TEXTREVISION: BENCHMARKING IMAGE TEXT EDITING TOWARDS INTENT FIDELITY WITH GATE-AND-CRAFT EVALUATION

Anonymous authors
Paper under double-blind review

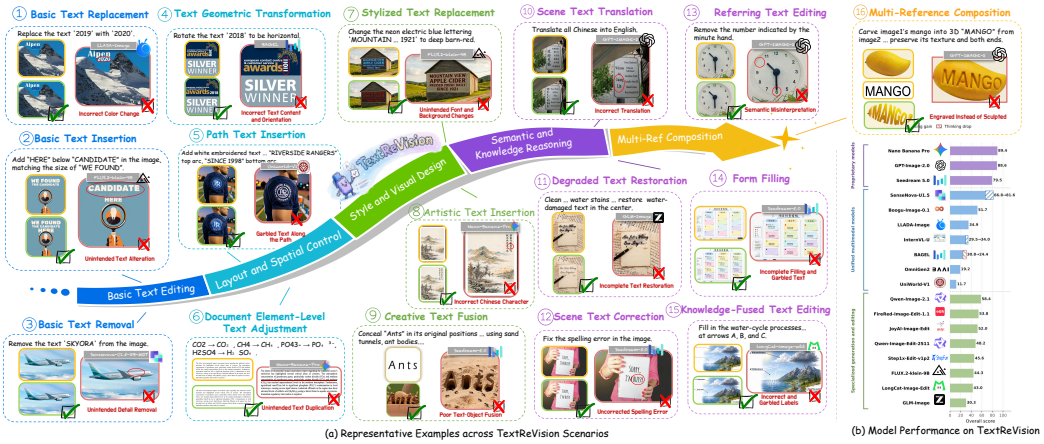


Figure 1: **Overview of TEXTREVISION. (a)** Examples from 16 scenarios in five families; orange, green, and red borders mark inputs, references, and outputs. **(b)** Mean of Chinese and English Overall scores for 18 models; hatching marks thinking-mode changes.

ABSTRACT

Reliable image text editing requires satisfying textual, spatial, stylistic, and semantic constraints while preserving non-target content. Existing benchmarks inadequately cover complex editing intents, while text recognition and image similarity alone cannot establish whether an instruction has been fulfilled. We present TEXTREVISION, a bilingual, scenario-driven benchmark comprising 1,200 human-curated samples across 16 scenarios and five capability families: basic editing, layout control, style design, semantic reasoning, and multi-reference composition. Sample-specific rubrics provide 6,426 verifiable criteria across eight evaluation dimensions. Furthermore, we propose *Gate-and-Craft* scoring, in which essential task completion bounds the attainable score and scales the contribution of execution quality. The scoring rule achieves a concordance correlation coefficient of 0.935 against human reference scores. Across 18 models and 21 configurations, the leading scenario-balanced scores reach 89.5/89.3 on Chinese/English samples, yet semantic reasoning remains a major weakness. SenseNova-U1.5-8B-MOT's official thinking mode raises its reasoning scores from 34.7/31.8 to 71.7/70.1, revealing that reasoning over user intent and translating it into actionable edits are critical. TEXTREVISION advances evaluation toward verifiable, intent-grounded image text editing.

1 INTRODUCTION

Image text editing modifies textual content and its visual presentation according to user instructions while preserving content and visual properties not targeted by the instruction. Bridging linguistic precision with visual coherence, this task underpins a broad range of applications, from product

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.

packaging and user interfaces to signage, presentation slides, and documents. Recent progress spans three major model families. *Proprietary systems*, including GPT-Image-2.0 (OpenAI, 2026b) and Nano Banana Pro (Google, 2025), have demonstrated increasingly strong instruction following and visual fidelity. *Open-source specialized editing models*, such as Qwen-Image-Edit (Qwen Team, Alibaba Group, 2025) and FireRed-Image-Edit-1.1 (Team et al., 2026), have further broadened access to high-quality text manipulation. Meanwhile, *unified multimodal generation models*, including BAGEL (Deng et al., 2025) and SenseNova-U1.5-8B-MOT (Diao et al., 2026), integrate visual understanding, reasoning, generation, and editing within a shared framework. Collectively, these advances have moved image text editing from constrained text rendering toward flexible, instruction-driven manipulation, calling for a corresponding shift in evaluation—from *isolated text replacement* to *faithful execution of diverse editing intents under complex visual constraints*.

Table 1: **Benchmarks with image text-editing tasks.** Coverage requires explicit tasks and separate evaluation. ∼: limited/related coverage or partial/proxy GT; –: insufficient reporting. Edited GT is human-created/verified; human validation compares automated scores with human judgments.

Benchmark	Scope	Data and Taxonomy					Explicit Capability Coverage				Evaluation Protocol		
		Size	Taxonomy	Zh/En Coverage	Human Curation	Edited GT	Layout	Style	Reason.	Multi-Ref.	Sample-Specific Rubric	Completion Gate	Human-Validated Evaluation
ScenePair	Scene text	1,280	1 basic operation	–	✓	∼	✗	✗	✗	✗	✗	✗	✗
MARIO-Eval-edit	Scene text	4,000	–	–	–	–	✗	✗	✗	✗	✗	✗	✗
TextEditBench	Text edit	1,196	12 fine-grained sub-tasks	✓	✓	∼	✓	✓	✓	✗	✗	✗	✓
OCRGenBench	Generation, edit, and OCR I2I	1,060	33 tasks (10 edit)	✓	✓	∼	∼	✓	✗	∼	✗	✗	✗
WeEdit	Text edit	4,000 (two splits)	8 operation categories	✓	–	–	✓	✓	✓	✗	✗	✗	✗
TEXTREVISION (Ours)	**Text edit**	**1,200**	**5 capabilities 16 scenarios**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Existing evaluations of image text editing remain fragmented and can be broadly grouped into four categories, with representative text-related benchmarks compared in Table 1 across task taxonomy, capability coverage, and evaluation protocol. **(i) Dedicated text-editing benchmarks.** ScenePair (Zeng et al., 2024) and MARIO-Eval-edit (Lan et al., 2025) primarily focus on basic operations such as text replacement, insertion, and removal, while often relying on localized text-region information or auxiliary conditions, such as quadrilateral regions, glyphs, or masks. This setup is increasingly misaligned with modern image editing models, which are designed to execute diverse editing intents from natural-language instructions alone. TextEditBench (Gui et al., 2025) further covers 12 sub-tasks, WeEdit (Zhang et al., 2026) organizes editing into eight operation categories, but this coarse taxonomy does not distinguish fine-grained user intents. **(ii) Broad OCR-generative benchmarks.** OCRGenBench (Zhang et al., 2025) unifies text generation, text editing, and OCR-oriented image restoration into 33 tasks, only 10 of which concern text editing, making it better suited to measuring general OCR-generative capability than diagnosing editing-specific skills. Its task-level metrics lack sample-specific criteria and allow visual similarity to compensate for failed edits. TextEditBench validates its automatic scores against human judgments but does not provide explicit sample-specific scoring rubrics. **(iii) General instruction-guided editing benchmarks.** MagicBrush (Zhang et al., 2023), Emu Edit (Sheynin et al., 2024), and ImgEdit (Ye et al., 2026) cover diverse visual transformations but treat text editing as only a limited subcategory. **(iv) Reasoning-aware editing benchmarks.** KRIS-Bench (Wu et al., 2026) introduces editing tasks that require knowledge and reasoning but offers only limited coverage of text editing and does not systematically examine text-specific requirements such as legibility, layout structure, and stylistic consistency. Overall, image text editing still lacks an expert-informed, multi-scenario benchmark centered on real-world editing intents.

To bridge this gap, we introduce **TEXTREVISION**, a benchmark combining a multi-scenario task suite with expert-informed evaluation, organized around *user intents* rather than low-level operations or text-carrier types. As shown in Figure 1(a), the suite contains *1,200* human-curated samples, evenly split between Chinese and English, spanning *16* real-world scenarios across five capability families: *basic editing*, *layout and spatial control*, *style and visual design*, *semantic and knowledge reasoning*, and *multi-reference composition*. Degraded text restoration, form filling, scene text translation, and multi-reference text design reflect practical user needs beyond basic text replacement.

To evaluate how faithfully models fulfill user requests across this task suite, we propose an expert-informed **Gate-and-Craft** protocol. Its design follows a simple principle: *essential task completion should bound the reward assigned to visual craftsmanship*. Conventional averaging treats correctness

and visual quality as compensatory, potentially favoring a polished but functionally incorrect output. Our protocol instead uses the *Gate* to measure essential task completion and to cap and scale the contribution of *Craft*, which measures execution quality. To accommodate diverse editing intents, we define eight general dimensions, including text accuracy, target localization, font accuracy, layout accuracy, visual naturalness, non-target preservation, semantic correctness, and multi-reference alignment. For each sample, we select the relevant dimensions, identify the *Gate* criterion, designate the appropriate dimensions as *Craft* criteria and construct a sample-specific rubric through expert annotation, specifying intent-aligned requirements for successful completion and execution quality. We validate the agreement of the automatic judge and scoring protocol with human ratings.

Using TEXTREVISION, we evaluate 18 leading proprietary and open-source systems, spanning specialized image editors and unified multimodal models. Table 3 reports overall scores, while Figure 2 compares representative capability profiles. Together, they reveal three key observations. **(i) Large performance gap.** Nano Banana Pro scores 89.5/89.3 on the Chinese/English subsets, whereas the open-source FLUX.2-klein-9B scores 40.3/48.2, revealing a substantial performance gap. **(ii) Capability imbalance.** Strong basic editing performance does not consistently extend to semantic reasoning, while multi-reference composition remains uneven across models, highlighting persistent challenges in handling complex editing intents. **(iii) Thinking can strengthen the understanding step before editing:** SenseNova-U1.5-8B-MOT's semantic reasoning scores rise from 34.7/31.8 to 71.7/70.1 on Chinese/English samples under thinking mode. *These findings expose a clear capability boundary: basic text editing proficiency does not yet translate into robust execution of complex user intents.*

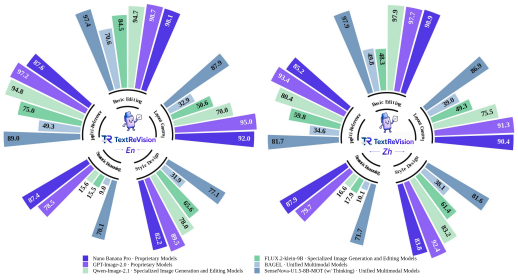


Figure 2: **Model performance on TEXTREVISION.** Scenario-balanced capability-level Total scores for six representative models on English (left) and Chinese (right).

In summary, our contributions are threefold:

- **An intent-driven, multi-scenario benchmark.** We introduce TEXTREVISION, comprising 1,200 curated bilingual samples across five capability families and 16 real-world scenarios, extending evaluation beyond basic text replacement.
- **A diagnostic and expert-informed evaluation protocol.** We propose *Gate-and-Craft* evaluation, which combines eight general dimensions with sample-specific rubrics to separate essential task completion from visual craftsmanship. We further assess both the scoring protocol and automatic judge against human ratings.
- **A comprehensive empirical study.** We evaluate 18 leading proprietary and open-source systems, revealing uneven capabilities, a clear performance gap, and persistent bottlenecks in layout control, semantic reasoning, and multi-reference composition. These results highlight that understanding and reasoning about user intent is just as critical as precise text rendering.

2 RELATED WORK

2.1 VISUAL TEXT GENERATION AND EDITING

Rendering legible, visually coherent text remains challenging in image synthesis. Methods including the TextDiffuser series (Chen et al., 2023; 2024), GlyphControl (Yang et al., 2023), the AnyText series (Tuo et al., 2024b;a), and Glyph-ByT5 (Liu et al., 2024), improve textual fidelity using character, glyph, and spatial guidance. Recent foundation models further support longer passages, multilingual content, and complex layouts, as demonstrated by Qwen-Image (Qwen Team, Alibaba Group, 2025). Image text editing additionally requires preserving non-target regions and following typography and background constraints. Early editors, including SRNet (Wu et al., 2019), MOSTEL (Qu et al., 2023), and TextCtrl (Zeng et al., 2024), rely on glyphs, quadrilateral regions, or masks, whereas FLUX.2-klein (Black Forest Labs, 2026) and FireRed-Image-Edit (Team et al., 2026) support

instruction-driven editing. Common OCR and visual metrics, such as edit distance, CLIPScore, and LPIPS, only partially capture each request's semantic, spatial, and stylistic requirements.

2.2 Instruction-Guided and Reasoning-Aware Image Editing

Instruction-guided image editing increasingly extends to tasks requiring reasoning. InstructPix2Pix (Brooks et al., 2023), MagicBrush (Zhang et al., 2023), and Emu Edit (Sheynin et al., 2024) enable edits from natural-language instructions without spatial annotations at inference time. Unified multimodal models, including BAGEL (Deng et al., 2025), OmniGen2 (Wu et al., 2025), InternVL-U (Tian et al., 2026), SenseNova-U1 (Diao et al., 2026), and JoyAI-Image (Song et al., 2026), further integrate understanding, reasoning, generation, and editing to support complex instructions and reference-conditioned tasks. KRIS-Bench (Wu et al., 2026) evaluates edits requiring knowledge and reasoning. Reliable transfer of these capabilities to text editing remains unclear.

