TRANSVIS: BENCHMARKING REAL-WORLD IN-IMAGE MACHINE TRANSLATION AT SCALE AND FINE GRANULARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In-image machine translation (IIMT) aims to translate the text inside an image while seamlessly preserving the original visual layout. Existing benchmarks either stop at textual translation or assess generated images with image-level metrics; even multi-aspect protocols do not trace each source text block through discovery, translation, text rendering and visual preservation. Consequently, omissions, mistranslations and visual damage remain difficult to attribute. We therefore present TRANSVIS-BENCH, a large-scale, fine-grained benchmark for real-world Chinese-to-English IIMT. The benchmark covers 5 major categories and 18 sub-categories of text–visual relations, comprising $2{,}591$ images and $19{,}453$ source text blocks. We design a source-block-anchored evaluation protocol that decouples model behaviour into four dimensions: translation coverage, translation quality, rendering accuracy and visual fidelity. An evaluation of 14 foundation image-generation models shows that a single aggregate score hides markedly different combinations of failure in text translation, text rendering and visual preservation. Controlled experiments further show that explicitly injecting linguistic information raises the translation score of an open-source backbone from 3.72 to 73.77, indicating that acquiring linguistic information remains a major limitation of current end-to-end translate-and-render models. TRANSVIS-BENCH provides a comprehensive testbed for real-world IIMT and supports auditable, fine-grained analysis of generated results.

1 INTRODUCTION

In-image machine translation (IIMT) converts source-language text inside an image into a target language and returns a reusable translated image. It supports cross-border content consumption, product localisation and digital-product internationalisation (Wu et al., 2026b;a; Qian et al., 2024). In packaging, interfaces and posters, a model must preserve reading order, visual hierarchy and context as well as meaning. Unified generation and editing models are shifting IIMT from an "OCR – translation – repainting" cascade (Cui et al., 2026; Ju et al., 2024) to direct image-to-image generation (Niu et al., 2024; Lan et al., 2024), coupling linguistic accuracy with visual realisation. This shift makes the output immediately useful, but also removes the clean module boundaries that once made recognition, translation and rendering errors relatively easy to isolate.

Real IIMT requires more than correct text recognition and translation: visual fidelity is equally important. Rendered text should remain consistent with surface perspective and material (Tuo et al., 2024), graphic designs need to accommodate cross-lingual length change (Hsu et al., 2023), and interface edits should respect component boundaries. Yet existing benchmarks remain limited in both data coverage and evaluation granularity. As Figure 1 illustrates, Translatotron-V (Lan et al., 2024) relies on synthetic paired scenes, PRIM (Tian et al., 2025a) translates only a local region, and WeEdit (Zhang et al., 2026) evaluates editing on generated images, underrepresenting the text densities, layouts and visual contexts found in real documents, captured scenes, social media, interfaces and graphic designs. Evaluation granularity is likewise limited: Translatotron-V and PRIM rely on conventional image-level visual metrics such as SSIM and FID (Wang et al., 2004; Heusel et al., 2017), which provide only a coarse view of visual quality, while WeEdit assesses visual effects only through background preservation. In contrast, TRANSVIS-BENCH covers diverse real-world

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.

