3	4.3	0.2	0.0	11.7	2.1	1.1	8.2	12.2
InternVL-U (Tian et al., 2026)	29.1	9.2	23.8	58.2	23.5	15.4	52.3	23.4	15.9	43.1	13.6	17.9	41.9	10.9	14.0	37.1	9.7	20.6	59.8	27.6	24.0	29.9	9.9	18.6	47.6	18.4	18.4	12.1
OmniGen2 (Wu et al., 2026)	40.7	17.8	21.2	34.1	16.9	13.3	23.3	9.5	6.5	29.2	8.4	6.9	27.6	6.9	11.4	35.9	13.8	20.6	23.0	9.4	8.2	39.8	10.7	14.3	31.6	12.4	12.6	18.0
Illume+ (Huang et al., 2025)	3.2	2.2	20.8	17.4	4.9	4.9	9.0	2.9	3.1	2.1	1.0	10.6	1.6	0.5	11.0	11.4	7.7	21.9	6.3	2.7	4.3	0.3	0.0	12.4	8.4	3.0	9.4	11.9

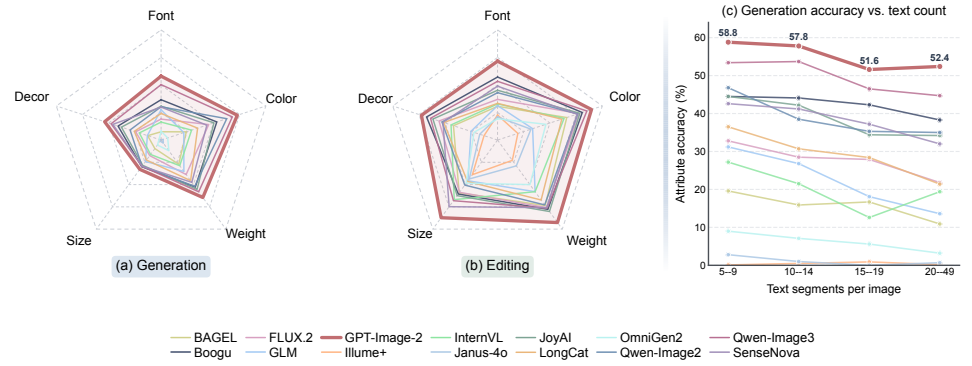

Figure 4: Panels a and b report attribute accuracy in generation and editing. Panel c reports generation attribute accuracy by text count.

number of outputs, as detailed in Table 2. We assess every returned output using the protocol in Sec. 3.3, with Qwen3.8-Max serving as the visual judge agent for the results in Table 2.

We report complementary measures of correctness and background preservation. Both abilities use AR, $Attr$ and All as defined in Sec. 3.3. AR preserves partial credit for recovered target characters, whereas $Attr$ requires every requested attribute, preserved attribute and harmony verdict to pass. All further requires every content, deletion and attribute request to pass, so deletion results enter All without a separate column. Editing additionally reports background $PSNR$ outside the resolved target regions. Undecidable checks remain in every relevant denominator. All reported font scores use the library-grounded protocol in Sec. 3.3, which compares the output with a same-content rendered reference whenever the requested font is available.

4.2 Main Results

Table 2 reports all three metrics for every scene and model. The results reveal several complementary strengths and weaknesses that no single aggregate score captures.

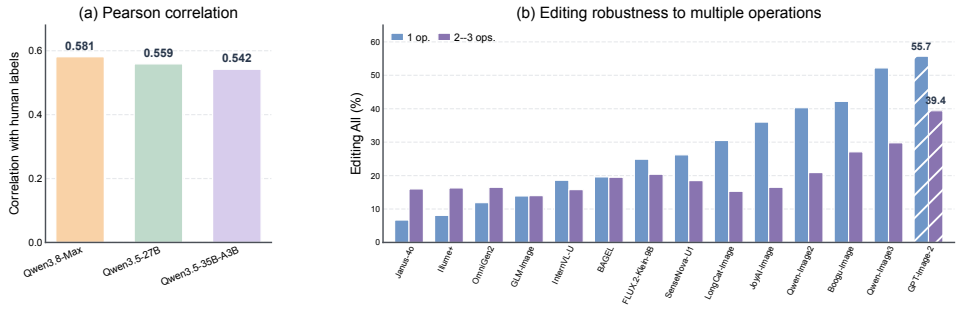

Figure 5: Panel a compares three Qwen visual judge agents with human judgments on 196 randomly sampled editing examples. Both assign binary labels to each attribute requirement, and Pearson correlation measures their association. Panel b reports editing All by operation count.