2.3 Benchmarks and Evaluation for Image Text Editing

Image text-editing benchmarks vary in both task coverage and evaluation design. ScenePair (Zeng et al., 2024) focuses on localized edits, while TextEditBench (Gui et al., 2025) covers more diverse operations. For evaluation, OCRGenBench (Zhang et al., 2025) combines OCR and perceptual metrics, and TextEdit (Tian et al., 2026) uses completion gating, but neither provides explicit sample-specific rubrics. TextReVision combines intent-driven scenarios with such rubrics and Gate-and-Craft scoring to assess completion and execution quality (Table 1).

3 The TextReVision Benchmark

3.1 Design Principles

TextReVision evaluates image text editing as a practical, intent-driven task rather than a set of isolated low-level operations. It follows three principles. **(i) Intent-driven coverage.** Scenarios reflect real-world user requests and intended outcomes. **(ii) Capability-oriented organization.** Scenarios are grouped by the core capabilities they require. **(iii) Diagnostic assessment.** Sample-specific success criteria reveal strengths and failure modes beyond generic similarity scores.

3.2 Multi-Scenario Task Suite

TextReVision comprises five capability families and 16 editing scenarios, with balanced English and Chinese coverage (Figure 3(a)).

Basic Text Editing covers *text replacement*, *text insertion*, and *text removal*, testing target accuracy, background restoration, and preservation of non-target content.

Layout Control includes *text geometric transformation*, *path text insertion*, and *document element-level text adjustment*, testing geometric, positional, and layout constraints.

Style Design comprises *stylized text replacement*, *artistic text insertion*, and *creative text fusion*, testing typographic fidelity, legibility, and integration with scene appearance.

Semantic and Knowledge Reasoning covers *scene-text translation*, *degraded text restoration*, *scene-text correction*, *referring text editing*, *form filling*, and *knowledge-fused text editing*. These tasks require contextual interpretation and reasoning over linguistic, visual, or external knowledge to determine the intended edit.

Multi-Reference Composition covers *multi-reference design*, testing whether outputs combine content, layout, and style cues according to each reference's instructed role.

Median instruction lengths range from 22 characters (referring text editing) to 431.5 (form filling; Figure 3(b)). References span artistic lettering, perspective/spiral layouts, letters formed by objects, calligraphic overlays, and dense forms (Figure 3(c)). Text density is the image-area fraction covered by the union of PP-OCRv4 (PaddleOCR Team, 2023) detection polygons. Distributions by capability (Figure 3(d)) use boundary-corrected kernel density estimates with a shared bandwidth over 1,180 inputs, including multi-reference conditioning images and excluding 64 images without detections.



Figure 3: **Diversity analysis of TEXTREVISION. (a)** 1,200 samples across five capability families and 16 scenarios. **(b)** Instruction lengths by scenario (Unicode characters, log scale); boxes show interquartile ranges, lines medians, and whiskers the 5th–95th percentiles. **(c)** Diverse English and Chinese reference examples. **(d)** Text density by capability family: percentage of image area covered by PP-OCRv4 detections.

3.3 DATA COLLECTION AND CURATION

As illustrated in Figure 4, the construction of TEXTREVISION consists of three stages. First, 16 annotators collect source- image candidates from diverse text-bearing domains, while VLM-derived captions and image generators supplement scenarios with limited real-world data. After manual screening, annotators formulate scenario-specific instructions to construct source–instruction pairs. Reference edits are then obtained through two complementary pipelines: an open-resource pipeline based on available source images, glyphs, editing pairs, or image-editing software, and a closed-resource pipeline in which advanced image editors generate candidates for human selection.. Second, triplets are automatically screened for OCR consistency, readability, visual quality, layout consistency, and safety, then verified by humans. Finally, accepted triplets receive relevant evaluation attributes and sample-specific rubrics (Sec. 4). Source provenance, licensing, safety review, and release restrictions are detailed in Appendix A.

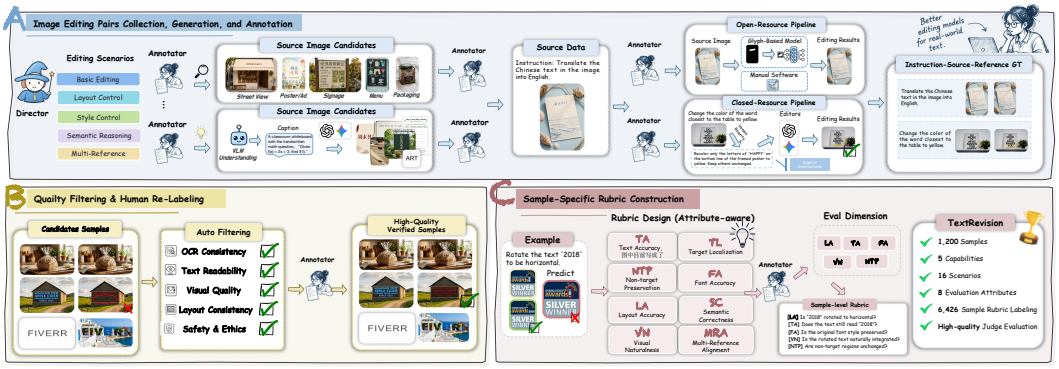


Figure 4: **Construction pipeline of TEXTREVISION. (A) Editing-pair construction.** Collected or generated images are paired with scenario-specific instructions and reference edits from open-resource or closed-resource pipelines. **(B) Quality control.** Triplets undergo automatic screening and human verification. **(C) Rubric construction.** Annotators build sample-specific rubrics using relevant criteria from eight evaluation dimensions introduced in Sec. 4.

4 EVALUATION PROTOCOL

4.1 GENERAL EVALUATION DIMENSIONS

Different editing intents impose distinct success criteria, making a uniform bundle of generic metrics insufficient. We therefore define eight reusable evaluation dimensions: (i) **Text Accuracy (TA):** whether the rendered text is correct and legible; (ii) **Target Localization (TL):** whether the edit accurately targets the intended region; (iii) **Font Accuracy (FA):** whether the edited text follows the source or requested typography; (iv) **Layout Accuracy (LA):** whether position, orientation and spacing satisfy the instruction; (v) **Visual Naturalness (VN):** whether the edit blends naturally into the

image; (vi) **Non-target Preservation (NTP):** whether regions outside the intended edit and necessary repair area remain free from unintended changes; (vii) **Semantic Correctness (SC):** whether the edited content is contextually and logically valid; and (viii) **Multi-Reference Alignment (MRA):** whether the output correctly integrates the required information from all references. These dimensions separate target selection, rendered-text integrity, semantic validity, final spatial arrangement, and visual quality to reduce duplicate penalties across criteria.

4.2 Sample-Specific Rubric Construction

Scenario-specific experts select applicable dimensions and write verifiable questions grounded in the instruction, source image, and human-verified reference edit (Figure 4(C)). The completion-critical criterion supplies the *Gate* score; at least one other criterion forms *Craft*, with scenario-specific weights (Appendix B.2). Replacement rubrics assess text, location, and typography; removal rubrics assess removal, background restoration, and non-target preservation. All criteria use $\mathcal{A} = \{0, 0.25, 0.50, 0.75, 1.00\}$, from unmet to fully satisfied (Appendix B.3).

4.3 Gate-and-Craft Scoring

For sample i, rubric criteria $\mathcal{R}_i$ are split into a Gate index k_i ($g_i = a_{ik_i} \in [0, 1]$, task completion) and Craft indices $\mathcal{C}_i = \mathcal{R}_i \setminus \{k_i\}$ with expert weights $w_{ij} > 0$. The Craft score $c_i \in [0, 1]$ (execution quality) and overall sample score s_i are defined as:

$$c_i = \frac{\sum_{j \in \mathcal{C}_i} w_{ij} a_{ij}}{\sum_{j \in \mathcal{C}_i} w_{ij}}, \qquad s_i = g_i \left[\beta + (1 - \beta) c_i\right], \tag{1}$$

where $\beta \in [0, 1]$ ($\beta = 0.5$ by default; reported as $100 s_i$). Under this design, failing the Gate ($g_i = 0$) completely voids the score ($s_i = 0$), while a satisfied Gate guarantees a base credit of β, with remaining points scaled by Craft quality c_i.

For each language subset, let $\mathcal{I}_j$ denote the set of samples in scenario j, and let $\mathcal{J}_\ell$ denote the set of scenarios in capability ℓ. With J scenarios, we compute the scenario-level, capability-level, and overall Total scores as.

$$S_j^{scn} = \frac{100}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} s_i, \qquad S_k^{cap} = \frac{1}{|\mathcal{J}_k|} \sum_{j \in \mathcal{J}_k} S_j^{scn}, \qquad S^{\text{overall}} = \frac{1}{J} \sum_{j=1}^{J} S_j^{cap}. \tag{2}$$

Gate scores are aggregated analogously, with g_i replacing s_i

4.4 Automatic Judging and Score Aggregation

To assess each edit, the VLM judge receives the instruction, source and output images, the human-verified reference edit, any conditioning references, and the sample-specific rubric. The reference illustrates a valid completion, while the instruction and rubric define the requirements. To inspect tiny text defects and subtle spatial shifts, the judge also uses enlarged crops and red/cyan overlays (Figure 5; Appendices D.2–D.3). Crops expose character differences; overlays locate potential changes, which the judge verifies against source and output. It records criterion-specific visual evidence, then an image-free request maps each reason to a scoring anchor. The resulting status labels map to ratings in $\mathcal{A}$, which Eq. 1 aggregates into s_i. Missing or invalid model outputs receive $g_i = s_i = 0$ and remain in scenario averages. Together, these steps help ensure that the judge's assessments are reliable and traceable.

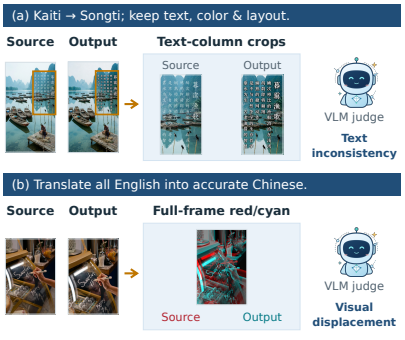


Figure 5: **Visual aids for automatic judging.**

5 EXPERIMENTS

5.1 SETUP

We evaluate 18 models in 21 configurations (Table 3) across three families: *proprietary*—GPT-Image-2.0 (OpenAI, 2026b), Nano Banana Pro (Google, 2025), and Seedream 5.0 (ByteDance Seed Team, 2026); *unified multimodal*—BAGEL (Deng et al., 2025), OmniGen2 (Wu et al., 2025), InternVL-U (Tian et al., 2026), UniWorld-V1 (Lin et al., 2025), SenseNova-U1.5-8B-MOT (Diao et al., 2026), Boogu-Image-0.1 (Chen et al., 2026b), and LLADA-Image (Chen et al., 2026a); *specialized editors*—Qwen-Image-2.1, Qwen-Image-Edit-2511 (Qwen Team, Alibaba Group, 2025), FireRed-Image-Edit-1.1 (Team et al., 2026), LongCat-Image-Edit (Team et al., 2025), GLM-Image (Z.AI, 2026), JoyAI-Image-Edit (Song et al., 2026), FLUX.2-klein-9B (Black Forest Labs, 2026), and Step1x-Edit-v1p2 (Yin et al., 2026). All models are evaluated under identical experimental settings. We also test official thinking modes for BAGEL, InternVL-U, and SenseNova-U1.5-8B-MOT.

5.2 RELIABILITY OF THE AUTOMATIC EVALUATOR

To assess automatic scoring reliability, we select 128 model-output units with bilingual coverage across all 16 scenarios. Two human evaluators with expertise in text editing and specialized training in the evaluation protocol score each output at the criterion level using the same sample-specific rubrics as the automatic judges. The human evaluation interface is provided in Appendix D.5. Their high inter-annotator agreement demonstrates the consistency and reliability of human judgments under our evaluation criteria. For judge comparison, each expert's criterion-level ratings are separately aggregated into Gate and overall scores, then averaged across the two experts to obtain the human reference scores. Then we measure agreement between automatic scores and human reference scores.

Aggregation rule. In Table 2, Panel A, we apply alternative aggregation rules to the criterion-level expert ratings and measure agreement between the resulting totals and human-aligned overall reference scores. Unweighted Average takes the arithmetic mean of all applicable criterion scores; Weighted Average uses their normalized rubric weights; Pure Product multiplies the Gate and Craft scores, $g \times c$. This comparison assesses how closely each rule aligns with the overall reference scores. We adopt $\beta = 0.5$, as Gate-and-Craft achieves the highest concordance correlation coefficient (CCC) of 0.935 among the settings in Table 2, the lowest mean absolute error (MAE) of 0.073, and the lowest root mean squared error (RMSE) of 0.129.

Table 2: **Reliability of automatic evaluation. Panel A:** Panel A: Agreement of Aggregation Rules with Overall Reference Scores. **Panel B:** Judge–human agreement (Gate/overall, $\beta = 0.5$). Dark/light blue: best/second per column (excluding Human–Human), ranked before rounding. †: selected configurations for evaluation.