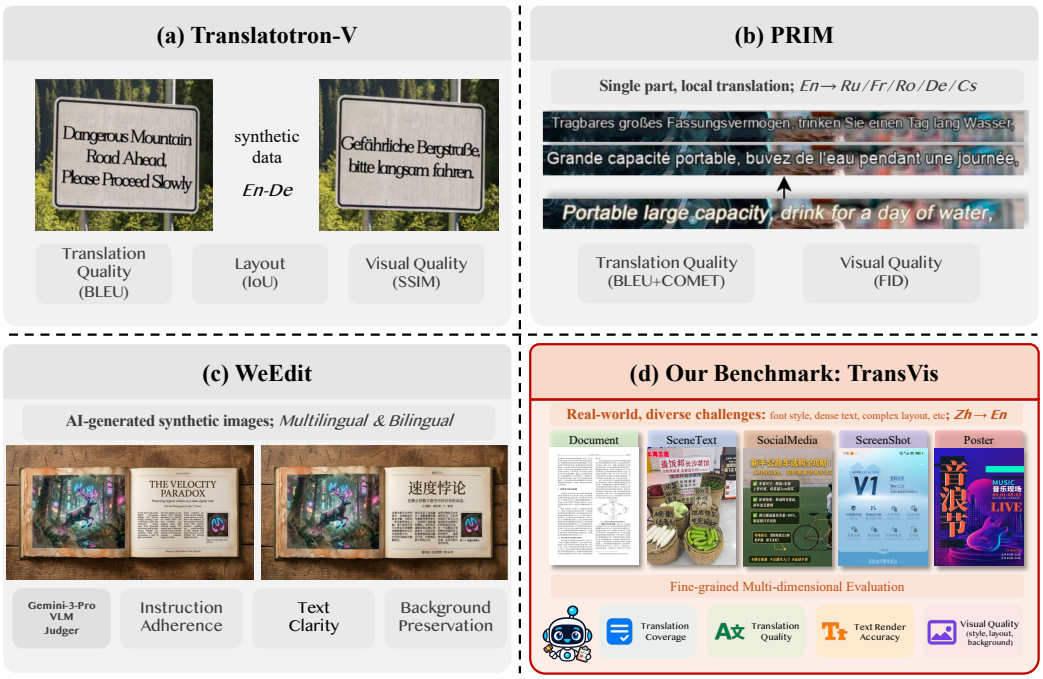


Figure 1: **Positioning of TRANSVIS-BENCH against existing benchmarks.** Prior benchmarks focus on synthetic data, local translation, or broad editing criteria. In contrast, TRANSVIS-BENCH covers diverse real-world Chinese-to-English scenarios and separately evaluates translation coverage, translation quality, text rendering and visual fidelity.

text–visual relations and separately diagnoses translation coverage, textual translation quality, text rendering and visual fidelity, further decomposing the latter into font style, layout plausibility and background preservation so that distinct failures are not conflated in one global score.

To fill this gap we propose TRANSVIS-BENCH, a real-world Chinese-to-English benchmark organised around text–visual relations rather than application names. It contains 2,591 images and 19,453 source blocks across 5 categories and 18 sub-scenarios. Its source-block-anchored evaluation has four layers: **coverage** checks whether each required block reaches the output; **translation** scores matched blocks; **rendering** measures character- and word-level legibility; and **visual fidelity** measures style, layout and background preservation. Coverage remains an independent first layer so omissions cannot be hidden by conditional scores. This design connects each score to a concrete source block while retaining whole-image checks for background damage, enabling a model's failure profile to be inspected rather than inferred from a single aggregate number.

We evaluate 14 foundation image-generation models and conduct controlled studies of end-to-end supervision and linguistic assistance. Similar aggregate scores conceal different coverage, translation, rendering and visual-fidelity profiles: Qwen-Image-3-Pro (Qwen Team, 2026b) and GPT-Image-2 (OpenAI, 2026b) score 84.43 and 83.08 overall, yet their translation-quality scores differ by 18.74 points, while GPT-Image-2 leads the rendering and visual metrics. Supplying linguistic information raises SenseNova-U1.5's (Diao et al., 2026a) translation score from 3.72 to 73.77 and Boogu-Image-0.1-Edit's (Chen et al., 2026) from 0.46 to 56.01, increasing their overall scores by 24.76 and 25.85 points, respectively. However, style and layout do not improve consistently. These results separate failures in acquiring target-language content from failures in executing it visually, supporting auditable comparison and targeted improvement.

Our main contributions are as follows.

- **We introduce TRANSVIS-BENCH, a comprehensive, large-scale real-world benchmark.** It is organised around text–visual relations into 5 categories and 18 sub-scenarios with 2,591 images and 19,453 text blocks, covering the visual constraints of IIMT.

- **We design a fine-grained, decoupled evaluation protocol.** The layered, text-block-anchored protocol separates coverage, translation, rendering and visual fidelity and uses coverage as an explicit gate, supporting evaluation from the pixel to the semantic level.
- **We reveal a language bottleneck in open-source end-to-end image editors.** Across 14 models and two controlled studies, the evaluated open-source systems show almost no text translation capability under coupled translate-and-render generation, leading to poor end-to-end performance.

2 THE TRANSVIS-BENCH BENCHMARK

2.1 TASK DEFINITION

Given image x containing Chinese text, a model outputs English image $\hat{x}$. From image alone, it must discover all translatable text, convert it into English and re-render at corresponding position. Existing English or other languages, numbers, formulas, URLs, logos, trademarks and non-text regions are preserved content. No human OCR, target translation or editing mask is provided, so the task tests complete image translation rather than replacement within a supplied box, unlike text rendering and editing settings with specified content or regions (Chen et al., 2024; Tai et al., 2025).