1. Accurate text content does not imply controllable typography. Table 2 shows that the strongest generation systems achieve high AR but remain much weaker on $Attr$ and All. This recurring gap indicates that models can recover the requested characters while failing to bind visual properties to the correct text instance. The sharper decline in $Attr$ and All further shows how atomic attribute errors accumulate into sample-level failure. Visual text evaluation should therefore separate content rendering from attribute control rather than use OCR accuracy as a proxy for both.

2. Generation and editing require different forms of visual text control. Table 2 shows that the two leading models stay the same across abilities, while the middle of the ranking shifts in attribute control. For example, SenseNova-U1 ranks fifth in generation $Attr$ but only eighth in editing $Attr$, whereas LongCat-Image rises from eighth to sixth. Generation requires models to synthesize text appearance from language, whereas editing adds localization and preservation constraints. Strong generation performance therefore does not guarantee precise conditional modification. Evaluating both abilities exposes failure modes that either one alone would hide.

3. Scene context changes the difficulty of the same nominal operation. Table 2 shows substantial performance variation across scenes, while the hardest and easiest scenes differ across systems. Text control therefore depends on perspective, background structure, local typography and neighboring visual elements. The scene taxonomy reveals where each model fails rather than merely partitioning an aggregate score.

4. Instruction success and image preservation expose a fundamental editing trade-off. Appendix Fig. 8 shows that the leaders in All, deletion success and background $PSNR$ are different models. A system may complete the requested text change through broad image regeneration, while another may preserve the background yet underperform on the requested operation. Consequently, a single editing score can reward one behavior while concealing failure on the other. Reporting task correctness together with background preservation distinguishes successful local editing from visually plausible repainting.

4.3 Attribute-Level Diagnosis

Atomic verdicts further reveal operation-level accuracy for each attribute. Fig. 4a and b show that GPT-Image-2 leads across both abilities and all five attributes. In generation, size is the weakest attribute for most models, while color or weight is the strongest for nearly all models. Editing is consistently stronger than generation across attributes. These patterns identify size control as a key generation bottleneck and reveal a broad gap between generation and editing attribute control.

4.4 Where Models Break Down

1. Text-dense images weaken attribute control. Fig. 4c shows that attribute accuracy declines for nearly all models as text density increases. The highlighted leader follows the same trend despite remaining strongest throughout. This broad decline shows that binding requested attributes to the correct text instances becomes harder as images contain denser text.

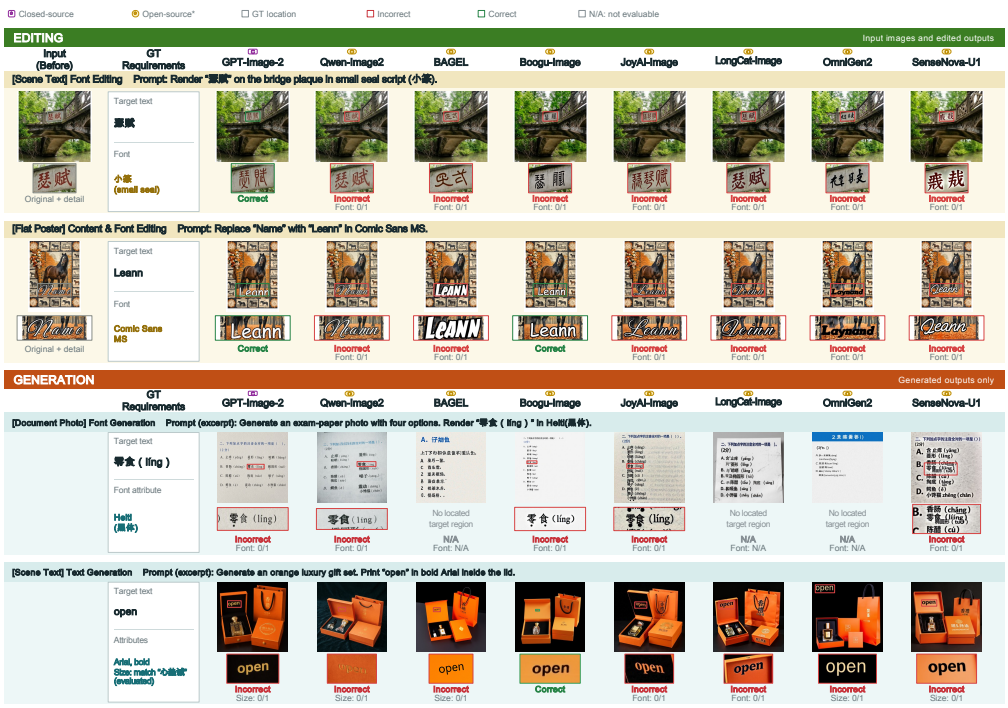

Figure 6: The upper panel shows editing, while the lower panel shows generation. Green and red boxes mark correct and incorrect checks, while unresolved regions remain errors.