Panel A: Aggregation Ablation

Aggregation Rule	Score Agreement		
	CCC ↑	MAE ↓	RMSE ↓
Unweighted Average	0.719	0.157	0.232
Weighted Average	0.870	0.116	0.167
Pure Product ($g \times c$)	0.920	0.092	0.142
Gate-and-Craft, $\beta = 0.3$	0.934	0.077	0.130
Gate-and-Craft, $\beta = 0.5$†	**0.935**	**0.073**	**0.129**
Gate-and-Craft, $\beta = 0.7$	0.930	0.084	0.135
Gate-and-Craft, $\beta = 0.9$	0.919	0.099	0.147

Panel B: Judge Score Agreement

Judge Model	Gate			Overall		
	CCC ↑	MAE ↓	RMSE ↓	CCC ↑	MAE ↓	RMSE ↓
Human–Human	0.970	0.031	0.094	0.969	0.044	0.092
Gemini-3.7-Flash†	0.902	0.086	0.166	0.905	0.100	0.161
Gemini-3.8-Flash	0.749	0.188	0.219	0.700	0.204	0.234
Claude-Opus-5	0.898	0.100	0.166	0.884	0.120	0.167
GPT-5.6-Sol	0.893	0.084	0.177	0.892	0.102	0.170
GPT-6-Astra	**0.916**	**0.078**	**0.160**	**0.912**	**0.090**	**0.155**
Qwen3.8-27B	0.747	0.145	0.282	0.784	0.161	0.256

Judge selection. Panel B evaluates judge–expert agreement using the selected aggregation rule, with CCC, MAE, and RMSE computed on 0–1 scores. GPT-6-Astra (OpenAI, 2026a) achieves the highest CCC values of 0.916 and 0.912 for Gate and overall scores, respectively. Gemini-3.7-Flash (Google, 2026) ranks second on both, with CCC values of 0.902 and 0.905, trailing by 0.014 and 0.007. Its corresponding MAEs are 0.086 and 0.100, and its RMSEs are 0.166 and 0.161. Given this strong agreement, we select Gemini-3.7-Flash to balance agreement with human ratings and cost, keeping the benchmark affordable for broader community use.

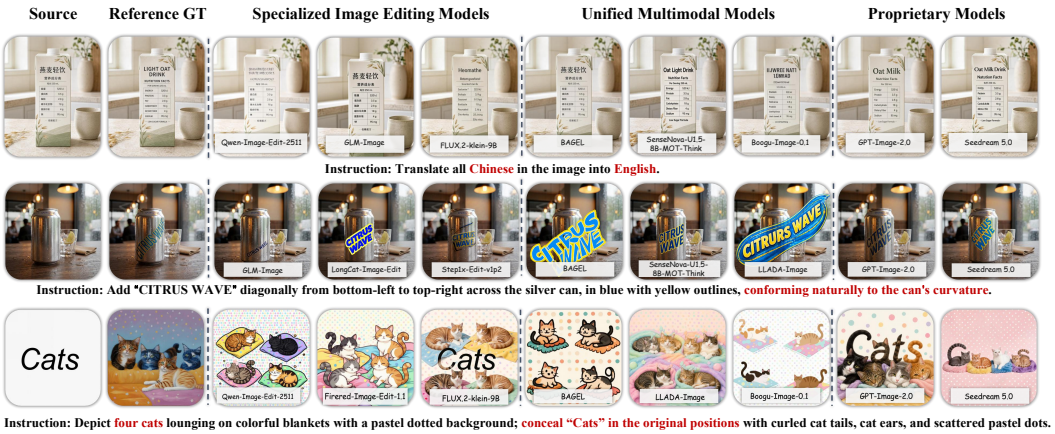


Figure 6: **Qualitative results of representative models.** Columns show the source, reference edit, and outputs from specialized image editing models, unified multimodal models, and proprietary models, separated by dashed lines. In the top row, the instruction translates all Chinese text on an oat-drink carton into English while preserving the nutrition panel and package layout; some outputs retain Chinese or render illegible small print. In the middle row, "CITRUS WAVE" must run diagonally across a silver can in blue with yellow outlines and follow the can's curvature; outputs differ in placement, scale, and surface alignment. In the bottom row, four cats lounge on colorful blankets against a pastel dotted background, with "Cats" concealed through curled tails, ears, and dots; some outputs omit or expose the intended hidden word. Red text marks the key instruction constraints.

5.3 Main Results

Overall. The leading proprietary models clearly outperform the strongest open configuration: Nano Banana Pro scores 89.5/89.3 and GPT-Image-2.0 scores 88.5/88.6, versus 82.0/81.1 for SenseNova-U1.5-8B-MOT with thinking. Yet even these leaders score lower on Semantic Reasoning than Basic Editing, revealing a remaining weakness. Figure 6 shows qualitative comparisons.

Table 3: **Main leaderboard on TextRevisionat Capability-level.** *Gate* and *Total* are reported for Chinese and English, with *Gate* denoting the mean gating score within each capability and *Total* combining *Gate* and *Craft* scores under $\beta = 0.5$; all scores are averaged at the scenario level. *w/ Thinking* denotes official thinking mode. Dark/light blue mark the best/second-best *Total* and *Overall* in each model family. Scenario-level results are in Appendix E (Table 8).

Model	#Params	Capability-Level Score																				Overall	
		Basic				Layout				Style				Reasoning				Multi-Ref.					
		Chinese		English		Chinese		English		Chinese		English		Chinese		English		Chinese		English		Zh	En
		Gate	Total	Gate	Total	Gate	Total	Gate	Total	Gate	Total	Gate	Total	Gate	Total	Gate	Total	Gate	Total	Gate	Total		
Proprietary Models																							
GPT-Image-2.0	–	98.5	97.7	99.7	**98.7**	92.2	**91.3**	96.0	**95.0**	93.2	**92.4**	90.0	**89.5**	82.0	79.7	80.4	78.5	95.0	**93.4**	98.3	**97.2**	88.5	88.6
Nano Banana Pro	–	99.7	**98.9**	99.2	98.1	92.6	90.4	93.1	92.0	84.9	83.8	83.8	82.2	89.4	**87.9**	88.6	**87.4**	90.0	85.2	91.7	87.6	**89.5**	**89.3**
Seedream 5.0	–	96.9	95.5	98.5	96.5	78.7	75.1	75.2	71.2	73.9	71.6	75.3	72.0	82.8	79.5	75.6	72.1	93.3	90.4	98.3	95.8	80.9	78.0
Unified Multimodal Models																							
BAGEL	7B+7B	57.4	49.8	82.5	70.6	48.0	39.0	39.0	32.9	45.1	38.1	37.1	31.9	12.8	10.1	10.5	9.0	46.7	34.6	60.0	49.3	29.7	31.8
BAGEL (*w/ Thinking*)	7B+7B	30.8	25.5	50.1	43.3	38.9	31.5	32.1	26.2	46.4	38.5	39.3	33.5	13.0	10.3	11.6	9.1	36.7	24.9	55.0	43.5	23.3	25.5
OmniGen2	3B+4B	26.0	21.4	59.0	48.6	20.9	15.8	27.9	23.9	29.0	22.6	34.3	27.7	8.8	6.5	8.9	7.0	31.7	20.9	41.7	31.1	15.0	23.3
InternVL-U	4B	88.4	74.4	84.0	68.4	36.5	29.3	38.2	32.6	43.1	35.2	48.3	40.4	12.8	9.2	10.4	8.0	0.0	0.0	0.0	0.0	29.5	29.5
InternVL-U (*w/ Thinking*)	4B	91.2	78.2	79.4	66.8	39.9	32.9	36.8	32.0	45.7	38.7	55.1	47.0	21.3	16.7	20.3	16.7	0.0	0.0	0.0	0.0	34.3	33.6
UniWorld-V1	7B+12B	2.8	2.2	21.7	17.4	12.3	8.4	27.9	22.5	7.6	6.3	36.0	29.8	9.2	5.9	12.6	9.8	11.7	7.2	18.3	13.7	5.8	17.6
SenseNova-U1.5-8B-MOT	8B+8B	98.8	97.1	98.8	96.6	88.5	85.8	89.3	86.5	82.1	80.3	80.0	**77.8**	38.8	34.7	34.4	31.8	85.0	78.7	90.0	85.2	67.3	66.2
SenseNova-U1.5-8B-MOT (*w/ Thinking*)	8B+8B	99.7	97.9	99.0	97.4	89.7	86.9	90.5	87.9	83.1	81.6	79.2	77.1	75.7	71.7	72.8	70.1	88.3	81.7	93.3	89.0	82.0	81.1
Boogu-Image-0.1	10B	97.2	95.2	95.4	91.9	73.5	69.2	67.2	64.4	80.1	76.8	73.8	70.7	26.8	23.3	20.4	18.1	0.0	0.0	0.0	0.0	54.0	49.4
LLADA-Image	6B	78.7	70.4	84.2	72.1	41.8	36.6	42.0	38.1	64.0	57.3	67.6	61.7	12.1	8.9	11.1	8.9	0.0	0.0	0.0	0.0	34.2	35.6
Specialized Image Generation and Editing Models																							
Qwen-Image-2.1	7B	98.8	**97.9**	96.7	**94.7**	77.8	**75.5**	72.3	**70.0**	84.4	**83.2**	80.1	**78.0**	19.1	16.6	17.3	15.6	85.0	**80.4**	96.7	**94.8**	**59.4**	**57.3**
Qwen-Image-Edit-2511	20B	95.5	89.4	94.5	88.2	65.0	59.2	59.7	54.6	67.4	64.2	61.9	57.7	19.6	16.6	19.5	16.1	50.0	38.8	73.3	67.3	48.6	47.8
FireRed-Image-Edit-1.1	20B	97.0	95.3	97.2	94.5	78.0	73.0	70.1	65.9	62.9	61.2	66.2	62.8	22.2	19.5	20.4	17.9	73.3	68.1	75.0	70.2	54.6	52.9
LongCat-Image-Edit	6B	98.5	94.6	94.0	90.4	60.5	54.9	53.5	49.6	58.8	55.1	48.9	45.0	19.7	17.2	18.9	17.1	0.0	0.0	0.0	0.0	44.8	41.1
GLM-Image	9B+7B	74.2	65.6	65.0	55.7	39.5	35.9	37.5	34.1	32.4	27.9	25.0	21.9	15.6	12.9	15.1	11.9	56.7	46.1	61.7	49.8	32.0	28.5
JoyAI-Image-Edit	8B+16B	98.1	94.3	96.8	92.7	74.4	69.4	71.6	68.2	60.7	56.8	70.8	66.8	22.6	18.7	24.2	21.3	38.3	32.9	53.3	47.1	50.4	53.6
FLUX.2-klein-9B	9B	51.9	48.3	89.4	84.5	57.6	49.3	54.3	50.6	69.2	61.4	69.4	65.6	20.2	17.9	17.3	15.5	70.0	59.8	81.7	75.0	40.3	48.2
Step1x-Edit-v1p2	7B+12B	90.8	87.6	93.5	89.9	45.8	41.1	45.9	42.3	53.2	47.9	43.3	38.5	26.6	**23.7**	38.5	**35.7**	33.3	23.9	45.0	36.5	43.5	47.7

Basic editing proficiency does not ensure strong reasoning. FireRed-Image-Edit-1.1 achieves 95.3/94.5 on Basic Editing but only 19.5/17.9 on Semantic Reasoning. This suggests that, for complex

real-world requests, basic text editing and text rendering serves as a foundation, while reasoning over user intent and translating it into actionable editing instructions is critical.

Open models exhibit uneven capability profiles. Boogu-Image-0.1 scores 95.2/91.9 on Basic Editing but only 69.2/64.4 on Layout Control. BAGEL scores lower still on Layout Control (39.0/32.9): in the curved-can example of Figure 6, it places much of the requested lettering outside the can, while LLADA-Image extends the lettering beyond the can's boundary. Multi-Reference performance also varies sharply: Qwen-Image-2.1 reaches 80.4/94.8, whereas Boogu-Image-0.1, and LLADA-Image, InternVL-U score 0.0, revealing a multi-reference editing lacking.

Thinking can strengthen the understanding step before editing. For SenseNova-U1.5-8B-MOT, thinking boosts Semantic Reasoning from 34.7/31.8 to 71.7/70.1 and Overall from 67.3/66.2 to 82.0/81.1, with the largest gains in translation, correction, and knowledge-fused editing. However, the effect is model-dependent: BAGEL drops from 29.7/31.8 to 23.3/25.5, suggesting that its improved reasoning is not sufficiently aligned with the downstream editing process.

5.4 Model Failure Analysis

Table 4 averages rubric deficits $100(1 - a_{ij})$ across 21 configurations, where a_{ij} scores criterion j on sample i. These values measure the shortfall from full credit, not error frequency.

Capability-specific bottlenecks. Layout Control struggles most with Target Localization (51.1% deficit), and Style Design with Font Accuracy (40.1%). Basic Editing has smaller deficits (13.3–25.9%); high normalized entropy (0.947–0.998) indicates deficits across attributes.