Coverage measures translation omissions: whether every source text block has a corresponding translated block in the output. Only blocks containing a translation are eligible to be matched; absent or untranslated blocks are counted as missing. **Translation** evaluates the semantic accuracy and completeness of the translated text within matched regions only; **rendering** measures glyph- and word-level legibility; and **visual fidelity** measures retention of font style, layout and non-target content. The four layers follow causal order: coverage is the prerequisite for the conditional translation score, yet remains separate. Thus, omissions cannot be hidden by high translation, rendering or visual scores based only on matches. TRANSVIS-BENCH defines IIMT as full generation constrained by coverage, translation, text rendering and visual fidelity, rather than producing a plausible image.

2.2 DATA CONSTRUCTION

TRANSVIS-BENCH does not simply combine different application names; it is constructed around how text and visual content constrain each other. Figure 2 summarises the flow from multi-source candidates to the released set: real collection, public benchmarks and controlled synthesis form a candidate pool; Qwen3-VL-235B-A22B-Instruct performs initial filtering and classification; and human reviewers remove unsuitable samples and correct erroneous labels. The retained samples form a unified Chinese-to-English task with block-level annotation. Here, "comprehensive" denotes different failure pressures that real IIMT may encounter, rather than image count alone.

Candidate collection. Self-collected photographs and screenshots, public document, poster and layout datasets, public base images, social-media content and controlled synthesis are combined into a candidate pool covering object text, interfaces, dense documents, graphic design, overlays, decorative typography and subtitles.

Model-assisted filtering and human review. Qwen3-VL-235B-A22B-Instruct (Team, 2025) first removes clearly unsuitable candidates and assigns preliminary labels under the predefined 5-category, 18-sub-scenario taxonomy. Human reviewers then inspect the filtered pool, discard remaining samples that do not satisfy the collection criteria and correct erroneous category or sub-scenario labels before the dataset is finalised.

2.3 DATA ANALYSIS

The 2,591 TRANSVIS-BENCH images are approximately balanced across five categories and 18 sub-scenarios (Appendix Table 6), but their difficulty does not come from repeated sampling; it comes from text density, block length and cross-lingual typographic pressure varying simultaneously. The set contains 19,453 source text blocks, averaging 7.51 blocks per image: screen interfaces average 17.51 blocks per image, concentrated in short labels and component boundaries, whereas the dense-document category has a mean single-block length of 52.4 source characters,

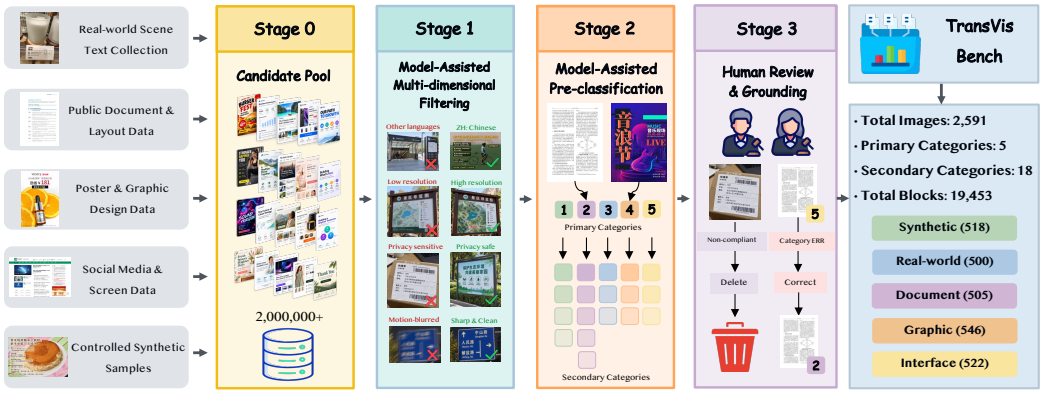
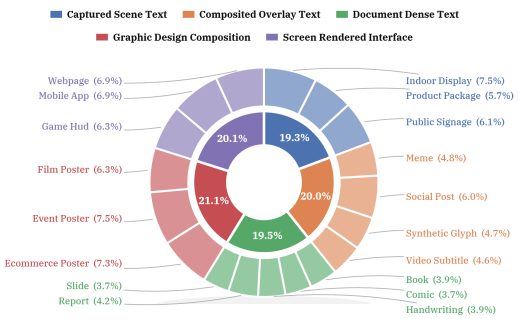
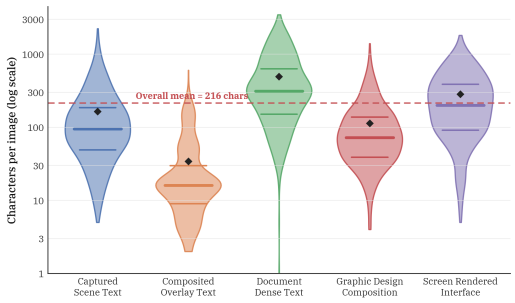