2. Multi-operation edits generally reduce the rate of fully correct outputs. Fig. 5b shows that most systems lose editing All when requests expand from a single operation to multiple operations. The highlighted leader also declines substantially, suggesting that coordinating several edits remains challenging. Some models nevertheless score higher on multi-operation edits, partly because these edits include deletion operations. Models with weak text-rendering ability may fail to reproduce the original text and thus satisfy the deletion criterion, even when they also alter the surrounding image. These apparent gains should therefore be interpreted jointly with background $PSNR$ to assess both editing correctness and background preservation.

4.5 Qualitative Comparison

Fig. 6 compares representative outputs from selected systems. Editing failures affect individual attributes despite a recognizable surrounding image, while text-dense generation scenes preserve broad composition but miss requested attributes.

5 Conclusion

We presented AttriTextBench, a fine-grained benchmark that evaluates visual text *generation* and *editing* under one attribute-level protocol. A shared operation schema and scene taxonomy align the analysis of both abilities, while region-grounded OCR evidence and binary visual judgments isolate content and attribute failures. The evaluation of 14 current models reveals that leading systems recover target content more reliably than they satisfy complete attribute requests. Performance also declines for dense text and multi-operation edits, and editing correctness does not imply strong background preservation. By exposing these differences at the scene, task and attribute levels, AttriTextBench supports unified comparison and precise diagnosis of controllable visual text models.

AI use statement

We used large language models to assist benchmark data construction and polish the manuscript text.

Ethics statement

AttriTextBench evaluates visual text generation and editing using real images and instructions derived from real user traffic. We review all released samples to remove personally identifiable or harmful content and exclude material that cannot be redistributed. The benchmark may still reflect biases in its source data and visual evaluators, so results should be interpreted within its documented scope.

Reproducibility statement

We describe data construction, evaluation and experimental settings in Sec. 3, Sec. 3.3 and Sec. 4.1. Upon publication, we will release all redistributable data and annotations, prompts, evaluation code, inference configurations and evaluation outputs.

References

Z ai Team. Glm-image: Auto-regressive for dense-knowledge and high-fidelity image generation, january 2026. *URL https://z. ai/blog/glm-image*.

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. *Advances in Neural Information Processing Systems*, 38, 2026.

Guoxuan Chen, Chufeng Xiao, Haoran Yang, Siyue Xie, Binxiao Huang, Ming Zhang, Cheuk Him Chau, Xinyu Fu, Yingzhao Lian, Tom SY Li, et al. Boogu-image-0.1: Boosting open-source unified multimodal understanding and generation. *arXiv preprint arXiv:2607.13125*, 2026.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025b.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv e-prints*, pp. arXiv–2503, 2025.

Rui Gui, Yang Wan, Haochen Han, Dongxing Mao, Fangming Liu, Min Li, and Alex Jinpeng Wang. Texteditbench: Evaluating reasoning-aware text editing beyond rendering. *arXiv preprint arXiv:2512.16270*, 2025.

Runhui Huang, Chunwei Wang, Junwei Yang, Guansong Lu, Yunlong Yuan, Jianhua Han, Lu Hou, Wei Zhang, Lanqing Hong, Hengshuang Zhao, et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv preprint arXiv:2504.01934*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

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.

Yiheng Lin, Siyu Jiao, Xiaohan Lan, Wei Zhou, Qi She, Fei Yu, Heyun Chen, Zhengwei Wang, Jinghuan Chen, Moran Li, et al. Textsculptor: Training and benchmarking scene text editing. *arXiv preprint arXiv:2605.21090*, 2026.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Yan Shu, Weichao Zeng, Fangmin Zhao, Zeyu Chen, Zhenhang Li, Xiaomeng Yang, Yu Zhou, Paolo Rota, Xiang Bai, Lianwen Jin, et al. Visual text processing: A comprehensive review and unified evaluation. *arXiv preprint arXiv:2504.21682*, 2025.

Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

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*, volume 2024, pp. 56783–56799, 2024.

Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, et al. Textatlas5m: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*, 2025.