Reasoning and output coverage. Semantic Reasoning's largest deficit is Semantic Correctness (72.4%). Multi-Reference Composition has 26.2% missing outputs, versus under 0.2% elsewhere; these are excluded from attribute deficits but receive zero leaderboard scores.

Qualitative example. Figure 7 shows that ireRed-Image-Edit-1.1's zero semantic Gate yields a zero sample score. This case illustrates that an image editing model needs sufficient reasoning to transform relevant knowledge into correctly rendered text.

Table 4: **Capability–attribute failure map.** Mean criterion-score deficits (%); darker is worse. *Miss.*: missing-output rate; –: inapplicable; H_{norm}: normalized deficit entropy.

Attr.	Basic	Layout	Style	Reason.	Multi-ref.
TA	19.8	30.8	25.0	65.6	27.7
TL	13.3	51.1	12.1	47.2	–
FA	25.9	31.3	40.1	58.7	32.5
LA	17.7	39.6	17.1	60.8	37.0
VN	17.2	27.8	36.1	47.4	33.8
NTP	19.1	27.3	10.0	29.6	32.0
SC	–	–	26.7	72.4	–
MRA	–	–	–	–	31.7
Miss.	0.1	0.1	0.0	0.1	26.2
H_{norm}	0.989	0.985	0.947	0.984	0.998

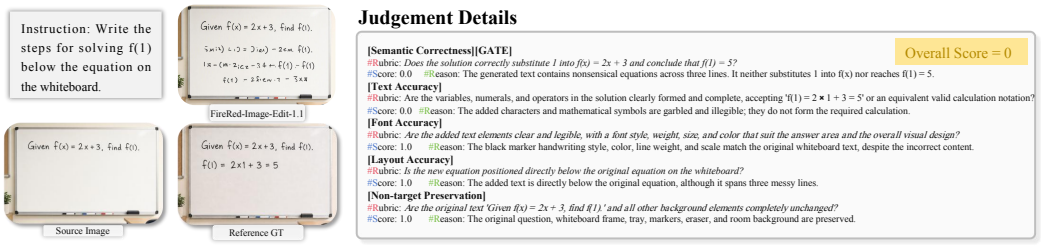


Figure 7: FireRed-Image-Edit-1.1: a zero semantic Gate yields a zero sample score.

6 Conclusion

In this paper, we introduce TextReVision, a bilingual benchmark of 1,200 human-curated samples spanning 16 image text editing scenarios and five capability families. To assess whether models fulfill the intended edit, we develop sample-specific rubrics and a completion-aware *Gate-and-Craft* scoring protocol that also captures execution quality. Experiments with 18 models in 21 configurations reveal persistent weaknesses in layout control, semantic reasoning, and multi-reference composition, despite strong performance on basic edits. By combining diverse tasks with criterion-level diagnostics, TextReVision provides a systematic foundation for measuring progress toward reliable, intent-faithful image text editing.

AI USE STATEMENT

Generative AI tools assisted with software implementation and debugging, and with the clarity, grammar, and presentation of the manuscript. VLMs also served as automated evaluators in the experimental pipeline; their inputs, scoring protocol, and human validation are described in Sections 4.4 and 5.2 and Appendix D. The authors reviewed AI-assisted code and text, checked the evaluation outputs and analyses, made the final decisions on methodology and conclusions, and take responsibility for this work.

REFERENCES

Black Forest Labs. Flux.2 [klein] 9b. https://huggingface.co/black-forest-labs/FLUX.2-klein-9B, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

ByteDance Seed Team. Seedream 5.0 lite. https://seed.bytedance.com/en/blog/deeper-thinking-more-accurate-generation-introducing-seedream-5-0-lite, 2026.

Chuyan Chen, Haoxing Chen, Kun Chen, Zhenglin Cheng, Long Cui, Ruishan Fang, Zhangxuan Gu, Zhicheng Huang, Zhenzhong Lan, Yuanting Lei, Haoquan Li, Jianguo Li, Rongchuan Li, Sidu Li, Tao Lin, Deyuan Liu, Jiacheng Liu, Lin Liu, Yuxuan Lou, Zhisheng Lu, Yuxin Ma, Shuheng Shen, Peng Sun, Chaoyang Wang, Hongjun Wang, Xiaomei Wang, Yongxin Wang, Chengzhang Wu, Hongru Wu, and Jun Xie. Llada-image: Building strong image generators with fully open training recipes. *arXiv preprint arXiv:2609.03796*, 2026a.

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 agentic multimodal generation via understanding under a minimal budget. *arXiv preprint arXiv:2607.13125*, 2026b.

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*, pp. 386–402. Springer, 2024.

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.

Google. Introducing nano banana pro, 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/.

Google. Gemini 3.7 Flash, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.7-flash.

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.

Rui Lan, Yancheng Bai, Xu Duan, Mingxing Li, Dongyang Jin, Ryan Xu, Dong Nie, Lei Sun, and Xiangxiang Chu. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329*, 2025.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, et al. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering. In *European Conference on Computer Vision*, pp. 361–377. Springer, 2024.

OpenAI. GPT-6 Astra, 2026a. URL https://developers.openai.com/api/docs/models/gpt-6-astra.

OpenAI. Introducing chatgpt images 2.0, 2026b. URL https://openai.com/index/introducing-chatgpt-images-2-0/.

PaddleOCR Team. PP-OCRv4. PaddleOCR technical documentation, 2023. URL https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.7/doc/doc_ch/PP-OCRv4_introduction.md.

Yadong Qu, Qingfeng Tan, Hongtao Xie, Jianjun Xu, Yuxin Wang, and Yongdong Zhang. Exploring stroke-level modifications for scene text editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2119–2127, 2023.

Qwen Team, Alibaba Group. Qwen-image-2511. https://qwen.ai/blog?id=qwen-image-2512, 2025. Apache-2.0 License.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *2024 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 8871–8879. IEEE, 2024.

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.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. FireRed-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

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

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv e-prints*, pp. arXiv–2506, 2025.

Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1500–1508, 2019.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *Advances in Neural Information Processing Systems*, 38, 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.

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.

Fukun Yin, Shiyu Liu, Yucheng Han, Zhibo Wang, Peng Xing, Rui Wang, Wei Cheng, Yingming Wang, Aojie Li, Zixin Yin, et al. Reasonedit: Towards reasoning-enhanced image editing models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23283–23293, 2026.

Z.AI. Glm-image: Auto-regressive for dense-knowledge and high-fidelity image generation. `https://z.ai/blog/glm-image`, 2026. Apache-2.0 License.

Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. Textctrl: Diffusion-based scene text editing with prior guidance control. *Advances in Neural Information Processing Systems*, 37: 138569–138594, 2024.

Hui Zhang, Juntao Liu, Zongkai Liu, Liqiang Niu, Fandong Meng, Zuxuan Wu, and Yu-Gang Jiang. Weedit: A dataset, benchmark and glyph-guided framework for text-centric image editing. In *European Conference on Computer Vision*, pp. 488–507. Springer, 2026.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 36:31428–31449, 2023.

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.

A BENCHMARK COMPOSITION AND CURATION

A.1 ENCHMARK DESIGN AND CONSTRUCTION

Scenario taxonomy. TEXTREVISION contains 1,200 samples across 16 scenarios and five capabilities, with 600 samples in each language. Table 5 defines every scenario and reports its bilingual composition. Scenario labels identify the primary capability; sample-specific rubrics capture all requirements of an edit, including those spanning multiple capabilities. Figure 1 illustrates a representative case for each scenario.

Table 5: **Taxonomy and composition of TEXTREVISION.** The 1,200 samples comprise 600 Chinese and 600 English samples. ZH/EN gives the Chinese/English sample counts for each scenario.

Scenario	Operational definition	Samples	ZH/EN
Basic Editing			280 samples
Text replacement	Replace specified source text with target text while retaining the intended local design.	100	50/50
Text insertion	Add new text at the instructed target without removing existing content.	100	50/50
Text removal	Remove the selected text and reconstruct its background.	80	40/40
Layout Control			260 samples
Geometric transformation	Transform existing text in scale, orientation, or geometry as instructed.	80	40/40
Path text insertion	Place text along a specified curve, surface, or trajectory.	100	50/50
Document element adjustment	Modify text-bearing document elements while respecting field and document structure.	80	40/40
Style Design			140 samples
Stylized replacement	Replace content while reproducing or changing a specified typographic style.	40	20/20
Artistic insertion	Insert display text with an instructed artistic treatment.	40	20/20
Creative fusion	Integrate text with objects, materials, or scene structure as a visual design.	60	30/30
Semantic Reasoning			490 samples
Scene-text translation	Translate selected scene text and render the translation in context.	90	45/45
Degraded restoration	Infer and restore text that is degraded, occluded, or incomplete.	80	40/40
Scene-text correction	Detect and correct erroneous text in the image.	90	45/45
Referring editing	Resolve a linguistic or visual referring expression and edit its target.	100	50/50
Form filling	Interpret a structured form and populate the requested fields.	50	25/25
Knowledge-fused editing	Derive the required content using calculation, factual knowledge, or contextual reasoning.	80	40/40
Multi-Reference			30 samples
Multi-reference design	Combine content, layout, or style cues from multiple references according to their stated roles.	30	15/15
Total		**1,200**	**600/600**

Source imagery. Source images comprise both real-world and synthetic imagery. Real-world images are collected primarily from Pexels[1]. covering natural scenes, posters and advertisements, documents and forms, and designed graphics. For scenarios with insufficient natural coverage, we construct synthetic source images through a VLM-assisted generation pipeline, where VLM-derived descriptions guide image generation.

Reference-edit construction. We construct reference edits through both open- and closed-resource routes. The open-resource route combines manual editing in Adobe Photoshop with a glyph-image-based workflow built on FLUX-Text (Lan et al., 2025), where glyph images provide visual conditioning for text editing. The closed-resource route uses advanced image editors to generate candidate reference edits for subsequent human selection.

[1] https://www.pexels.com/

Reference acceptance. References from both routes are retained only after human verification against the source image and editing instruction. Each accepted reference represents a valid outcome of the requested edit and supports task construction and quality auditing. During evaluation, the automatic judge receives the human-verified reference for each sample as a correct-edit example.

A.2 Filtering and Human Verification

We first apply automatic checks for text accuracy, instruction–image consistency, image quality, layout, duplication, and safety. Human reviewers then verify the source text, editing target, instruction feasibility, correctness of the edited content, visual quality, and preservation of unedited regions. Samples are rejected if they contain ambiguous instructions, incorrect reference images, severe visual artifacts, or unresolved safety issues.

A.3 Dataset Characteristics

We analyze OCR-detected text coverage across 1,200 input samples, including additional reference images; no text is detected in 64 images. Text coverage is measured as the fraction of image area covered by the union of PP-OCRv4 detection polygons, with capability-level medians ranging from 3.71% to 10.54%. Among 400 samples with explicit target text (200 per language), median target length is 8 Unicode code points in English and 4 in Chinese, excluding whitespace and counting each target once per sample. These statistics describe textual content and coverage rather than editing difficulty.

B Rubric Construction and Scoring Guidelines

B.1 Rubric Construction and Gate Selection

The curated annotations comprise 1,200 sample-specific rubrics with 6,426 criteria. For each sample, annotators write concrete evaluation criteria covering the capabilities required by the instruction, with each criterion assigned to one of the eight dimensions. Relevant criteria are selected to form the sample-specific rubric, with one designated as the Gate and the others as Craft criteria. The Gate captures the primary requirement for task completion, while Craft criteria assess the remaining aspects of edit quality.

B.2 Scenario-Specific Dimension Weights

Table 6 reports the Craft weights for each scenario and its applicable dimensions. Among the 16 scenarios, document adjustment, referring editing, and knowledge-fused editing each use two dimension variants, yielding 19 configurations in total. Weights are assigned by scenario and selected dimensions, are shared across models, and sum to 100%. They are used to compute C_i in Eq. 1, separately from β.

Applicability of the variants. Variants are used when a sample requires a different set of evaluation dimensions. Document deletion and referring removal omit criteria related to new text and typography, while knowledge-based marking uses layout and preservation criteria. Variant selection depends only on the requirements of the edit, not on model outputs.

Interpretation and sensitivity. The β sweep (Appendix C) holds Craft weights fixed, so we separately characterize sensitivity to the Craft weights, For normalized weight vectors w, w' over the same dimensions, with fixed ratings and Gate,

$$|100s_i(w) - 100s_i(w')| \leq 50(1-\beta)g_i \, \|w - w'\|_1. \tag{3}$$

This follows from ratings in $[0, 1]$ and $\sum_j (w_j - w'_j) = 0$. At $\beta = 0.5$, transferring five percentage points between two Craft dimensions changes a sample's score by at most $2.5g_i$ points. This bound quantifies local score sensitivity; it does not guarantee unchanged rankings between closely scoring models.

Table 6: **Scenario-specific dimension weights.** Craft weights sum to 100% within each row; Gray rows identify the three reduced-rubric variants.