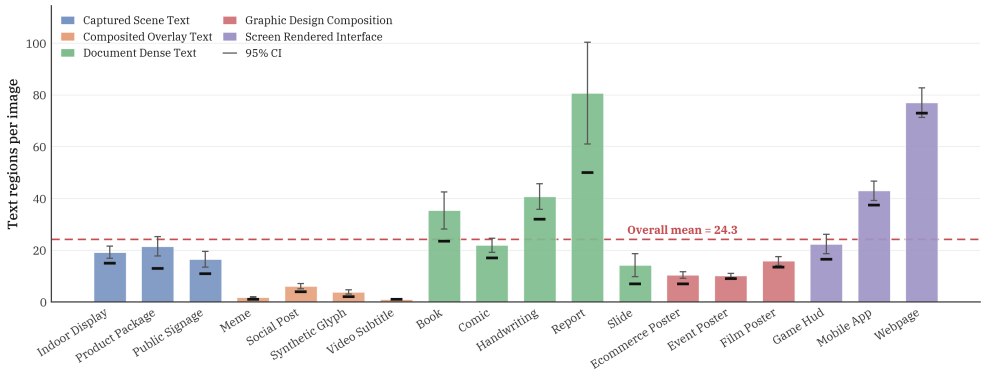
Figure 2: The data-construction pipeline of TRANSVIS-BENCH: multi-source candidates converge, pass initial filtering and category labelling by Qwen3-VL-235B-A22B-Instruct, and undergo human review to remove unsuitable samples and correct erroneous labels. The final set is sealed with metadata and block-level annotation. The released set contains 2,591 images over 5 categories and 18 sub-scenarios.



(a) Category and subcategory distribution.



(b) Characters per image by category.



(c) Text regions per image by subcategory.

Figure 3: Dataset composition and text density analysis in TRANSVIS-BENCH.

closer to paragraph-level reading and reflow. For alignable blocks the median EN/ZH character ratio is $3.31\times$ and the $p90$ is $5.0\times$, meaning that even a semantically correct translation may exhaust the original space budget. Figure 3 summarises the benchmark category composition alongside two OCR-based measures of text density at the image and subcategory levels.

These statistics directly motivate the need for diagnosis: high-density interfaces readily expose missed translations and correspondence failures, long text exposes legibility problems, and length

Table 1: Capability comparison of representative IIMT and text-centric editing benchmarks. *Discovery* denotes autonomous identification of all source text; ✓ / ✗ indicate whether a capability is explicitly evaluated, and "—" is not applicable.

Benchmark	Task setup			Diagnostic evaluation				Test set
	Output	Discovery	Scope	Coverage	Translation	Legibility	Visual fidelity	
MTIT6 (Qian et al., 2024)	Image	✓	6 pairs	✗	✓	OCR	Human/GPT	6,000 inst.
PRIM (Tian et al., 2025a)	Image	✓	1 scene; 6 lang.	✗	✓	OCR	FID	3,000 img.
PATIMT-Bench (Zhuang et al., 2025)	Text+boxes	✓	10 scen.; 2 tasks	✗	✓	—	—	1,200 img.
MMTIT-Bench (Li et al., 2026)	Text	✓	Multi-scen.; 14 lang.	✗	✓	—	—	1,400 img.
IMTBench (Lyu et al., 2026)	Image+text	✓	4 scen.; 9 lang.	✗	✓	Cross-modal	LPIPS/PQ	2,500 inst.
Vectra-Bench (Wu et al., 2026b)	Score	-	E-comm.; 5 targets	✗	✗	✓	14 defects	2,000 sets
MultiTextEdit (Cheng et al., 2026)	Image	✗	7 tasks; 12 lang.	✗	✗	✓	✓	3,600 inst.
WeEdit (Zhang et al., 2026)	Image	✗	8 tasks; 15 lang.	✗	1 task	✓	✓	4,000 cases
TRANSVIS-BENCH (ours)	**Image**	✓	**18 scen.; 5 cats.**	✓	✓	✓	**Style/layout/BG**	**2,591 img.; 19,453 blocks**

expansion raises the risk of out-of-bounds and layout failures. TRANSVIS-BENCH therefore does not compress text–visual relations into one average difficulty, but measures coverage, translation, rendering and visual fidelity separately in Section 3 to track the breaking point under each pressure.