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.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

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

Hongzhu Yi, Yujia Yang, Yuanxiang Wang, Tong Li, Zhenyu Guan, Tianyu Zong, Jiahuan Chen, Chenxi Bao, Tiankun Yang, Haopeng Jin, et al. Vde bench: Evaluating the capability of image editing models to modify visual documents. *arXiv preprint arXiv:2602.00122*, 2026.

Yunjie Yu, Jingchen Wu, Junchen Zhu, Chunze Lin, and Guibin Chen. Skyreels-text: Fine-grained font-controllable text editing for poster design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14875–14884, 2026.

Peirong Zhang, Haowei Xu, Jiaxin Zhang, Xuhan Zheng, Guitao Xu, Yuyi Zhang, Junle Liu, Zhenhua Yang, Wei Zhou, and Lianwen Jin. Ocrgenbench: A comprehensive benchmark for evaluating ocr generative capabilities. *arXiv preprint arXiv:2507.15085*, 2025.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026.

Shitian Zhao, Qilong Wu, Xinyue Li, Bo Zhang, Ming Li, Qi Qin, Dongyang Liu, Kaipeng Zhang, Hongsheng Li, Yu Qiao, et al. Lex-art: Rethinking text generation via scalable high-quality data synthesis. *arXiv preprint arXiv:2503.21749*, 2025.

A Appendix

A.1 Data Construction Overview

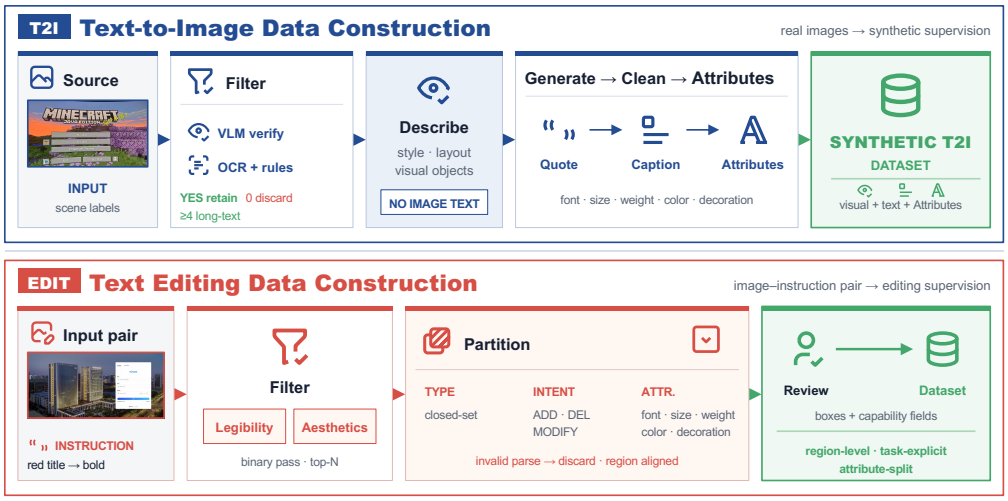

Figure 7: The pipeline derives generation prompts from real images and region-level editing supervision from image-instruction pairs.

A.2 Background Preservation

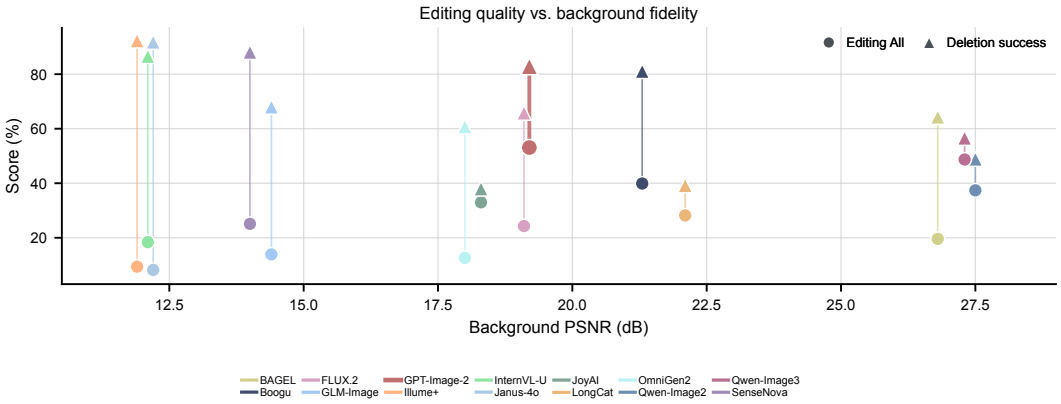

Figure 8: The plot compares editing correctness with background preservation. Circles report editing *All*, while triangles report deletion success.