Scenario / rubric variant		n	Gate	Craft weights (%)						
				TA	TL	FA	LA	VN	NTP	SC
Basic Editing										
1.1	Text replacement	100	**TA**	–	20	30	5	25	20	–
1.2	Text insertion	100	**TA**	–	20	30	25	15	10	–
1.3	Text removal	80	**TL**	–	–	–	–	70	30	–
Layout Control										
2.1	Geometric transformation	80	**LA**	30	–	35	–	25	10	–
2.2	Path text insertion	100	**LA**	25	–	25	–	35	15	–
2.3	Document adjustment: text	77	**LA**	25	40	10	–	–	25	–
2.3	Document adjustment: removal	3	LA	–	65	–	–	–	35	–
Style Design										
3.1	Stylized replacement	40	**FA**	35	25	–	10	15	15	–
3.2	Artistic insertion	40	**FA**	35	5	–	25	25	10	–
3.3	Creative fusion	60	**VN**	30	–	30	20	–	–	20
Semantic Reasoning										
4.1	Scene-text translation	90	**SC**	25	10	25	–	30	10	–
4.2	Degraded restoration	80	**TA**	–	20	35	–	30	15	–
4.3	Scene-text correction	90	**SC**	35	25	20	–	10	10	–
4.4	Referring editing: text	81	**TL**	30	–	20	–	10	20	20
4.4	Referring editing: removal	19	TL	–	–	–	–	30	30	40
4.5	Form filling	50	**SC**	30	35	10	15	–	10	–
4.6	Knowledge-fused editing: text	75	**SC**	40	–	15	25	–	20	–
4.6	Knowledge-fused editing: marking	5	SC	–	–	–	55	–	45	–
Multi-Reference										
5.1	Multi-reference design	30	**MRA**	20	–	20	20	20	20	–

n: sample count. A dash denotes a dimension excluded from Craft, including the Gate dimension.

TA: Text Accuracy; TL: Target Localization; FA: Font Accuracy; LA: Layout Accuracy; VN: Visual Naturalness; NTP: Non-target Preservation; SC: Semantic Correctness; MRA: Multi-Reference Alignment (Gate only).

B.3 Rubric Scoring and Interpretation

Five-level scoring scale. Each selected criterion is rated on a five-level scale: `fully_met` (1.00), `mostly_met` (0.75), `partially_met` (0.50), `minimally_met` (0.25), and `not_met` (0.00). The rating reflects how well the criterion's requirements are satisfied, rather than evaluator confidence, with scene-specific anchors defining each level.

In the two-stage judge (Appendix D.4), Stage 1 uses the Question, Focus, and requirement-level anchors to produce visual evidence and a conclusion without assigning a rating. Stage 2 maps this evidence to one of the five levels. A rating below `fully_met` requires a concrete defect, omission, or failure supported by the corresponding reason. If the evidence is insufficient or contradictory, the response is treated as an `evaluation_error` rather than assigned a score.

The scene-specific anchors determine the meaning of each level; intermediate ratings do not correspond to fixed percentages of correct pixels or characters, and `not_met` does not necessarily imply that no edit was attempted. Criterion ratings remain on the five-point scale, while aggregation may produce any value in $[0, 1]$.

Appendix H presents the shared template for scene-specific dimension prompts. Table 7 summarizes the scope of the eight evaluation dimensions and clarifies their respective roles in the rubric.

Scoring guidelines. Only the dimensions selected for a sample are scored. Each score is judged against the instruction and its rubric criteria; reference images serve as supporting examples and need not be reproduced exactly unless explicitly required. A score below 1.00 must be supported by a specific defect in that dimension. The same defect should not be penalized across multiple dimensions

Table 7: **Evaluation dimensions and their scopes.**

Dimension	Scope
Text Accuracy (TA)	Correctness, completeness, and legibility of rendered text.
Target Localization (TL)	Correct selection and coverage of the instructed target.
Font Accuracy (FA)	Typeface, size, color, weight, and other font attributes.
Layout Accuracy (LA)	Position, alignment, spacing, orientation, and geometric layout.
Visual Naturalness (VN)	Visual integration and local appearance of the edited region.
Non-target Preservation (NTP)	Preservation of content outside the intended edit region.
Semantic Correctness (SC)	Correctness of the intended meaning, value, or concept.
Multi-Reference Alignment (MRA)	Correct use and alignment of multiple reference images.

unless it causes separate failures. Missing or corrupted judge inputs are treated as evaluation errors. If no model output is produced, the score is zero; otherwise, an unmet requirement follows the corresponding 0.00 anchor.

B.4 Worked Example of Rubric Scoring

Figure 8 shows the four image views together with all selected rubric questions (Q), their scores, and the corresponding explanations (R). The questions, scores, and explanations are taken from an archived Gemini-3.7-Flash evaluation. We recompute the Craft and overall scores using Table 6 and $\beta = 0.5$, without re-scoring any rubric question.

CASE 01 / ARTISTIC TEXT INSERTION

Sample 594 | English | Generator: GPT-Image-2

In the foreground lantern field in front of the Treasury facade, arrange the glowing amber paper bag lanterns to spell out 'CARVED FROM' on the top line and 'TIME ITSELF' on the bottom line.



Red: original luminance | Cyan: prediction luminance

Gate 1.00 **Craft 1.000** **Overall 1.000** Judge: Gemini-3.7-Flash

Font Accuracy GATE **1.00 / 1.00**

Q Are the letters of 'CARVED FROM' and 'TIME ITSELF' formed from glowing amber paper bag lanterns?

R All letters of 'CARVED FROM' and 'TIME ITSELF' are constructed from glowing amber paper bag lanterns matching the scene's existing lantern material and lighting.

Text Accuracy Craft w=0.35 **1.00 / 1.00**

Q Does the added text contain exactly 'CARVED FROM' and 'TIME ITSELF', with no missing, extra, or incorrect characters?

R The text accurately renders 'CARVED FROM' on the top line and 'TIME ITSELF' on the bottom line with no missing, extra, or misspelled characters.

Layout Accuracy Craft w=0.25 **1.00 / 1.00**

Q Are the words arranged in two distinct horizontal lines with 'CARVED FROM' positioned directly above 'TIME ITSELF' in the foreground?

R The text is arranged in two horizontal lines in the foreground, with 'CARVED FROM' centered directly above 'TIME ITSELF' following proper ground perspective.

Visual Naturalness Craft w=0.20 **1.00 / 1.00**

Q Do the lanterns forming the letters blend realistically with the sandy ground, showing natural light glow and shadows?

R The lanterns blend seamlessly with the sandy ground, showing realistic warm light casts, contact shadows, and perspective integration.

Non-target Preservation Craft w=0.15 **1.00 / 1.00**

Q Are the Treasury facade, the rock cliffs, and the distant background lanterns preserved without any alterations, and are all remaining pre-existing non-target contents—including every other unedited text element, person, object or illustration, background detail, color, lighting, texture, and the overall composition—preserved unchanged from the original image, with no omission, addition, substitution, distortion, or unintended restyling?

R The Treasury facade, surrounding rock cliffs, canyon passageway, and night sky are preserved without any unintended alterations.

Target Localization Craft w=0.05 **1.00 / 1.00**

Q Is the foreground lantern field correctly identified as the target region for the two added text lines?

R The foreground lantern field was correctly identified and utilized to form the two lines of text.

Figure 8: **Complete scoring example: English artistic text insertion.** Sample 594, generated by GPT-Image-2, receives 1.00 on all six criteria. Font Accuracy is the Gate: the lettering must be formed from glowing amber paper bag lanterns. Gate $= 1.00$ and Craft $= 1.00$ yield an overall score of 1.00.

C GATE-AND-CRAFT ANALYSIS

Analysis setup. We use the paired annotations from Table 2, Panel A: 32 prompts evaluated on four generation systems by two annotators, yielding 128 outputs and 256 annotated results. For each result, the rubric scores are aggregated and compared with the annotator's overall score. All aggregation methods are evaluated on the same data.

Sensitivity to β. Figure 9(a,b) varies β from 0 to 1 in steps of 0.05. CCC remains stable around $\beta = 0.4$–0.5, reaching 0.9351 at $\beta = 0.45$ and 0.9348 at $\beta = 0.5$. MAE is lowest at $\beta = 0.5$ (0.0725), compared with 0.0922 for the product form ($\beta = 0$) and 0.1065 for Gate alone ($\beta = 1$). We therefore use $\beta = 0.5$ throughout the benchmark. Confidence intervals are obtained from 2,000 bootstrap resamples over the 32 prompt groups.

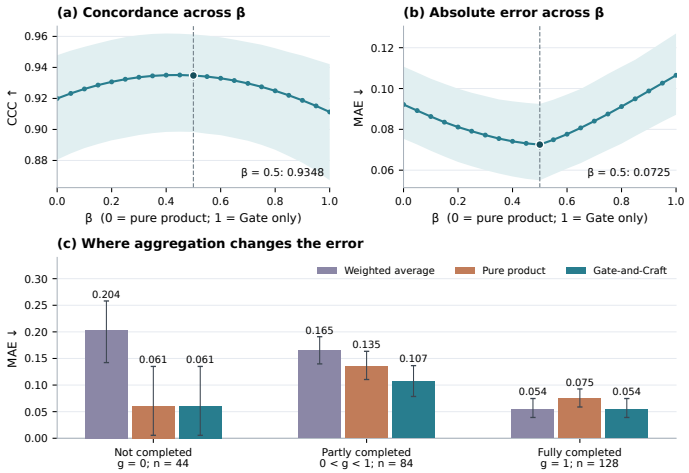


Figure 9: **Sensitivity to β and comparison across Gate levels. (a,b)** Agreement with annotator overall scores as β varies; the dashed line marks $\beta = 0.5$. **(c)** MAE of the three aggregation methods for outputs with $g = 0$, $0 < g < 1$, and $g = 1$. Error bars show 95% confidence intervals for the MAE estimates.

Performance across Gate levels. Figure 9(c) separates the results into $g = 0$ ($n = 44$), $0 < g < 1$ ($n = 84$), and $g = 1$ ($n = 128$). Gate-and-Craft achieves MAEs of 0.061, 0.107, and 0.054 across the three groups, respectively. Its main advantage appears for partially completed edits, where it outperforms both weighted averaging (0.165) and the product form (0.135), while remaining competitive at $g = 0$ and $g = 1$.

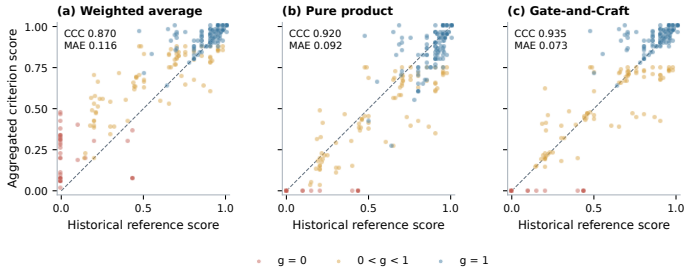


Figure 10: **Agreement between aggregation scores and annotator overall scores.** Each panel shows the same 256 annotated results. Colors indicate the Gate score, and the diagonal indicates exact agreement. Gate-and-Craft uses $\beta = 0.5$.

Figure 10 further shows that weighted averaging can assign substantial scores when the Gate is zero, whereas the product form tends to give lower scores to partially completed edits. Gate-and-Craft provides a middle ground between these two behaviors.

D AUTOMATIC JUDGING AND HUMAN VALIDATION

D.1 JUDGE EVALUATION SETUP

The judge receives the instruction, source and output, human-verified reference edit, conditioning references, and selected rubric questions and anchors. The instruction and rubric define requirements; the reference is a correct-edit example. The judge is not informed which model produced the output. Criterion-level reasons and five-level ratings are validated for completeness and schema compliance before aggregation (Appendix D.4). Invalid judge responses are retried, while missing or unreadable model outputs receive a score of zero.

D.2 RED/CYAN ALIGNMENT AID FOR VISUAL AUDIT

In the alignment aid, ORIGINAL is shown in red and GENERATED in cyan. Aligned regions appear gray, while red/cyan separation highlights areas that may have changed. Such separation is only a visual cue and does not by itself indicate an error. Changes are evaluated against the instruction and criterion anchors. For Non-target Preservation, only unintended changes outside the target and necessary repair regions are considered.

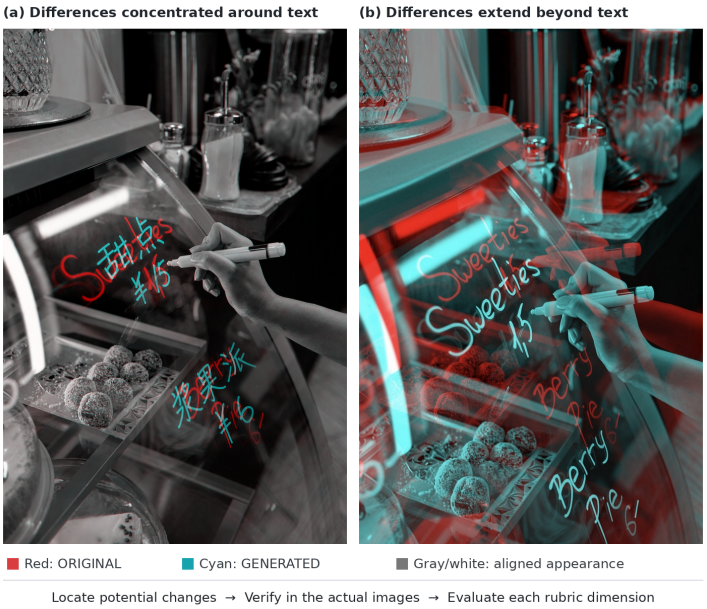


Figure 11: **Red/cyan alignment aids.** (a) Separation is concentrated near edited text. (b) Separation also appears around the hand, pen, pastry tray, and background containers.

D.3 ENLARGED CROPS FOR FINE-GRAINED VISUAL AUDIT

Figure 12 presents sample 0568, where the instruction requires changing the font from Kaiti to Songti while preserving the text content and layout. The generated image contains two character substitutions. Without enlarged text views, the judge identifies only one of them; with enlarged column views, both are detected. The Text Accuracy score remains 0.50 in both cases because both evaluations fall within the same scoring anchor. Thus, the enlarged views improve error detection rather than the numerical score.

(a) Full natural-scene images

Original

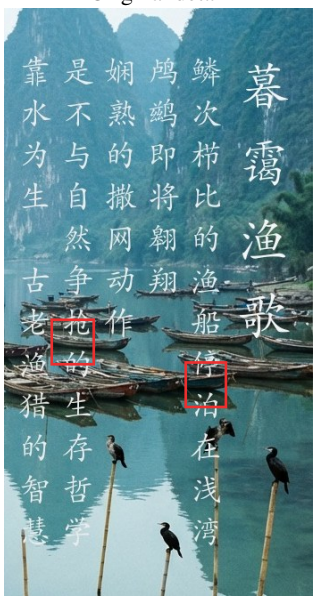


Generated



(b) Paired text-column details

Original detail



Generated detail

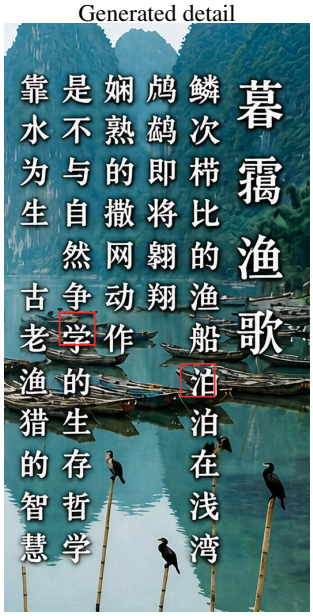


Figure 12: **Enlarged text views reveal a missed character error.** Sample 568, generated by SenseNova-U1.5-8B-MOT (thinking). (a) Original and generated images for a Kaiti-to-Songti conversion that should preserve all text content. (b) Enlarged views of the six text columns. The generated image contains two character substitutions: "tingbo"→"bobo" and "zhengqiang"→"zhengxue". The full-image judgment identifies only the first, whereas the enlarged views identify both. Red boxes are added for visualization only and were not shown to the judge. Text Accuracy remains 0.50 in both cases.

D.4 Two-Stage Judging and Prompt Examples

In diagnostic experiments, even advanced VLMs sometimes assigned status labels or scores that contradicted their own reasons when producing both in one response. To separate visual observation from anchor selection, we use two requests: Stage 1 records visual evidence and criterion-level conclusions; Stage 2 maps these fixed reasons to scoring anchors without images. This makes the evidence-to-anchor decision explicit and auditable; it adds no visual evidence.

D.4.1 STAGE 1: VISUAL EVIDENCE AND REASON GENERATION

Stage 1 uses three system-prompt variants to match the visual inputs used in different scenarios (Table 5). The four-image variant is used for *Creative fusion* (60 samples) and *Form filling* (50 samples), which are evaluated without supplementary paired detail views. The multiple-source variant is used for *Multi-reference design*, identifying the source inputs so that each element can be compared with its corresponding source. The remaining 13 scenarios use the five-image variant, which adds enlarged original–generated pairs for fine-detail inspection. In all three variants, each dimension's Focus specifies what to inspect and compare.

STAGE 1 SYSTEM PROMPT: FIVE-IMAGE INPUT

You are an image editing expert. Evaluate GENERATED against the instruction and labeled images.

- IMAGE 1 — ORIGINAL: The image before editing.
- IMAGE 2 — GENERATED: The edited result to evaluate.
- IMAGE 3 — RED/CYAN ALIGNMENT AID: ORIGINAL in red, GENERATED in cyan.
- IMAGE 4 — REFERENCE: A correct-edit example showing what the intended result should look like.
- IMAGE 5 — SUPPLEMENTARY PAIRED DETAIL: Enlarged target regions at the same scale, paired by row: left ORIGINAL, right GENERATED.

For each dimension, assess Question independently, follow Focus to collect evidence, and use the full Requirement levels to interpret it. Base each reason on concrete visual evidence; do not invent differences or penalize uncertainty alone.

Return each selected dimension exactly once in the specified reason JSON, with evidence and a clear conclusion. Do not output status labels, scores, extra fields, or prose.

Use evaluation_error only for missing/corrupt required inputs or irreconcilable requirements; evaluate poor edits normally.

STAGE 1 SYSTEM PROMPT: FOUR-IMAGE INPUT

You are an image editing expert. Evaluate GENERATED against the instruction and labeled images.

- IMAGE 1 — ORIGINAL: The image before editing.
- IMAGE 2 — GENERATED: The edited result to evaluate.
- IMAGE 3 — RED/CYAN ALIGNMENT AID: ORIGINAL in red, GENERATED in cyan.
- IMAGE 4 — REFERENCE: A correct-edit example showing what the intended result should look like.

For each dimension, assess Question independently, follow Focus to collect evidence, and use the full Requirement levels to interpret it. Base each reason on concrete visual evidence; do not invent differences or penalize uncertainty alone.

Return each selected dimension exactly once in the specified reason JSON, with evidence and a clear conclusion. Do not output status labels, scores, extra fields, or prose.

Use evaluation_error only for missing/corrupt required inputs or irreconcilable requirements; evaluate poor edits normally.

STAGE 1 SYSTEM PROMPT: MULTIPLE SOURCE IMAGES

You are an image editing expert. Evaluate GENERATED RESULT against the instruction and labeled images.

- SOURCE: An original input (image1, image2, etc.) named in the instruction.
- GENERATED RESULT: The edited result to evaluate.
- CORRECT-EDIT EXAMPLE: A correct-edit example showing what the intended result should look like.

For each dimension, assess Question independently, follow Focus to collect evidence, and use the full Requirement levels to interpret it. Base each reason on concrete visual evidence; do not invent differences or penalize uncertainty alone.

Return each selected dimension exactly once in the specified reason JSON, with evidence and a clear conclusion. Do not output status labels, scores, extra fields, or prose.

Use evaluation_error only for missing/corrupt required inputs or irreconcilable requirements; evaluate poor edits normally.

Sample 0001 uses five labeled images and the corresponding five-image system prompt, summarized above. The user request below returns one reason per criterion, without a status or score. Image payloads are omitted; the prompt text is reproduced from the supplied example.

STAGE 1: USER PROMPT

IMAGE 1 — ORIGINAL

IMAGE 2 — GENERATED

IMAGE 3 — RED/CYAN ALIGNMENT AID

IMAGE 4 — REFERENCE

IMAGE 5 — SUPPLEMENTARY PAIRED DETAIL

Instruction: Replace the text ’2019’ with ’2020’.

Text Accuracy
Question: Is the replacement text rendered exactly and legibly as ”2020”, with no missing, malformed, or extra characters?

Focus: Read IMAGE 2 (GENERATED), using the RIGHT GENERATED column of IMAGE 5 (SUPPLEMENTARY PAIRED DETAIL) for small characters when supplied; check the full image for text outside the crops. Read the replacement against the required string; check character completeness and legibility. Consult IMAGE 4 (REFERENCE) for an example of completed text. For Text Accuracy, judge requested text identity and legibility. Differences in brush style, stroke thickness, or font design belong to Font Accuracy unless a required character is genuinely malformed or ambiguous.

Requirement levels:
Use the evidence gathered under Focus to judge how fully the requirement in Question is satisfied.
- fully_met: For replacement text, all required text content is exactly correct and clearly legible, with no missing, extra, substituted, or malformed characters.
- mostly_met: For replacement text, all required text content is present and textually correct; only a minor glyph defect is visible, and every character remains unambiguous.
- partially_met: For replacement text, the required text remains meaningfully recognizable, but a key character or required string is wrong, missing, extra, or ambiguous, or roughly half of multiple required strings are correct.
- minimally_met: For replacement text, only a small portion of the required text is correct or legible, so the complete required content cannot be read reliably.
- not_met: For replacement text, the required text is absent, unrelated, or completely unreadable.

Target Localization
Question: Is the edit applied to the original ’2019’ text region?

Focus: Start with full IMAGE 1 (ORIGINAL) and IMAGE 2 (GENERATED); use corresponding LEFT ORIGINAL / RIGHT GENERATED rows of IMAGE 5 only to confirm local target coverage when supplied. Identify every original text instance designated for replacement and verify that each received the edit; judge final placement under Layout Accuracy. Consult IMAGE 4 (REFERENCE) for example target coverage.

Requirement levels:
Use the evidence gathered under Focus to judge how fully the requirement in Question is satisfied.
- fully_met: Selection of source text instance designated for replacement is fully correct: every required instance receives the intended edit, with no target confusion or omission.
- mostly_met: Selection of source text instance designated for replacement is correct overall across all required instances, with only a minor boundary or coverage imperfection that does not omit or confuse a required target.
- partially_met: Selection of source text instance designated for replacement identifies the correct target class or broad object, but one important required instance is missed or an incorrect instance is included, or only about half of multiple required instances are selected.
- minimally_met: Selection of source text instance designated for replacement is mostly incorrect: most required instances are missed or most selections refer to unrelated instances, though some correct localization remains evident.
- not_met: Selection of source text instance designated for replacement fails completely: no correct target instance is selected, or no edit is performed on any required target.

Font Accuracy
Question: Does the font of ’2020’ (family, weight, color, and style) match that of the original ’2019’ it replaced?

Focus: Start with IMAGE 5 (SUPPLEMENTARY PAIRED DETAIL): compare LEFT ORIGINAL with RIGHT GENERATED at the same scale when supplied. Use full IMAGE 1 (ORIGINAL) and IMAGE 2 (GENERATED) for source lettering or edited text outside the crops. Compare replacement lettering with the corresponding ORIGINAL lettering for glyph style, stroke thickness, nominal letter height, color/gradient, and outline/shadow. State the attributes on each side and apply requested font changes; text content and spacing are separate. Consult IMAGE 4 (REFERENCE) for an example of the requested typography. For lettering covered by IMAGE 5, base the ORIGINAL-versus-GENERATED font comparison on its paired detail row rather than full-scene similarity. Use full images for relevant lettering outside the crops as specified above.

Requirement levels:

STAGE 1: USER PROMPT (CONTINUED)

Use the evidence gathered under Focus to judge how fully the requirement in Question is satisfied.
- fully_met: For replacement lettering, every explicitly requested font attribute is satisfied. Attributes not explicitly changed by the instruction match the corresponding ORIGINAL lettering, with no clearly visible deviation in font family, glyph style, stroke weight, nominal letter size, color, or applicable effects. Broad stylistic similarity or compatibility with the local design alone is insufficient for full credit.
- mostly_met: For replacement lettering, all essential font requirements are satisfied, with one minor visible deviation in a non-critical typographic attribute.
- partially_met: For replacement lettering, one important requested font attribute is wrong, or several moderate typographic mismatches are visible, while substantial stylistic similarity remains.
- minimally_met: For replacement lettering, only a small part of the required or contextually appropriate typography is matched, and most font attributes are clearly inconsistent.
- not_met: For replacement lettering, there is no meaningful typographic match to the requirement or relevant source style, or the font appearance cannot be assessed.

Visual Naturalness
Question: Is the edited region around '2020' visually natural, with clean edges where the blue numbers overlap the white letters of 'Alpen' and the blue sky background, free of smudging or ghosting?

Focus: Start with IMAGE 5 (SUPPLEMENTARY PAIRED DETAIL): compare the LEFT ORIGINAL surface with RIGHT GENERATED at edit boundaries when supplied. Check full IMAGE 1 (ORIGINAL) and IMAGE 2 (GENERATED) for continuity with surrounding surfaces. Inspect the replaced text and underlying surface for old strokes, seams, texture breaks, and inconsistent lighting or perspective; font choice is separate. Consult IMAGE 4 (REFERENCE) for an example of plausible surface integration.

Requirement levels:
Use the evidence gathered under Focus to judge how fully the requirement in Question is satisfied.
- fully_met: The visual integration of region where the source text is removed and the replacement is rendered is seamless and physically or stylistically coherent, with consistent texture, lighting, perspective, and boundaries and no visible editing artifacts.
- mostly_met: The visual integration of region where the source text is removed and the replacement is rendered is convincing, with only one subtle localized artifact that does not disrupt the edit.
- partially_met: The visual integration of region where the source text is removed and the replacement is rendered remains coherent, but one obvious artifact or several moderate local inconsistencies are visible.
- minimally_met: The visual integration of region where the source text is removed and the replacement is rendered shows severe artifacts, ghosting, smudging, broken boundaries, or implausible surface integration, though the attempted edit remains recognizable.
- not_met: The required edit is absent so integration cannot be assessed, or the visual integration of region where the source text is removed and the replacement is rendered is incoherent, unrepaired, or dominated by artifacts such that no credible integration is achieved.

Non-target Preservation
Question: Apart from the intended replacement of "2019" with "2020", are the white 'Alpen' text, the name 'Franz Aßhauer', the mountain landscape, and the 'SIEGBERT LINNEMANN VERLAG' logo in the bottom right corner, together with all remaining pre-existing non-target contents—including every other unedited text element, logo or symbol, person, object or illustration, background detail, color, texture, and the overall composition—preserved unchanged from the original image, with no omission, addition, substitution, distortion, or unintended restyling?

Focus: Compare full IMAGE 1 (ORIGINAL) and IMAGE 2 (GENERATED). Use IMAGE 3 (RED/CYAN ALIGNMENT AID) to locate suspected changes, then confirm them in images 1 and 2; color separation alone is not a defect. Check retained objects, other text, surface details, and image boundaries outside the replacement and necessary repair region. Consult IMAGE 4 (REFERENCE) to clarify the intended edit. Judge preservation from the full images, not IMAGE 5 target crops. Ignore differences caused solely by compression, resampling, resolution, or transfer/encoding. Count only a concrete unintended content change confirmed in the relevant source and generated images.

Requirement levels:
Use the evidence gathered under Focus to judge how fully the requirement in Question is satisfied.
- fully_met: All non-target content in the original image is preserved without visible unintended change.
- mostly_met: Non-target content in the original image is preserved except for one small, localized, non-critical unintended variation.
- partially_met: Several localized unintended changes, or one important non-target alteration, are visible in the original image, while most non-target content is preserved.
- minimally_met: Non-target content in the original image shows widespread redraw, distortion, omission, addition, or restyling across multiple important areas.

STAGE 1: USER PROMPT (CONTINUED)

- not_met: Most non-target content in the original image is replaced, lost, or fundamentally changed beyond the requested edit.

Layout Accuracy
Question: Does '2020' fit the original '2019' text region with matching position, alignment, and orientation, without overflow or crowding?

Focus: Start with full IMAGE 1 (ORIGINAL) and IMAGE 2 (GENERATED) for spatial relationships; use LEFT ORIGINAL / RIGHT GENERATED rows of IMAGE 5 for local spacing and boundaries when supplied. Compare replacement position, alignment, spacing, orientation, and fit within the original text-bearing region against the requested placement; font attributes are separate. Consult IMAGE 4 (REFERENCE) for the intended spatial relationship.

Requirement levels:
Use the evidence gathered under Focus to judge how fully the requirement in Question is satisfied.
- fully_met: For replacement text within the replaced text's role and region, every explicit layout relationship in the instruction and rubric question is satisfied.
- mostly_met: For replacement text within the replaced text's role and region, all essential layout relationships are satisfied, with only one minor visible offset in position, alignment, spacing, orientation, or scale.
- partially_met: For replacement text within the replaced text's role and region, the overall layout direction is correct, but one important relationship is violated, or only about half of multiple explicit layout requirements are satisfied.
- minimally_met: The intended layout is only weakly recognizable; most explicit spatial or geometric relationships are incorrect.
- not_met: The requested layout operation is absent, reversed, or placed in an unrelated configuration.

Output:
Return exactly one JSON object with only the reason field and every selected dimension exactly once. Each reason must contain a clear conclusion about whether its requirements are fully met, partly met, or unmet, followed or preceded by concrete visual evidence relevant to that same dimension. For Font Accuracy, describe the required or corresponding source typography and the generated typography using the image roles named in that dimension's Focus, then give the conclusion inside the same reason. For text covered by IMAGE 5, cite what LEFT ORIGINAL and RIGHT GENERATED visibly show; use full images only for relevant text outside its crops. If not fully met, name the actual visible defect or omission and its consequence. Do not infer a match from broad similarity, invent minor defects, or change facts to fit a score. Do not output any score, five-level status label, matched_requirements, visible_differences, visible_defects, or requirement_status.

```
{
 "reason": {
  "Text Accuracy": "<conclusion with concrete visual evidence>",
  "Target Localization": "<conclusion with concrete visual evidence>",
  "Font Accuracy": "<conclusion with concrete visual evidence>",
  "Visual Naturalness": "<conclusion with concrete visual evidence>",
  "Non-target Preservation": "<conclusion with concrete visual evidence>",
  "Layout Accuracy": "<conclusion with concrete visual evidence>"
 }
}
```

For missing or corrupt required input or irreconcilable requirements only, return
{"evaluation_error":{"type":"input_error|configuration_conflict","reason":"<specific cause>"}}

D.4.2 Stage 2: Mapping Reasons to Scoring Anchors

The second request receives the instruction, the same criterion questions and requirement-level anchors, and the complete Stage 1 reason JSON. It receives no images and introduces no new visual observations. The returned status labels map to scores as specified in Appendix B.3. For compactness, the two marked placeholders replace the repeated Question-and- Requirement-level sections and the Stage 1 reasons. These are supplied in full at runtime; the literal "example" strings in the source template are dummy values, not judge observations.

STAGE 2: SYSTEM PROMPT

You classify rubric dimensions from supplied written reasons and scene-specific scoring anchors. You do not receive images. Use only the facts explicitly stated in each reason; do not infer, invent, or request visual details. Judge dimensions independently.

Choose fully_met when a reason affirmatively says that the dimension's requirements are met and identifies no relevant defect. A score below fully_met requires a concrete defect, omission, or failure within that same dimension's reason, with enough detail to support its anchor. Words such as 'mostly', 'closely', 'minimal', or 'minor' alone do not establish a defect. Do not turn a positive reason into mostly_met merely because the original result may have been imperfect. If the reasons genuinely cannot support any of the five anchors, return an evaluation_error explaining which dimension lacks evidence.

Return only the specified JSON object. Do not add prose or fields.

STAGE 2: USER PROMPT

Instruction: Replace the text '2019' with '2020'.

Rubric questions and anchors:
<the same six Question-and-Requirement-level sections shown above, without Focus lines>

First-stage reason JSON (the only visual evidence available):
<the complete reason JSON returned by the first-stage request>

Output:
Map each supplied reason to its scene-specific anchor. Return exactly one JSON object with only requirement_status; include every dimension exactly once. The reason is the entire visual evidence available to you. A label below fully_met requires a concrete corresponding defect in that reason. If a reason says requirements are fully met with no relevant defect, select fully_met.

```
{
 "requirement_status": {
  "Text Accuracy": "<fully_met|mostly_met|partially_met|minimally_met|not_met>",
  "Target Localization": "<fully_met|mostly_met|partially_met|minimally_met|not_met>",
  "Font Accuracy": "<fully_met|mostly_met|partially_met|minimally_met|not_met>",
  "Visual Naturalness": "<fully_met|mostly_met|partially_met|minimally_met|not_met>",
  "Non-target Preservation": "<fully_met|mostly_met|partially_met|minimally_met|
not_met>",
  "Layout Accuracy": "<fully_met|mostly_met|partially_met|minimally_met|not_met>"
 }
}
```

If a reason is insufficient or contradictory so no anchor can be justified, return only {"evaluation_error":{"type":"insufficient_evidence","reason":"<dimension and missing evidence>"}}

D.5 Human Evaluation

The validation set contains 128 outputs: 32 bilingual prompts spanning all 16 scenarios, each evaluated on four generation systems. Two evaluators with text-editing expertise and protocol training score every output with the same sample-specific rubrics and five-level anchors as the judges. Figure 13 shows the interface controls. Each evaluator's criterion ratings are aggregated separately; their Gate and overall scores are then averaged for judge comparison. Human–Human agreement compares the two evaluators' separately aggregated scores.

Table 2, Panel B reports agreement on the original 0–1 scale. These estimates apply to the sampled outputs and do not establish universal validity.

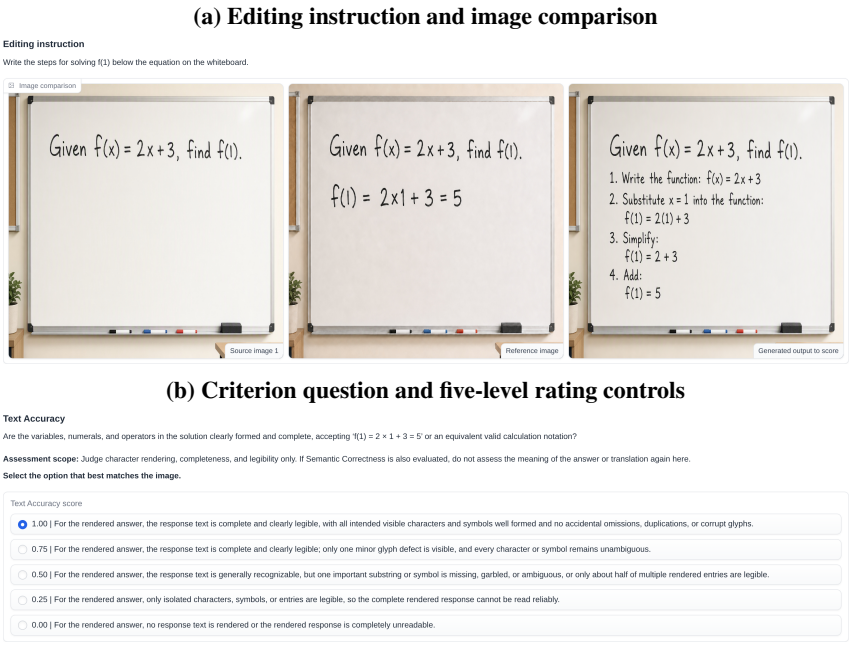


Figure 13: **Gradio interface for human evaluation. (a)** Whiteboard-editing instruction and source, reference, and generated images. **(b)** A criterion-specific question and five-level rating controls. This English-language reconstruction of the archived interface uses the same sample and recorded rating; it illustrates the controls rather than an original annotation session.

E Extended Quantitative Results

Table 8 reports bilingual-average Total scores for all 16 scenarios and all 21 leaderboard configurations, using the same per-sample scoring rule as Table 3. Scenario numbers follow Figure 1(a).

Table 8: **Performance across all 16 editing scenarios.** We report the Total score (0–100) averaged over Chinese and English examples, using the same model outputs as in Table 3. Missing outputs are assigned a score of zero. Bold numbers indicate the best performance in each scenario, and Thinking refers to the official thinking mode.

(a) Scenarios 1–8

Model	S1	S2	S3	S4	S5	S6	S7	S8
Proprietary models								
GPT-Image-2.0	**98.6**	98.5	97.6	**89.1**	**97.8**	**92.4**	**97.1**	**99.9**
Nano Banana Pro	97.8	**99.5**	98.2	88.0	96.0	89.6	81.1	95.4
Seedream 5.0	94.9	98.7	94.5	72.3	89.6	57.7	83.8	83.2
Unified multimodal models								
BAGEL	48.6	63.5	68.4	26.8	51.6	29.4	36.7	36.1
BAGEL (Thinking)	17.8	21.2	64.3	22.4	43.2	20.9	42.2	35.2
OmniGen2	26.3	18.8	59.8	17.5	33.2	8.8	24.6	24.3
InternVL-U	83.2	71.0	60.0	26.4	55.3	11.3	29.1	48.4
InternVL-U (Thinking)	81.9	73.9	61.7	22.5	60.3	14.5	39.2	48.0
UniWorld-V1	3.9	1.2	24.2	16.6	21.5	8.2	19.9	13.3
SenseNova-U1.5-8B-MOT	96.6	97.8	96.2	85.9	90.2	82.3	87.0	89.5
SenseNova-U1.5-8B-MOT (Thinking)	95.4	98.9	**98.5**	87.1	91.4	83.7	86.4	92.8
Boogu-Image-0.1	94.4	94.4	91.8	67.3	87.2	45.9	87.8	91.2
LLADA-Image	68.7	73.4	71.6	31.5	68.9	11.6	51.1	81.8
Specialized generation/editing models								
Qwen-Image-2.1	97.7	96.8	94.4	69.7	93.2	55.4	76.6	96.9
Qwen-Image-Edit-2511	88.4	94.5	83.4	51.5	87.9	31.3	60.4	81.9
FireRed-Image-Edit-1.1	96.7	95.6	92.3	74.0	87.3	47.1	67.0	69.7
LongCat-Image-Edit	93.0	91.7	92.8	43.6	77.4	35.8	55.1	62.2
GLM-Image	61.6	71.4	49.0	34.3	54.4	16.4	20.1	34.1
JoyAI-Image-Edit	94.1	94.0	92.4	65.9	88.3	52.3	64.7	86.7
FLUX.2-klein-9B	62.0	57.6	79.6	42.3	80.0	27.6	67.0	78.8
Step1x-Edit-v1p2	91.7	81.2	93.4	27.3	70.9	27.0	55.3	44.9

(b) Scenarios 9–16

Model	S9	S10	S11	S12	S13	S14	S15	S16
Proprietary models								
GPT-Image-2.0	**75.8**	86.2	71.9	65.0	70.9	93.1	87.7	**95.3**
Nano Banana Pro	72.5	**87.7**	**82.7**	**78.2**	81.7	**97.5**	**98.2**	86.4
Seedream 5.0	48.5	79.1	74.7	72.8	**82.0**	57.6	88.4	93.1
Unified multimodal models								
BAGEL	32.2	0.0	17.4	2.4	35.1	0.9	1.4	42.0
BAGEL (Thinking)	30.5	0.6	22.5	3.0	31.5	0.3	0.4	34.2
OmniGen2	26.5	0.4	11.1	4.1	24.4	0.3	0.2	26.0
InternVL-U	36.0	0.6	10.1	6.8	29.3	3.2	1.6	0.0
InternVL-U (Thinking)	41.3	13.3	19.8	16.2	34.1	2.9	14.0	0.0
UniWorld-V1	20.9	0.0	16.5	4.3	26.4	0.0	0.0	10.4
SenseNova-U1.5-8B-MOT	60.7	26.0	61.5	5.0	63.2	39.3	4.7	81.9
SenseNova-U1.5-8B-MOT (Thinking)	58.9	78.3	65.1	62.4	71.6	62.3	85.7	85.4
Boogu-Image-0.1	42.1	9.3	37.0	3.0	53.9	19.7	1.5	0.0
LLADA-Image	45.6	0.2	14.1	2.3	29.3	6.4	1.1	0.0
Specialized generation/editing models								
Qwen-Image-2.1	68.3	0.9	30.7	3.0	45.2	14.1	2.6	87.6
Qwen-Image-Edit-2511	40.6	2.0	35.6	2.9	49.5	8.2	0.0	53.1
FireRed-Image-Edit-1.1	49.3	0.0	46.8	4.8	50.2	9.1	1.0	69.1
LongCat-Image-Edit	32.8	1.7	42.7	5.5	44.0	5.7	3.2	0.0
GLM-Image	20.6	4.8	32.0	3.7	24.2	6.6	2.8	47.9
JoyAI-Image-Edit	34.1	1.3	37.3	6.8	49.1	22.5	2.8	40.0
FLUX.2-klein-9B	44.7	0.5	48.8	4.0	39.4	6.6	1.0	67.4
Step1x-Edit-v1p2	29.4	19.6	38.4	19.8	52.0	4.2	44.2	30.2

Scenario numbering follows Fig. 1(a). **S1**: Text replacement; **S2**: Text insertion; **S3**: Text removal; **S4**: Geometric transformation; **S5**: Path text insertion; **S6**: Document element adjustment; **S7**: Stylized text replacement; **S8**: Artistic text insertion; **S9**: Creative text fusion; **S10**: Scene-text translation; **S11**: Degraded text restoration; **S12**: Scene-text correction; **S13**: Referring text editing; **S14**: Form filling; **S15**: Knowledge-fused text editing; **S16**: Multi-reference composition.

Missing outputs count as zero: Multi-Reference has a 30.6% missing rate across 18 open configurations (less than 0.1% elsewhere), including all outputs from InternVL-U, Boogu-Image-0.1, and LLADA-Image. These fixed-test-set scores include output availability and editing quality.

E.1 Model Profiles across Evaluation Dimensions

Figure 14 compares all 21 configurations on the eight evaluation dimensions defined in Section 4. Each cell averages the applicable criterion ratings across Chinese and English samples with valid outputs. These dimension scores are distinct from the Gate-and-Craft leaderboard scores.

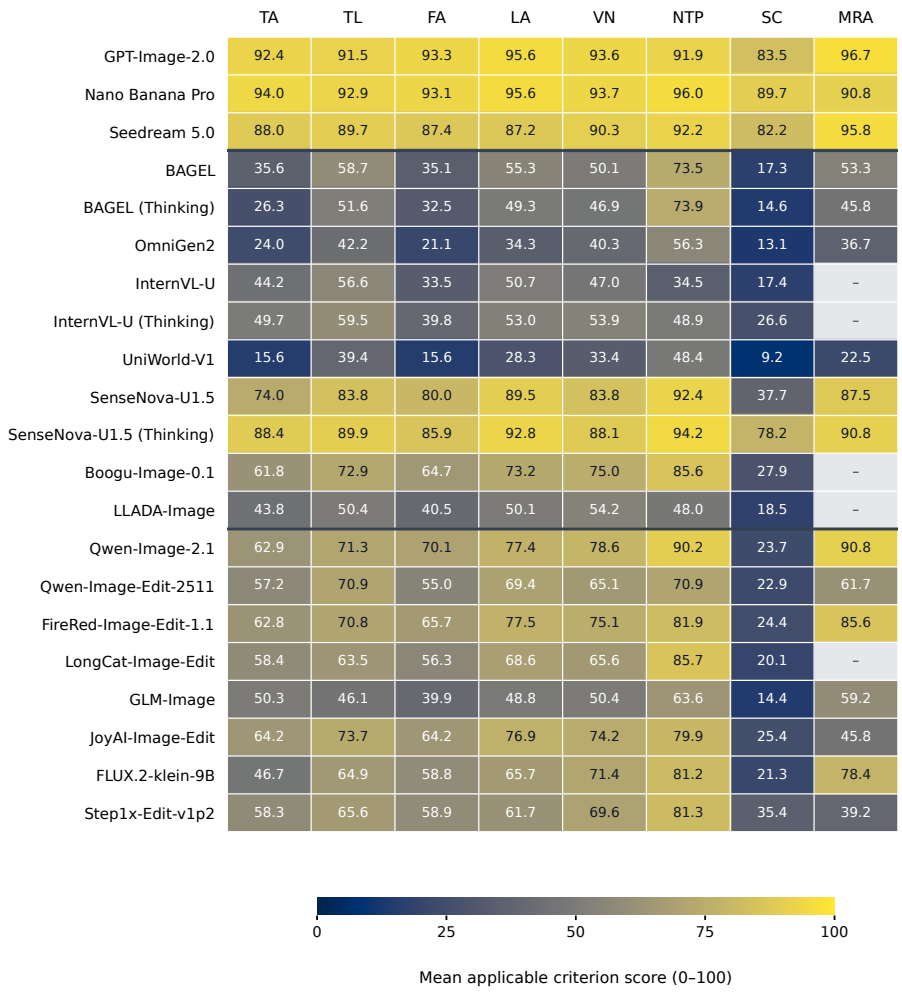


Figure 14: **Model profiles across eight evaluation dimensions on TextReVision.** Mean applicable criterion scores on a 0–100 scale for Text Accuracy (TA), Target Localization (TL), Font Accuracy (FA), Layout Accuracy (LA), Visual Naturalness (VN), Non-target Preservation (NTP), Semantic Correctness (SC), and Multi-Reference Alignment (MRA). Scores pool Chinese and English samples. Missing outputs are excluded from dimension means, as in Table 4; an en dash indicates no valid ratings, not a zero score.

The proprietary models score strongly across all eight dimensions, while SC is a marked weakness for many open configurations. SenseNova-U1.5-8B-MOT's thinking mode raises SC from 37.7 to 78.2 and TA from 74.0 to 88.4. Five configurations have no valid MRA ratings because they produce no valid outputs in the multi-reference scenario.

F LIMITATIONS AND RELEASE PLAN

Limitations. The benchmark covers 16 editing scenarios, while broader editing domains remain unexplored. Rubric construction involves human judgment in defining sample-specific evaluation requirements, and VLM-based evaluation may overlook fine-grained glyph or typography defects. Human validation on 128 outputs provides evidence only for the sampled systems and scenarios. Although references are used only as correct-edit examples, residual reference bias may remain for highly open-ended edits.

Release plan. We plan to release the benchmark metadata, instructions, all 1,200 rubrics, scoring guidelines, evaluation code, and per-sample results. Images and model outputs will also be released where redistribution is permitted. For assets that cannot be redistributed, we will provide source metadata and retrieval information where permitted.

G FULL PROMPTS FOR FIGURE 1

The following reproduces the complete editing instructions for the 16 examples in Figure 1(a), in the same numbered order. Instructions retain their original language and wording; image1 and image2 refer to the inputs shown in the multi-reference example.

(1) Basic Text Replacement. Replace the text '2019' with '2020'.

(2) Basic Text Insertion. Add "HERE" below "CANDIDATE" in the image, matching the size of "WE FOUND."

(3) Basic Text Removal. Remove the text 'SKYORA' from the image.

(4) Text Geometric Transformation. Rotate the text '2018' to be horizontal in the image.

(5) Path Text Insertion. Add white embroidered curved ring text to the empty dark blue ring band of the round embroidered patch on the shoulder, placing "RIVERSIDE RANGERS" on the top arc and "SINCE 1998" on the bottom arc, aligned to the ring with consistent kerning.

(6) Document Element-Level Text Adjustment. Format the chemical formulas in the text with proper subscripts and superscripts: change 'CO2' to 'CO_2', 'CH4' to 'CH_4', 'PO43-' to 'PO_4^{3-}', and 'H2SO4' to 'H_2SO_4'.

(7) Stylized Text Replacement. Change the neon electric blue lettering 'MOUNTAIN VIEW APPLE CIDER / PRESSED FRESH DAILY / SINCE 1921' to deep barn-red.

(8) Artistic Text Insertion.
在图片右上方空白处，用黑色书法字体竖向书写两列诗句：右列为'落霞与孤鹜齐飞'，左列为'秋水共长天一色'。

(9) Creative Text Fusion. Render several ants gather on a sandy structure with arched tunnels; conceal "Ants" in the original positions with arched sand tunnels, ant bodies, and brown sandy surface.

(10) Scene Text Translation. Translate all Chinese in the image into English.

(11) Degraded Text Restoration. Clean up the water stains and restore the water-damaged text in the center.

(12) Scene Text Correction. Fix the spelling error in the image.

(13) Referring Text Editing. Remove the number indicated by the minute hand.

(14) Form Filling.
完成“项目任务看板”中除左上角已有高优先级卡片之外的11张空白便签。任务尚未开始时归入“待处理”，已经开始但尚未验收时归入“进行中”，已经验收时归入“已完成”。优先级分数3、2、1分别转换为“高”“中”“低”；每栏按高到低排列，同分时保持下列记录的先后顺序。任务记录为：设计页面原型/完善交互流程—尚未开始，分数2；调研产品功能/整理参考资料—尚未开始，分数1；准备测试用例/编写测试数据—尚未开始，分数1；开发核心功能/实现关键逻辑—已经开始，分数3；搭建接口服务/联调相关接口—已经开始，分数2；设计数据库表/优化数据结构—已经开始，分数2；编写单元测试/保证代码质量—已经开始，分数1；项目计划制定/明确里程碑—已经验收，分数3；环境搭建完成/工具安装配置—已经验收，分数2；团队规范制定/编写开发规范—已经验收，分数1；创建项目仓库/初始化代码结构—已经验收，分数1。每张便签依次填写“优先级：”加推导结果和两行任务文字。彩色图钉视为不可修改的既有标记，不用于本次文字优先级判断。匹配现有手写字体、字号、颜色和对齐方式，并保持看板布局不变。

(15) Knowledge-Fused Text Editing. Based on the image, fill in the names of the three water-cycle processes indicated by arrows A, B, and C on the corresponding lines.

(16) Multi-reference Composition. Carve the middle of the mango in image1 into the 3D text ’MANGO’ from image2, retaining the mango’s texture and leaving the original mango ends on the sides.

H SCENE-SPECIFIC DIMENSION PROMPT TEMPLATE

All scene-specific dimension prompts follow the shared template below. For each selected scene–dimension pair, the prompt specifies the dimension scope, scenario-specific focus, and five scoring anchors.

Only dimensions selected in the sample rubric are scored; Each sample is evaluated only on the dimensions relevant to its editing task. The instruction and rubric question jointly define the requirements, while the anchors specify how different levels of satisfaction are scored. Each sample's Gate designation and Craft weights follow its scenario and applicable dimension set, as specified in Table 6.

SCENE-SPECIFIC DIMENSION PROMPT TEMPLATE

Sample instruction: [Editing instruction.]

Scenario: [Scenario used to select the corresponding prompt fields.]

Selected dimension: [Dimension name, sample-specific Gate designation, and weight.]

Boundary: [What this dimension evaluates and what belongs to other dimensions.]

Scene focus: [How this dimension applies to the current scenario.]

Question: [Concrete sample-specific requirement to assess.]

Scene-specific scoring anchors:

1.00: [Scenario- and dimension-specific description of full satisfaction.]

0.75: [Description of largely satisfied requirements with minor deviations.]

0.50: [Description of partial satisfaction with important remaining errors.]

0.25: [Description of limited satisfaction with major errors.]

0.00: [Description of an unmet requirement or failed edit.]

Bracketed text indicates template placeholders and is replaced by the corresponding scene-specific content during evaluation. For each selected criterion, the dimension boundary, scenario-specific focus, and five scoring anchors are filled accordingly, while the instruction, rubric question, Gate designation, and Craft weight are specified by the sample rubric. The block is repeated for each selected criterion. Appendix D.4 provides a complete user-prompt example, including the image inputs and required output format.