Table 1 compares representative benchmarks by output form, autonomous text discovery, scenario coverage and diagnostic dimensions. Existing resources broaden languages or scenarios, or assess selected aspects of legibility and visual quality, but do not jointly measure whether every source block is translated and diagnose subsequent translation, rendering and visual-fidelity failures. TRANSVIS-BENCH combines 18 text–visual scenarios with source-block coverage and seven fine-grained score components.

3 EVALUATION FRAMEWORK

3.1 LAYERED EVALUATION AND AGGREGATION

TRANSVIS-BENCH evaluates coverage, translation, rendering and visual fidelity on a 0–100 scale.

Coverage (s_{match}) is the proportion of source blocks with a unique English block in the output; missing, untranslated or unmatchable blocks count as failures. **Translation quality** (s_{trans}) aggregates 0–3 judgements from a translation judge on matched blocks. Higher scores indicate more complete, semantically correct translations and must be interpreted jointly with s_{match}; this complements established corpus-level translation metrics such as BLEU and COMET (Papineni et al., 2002; Rei et al., 2020). **Rendering accuracy** (s_{char}, s_{word}) uses TextPecker OCR (Zhu et al., 2026) to inspect generated text. The character score measures the proportion of correctly recognised target characters, thereby penalising malformed or garbled glyphs, while the word score measures the proportion of standard words recognised correctly. **Visual fidelity** (s_{style}, s_{layout}, s_{bg}) measures, respectively, preservation of the original font and visual style, the positions and reading relations of text blocks, and whether the non-target background remains undamaged, separating these factors rather than collapsing them into conventional image-level similarity measures (Wang et al., 2004; Zhang et al., 2018; Heusel et al., 2017).

TRANSVIS-JUDGER supplies (s_{match}, s_{style}, s_{layout}, s_{bg}), CULTURE-MT (Wu et al., 2026a) (s_{trans}), and legibility evaluation (s_{char}, s_{word}), respectively:

$$S_{\text{final}} = \underbrace{0.15\, s_{\text{match}} + 0.25\, s_{\text{trans}}}_{\text{coverage and translation}} + \underbrace{0.15\,(s_{\text{char}} + s_{\text{word}})}_{\text{text rendering}} + \underbrace{0.10\,(s_{\text{style}} + s_{\text{layout}} + s_{\text{bg}})}_{\text{visual fidelity}}. \tag{1}$$

3.2 SOURCE-BLOCK ANCHORING AND CONDITIONAL METRICS

Evaluation is anchored on source blocks rather than output text that "looks fine." The judge finds Chinese text in reading order, excluding English, numbers, formulas, URLs and brands,

Table 2: Scoring-eligibility agreement (%, higher is better) and Cohen's κ of TRANSVIS-JUDGER and candidate visual judges on 700 validation pairs. Agreement checks whether automatic judges and humans select the same valid objects; κ checks whether its judgements agree with humans on the valid ones. Bold and underlining mark the best and second-best values per metric subcolumn.

Judge	Text coverage agree / κ	Style agree / κ	Layout agree / κ	Background agree / κ	Mean agree / κ
GPT-5.5 (teacher)	<u>90.7</u> / <u>0.812</u>	**76.3 / 0.461**	69.6 / **0.272**	72.9 / 0.453	**77.4 / 0.500**
TRANSVIS-JUDGER (ours)	87.7 / 0.751	<u>72.6</u> / <u>0.380</u>	**71.3** / <u>0.267</u>	**74.9** / <u>0.484</u>	<u>76.6</u> / <u>0.470</u>
Gemini-3.1-Pro	**91.3 / 0.826**	64.6 / 0.245	<u>70.6</u> / 0.221	<u>74.0</u> / **0.489**	75.1 / 0.445
Qwen3.6-27B-GPT-5.5 (unified)	86.9 / 0.733	72.0 / 0.370	70.3 / 0.225	69.7 / 0.425	74.7 / 0.438
Gemini-3.1-Flash	85.6 / 0.708	58.9 / 0.146	66.3 / 0.150	62.6 / 0.296	68.3 / 0.325
Qwen3.6-27B (base)	86.1 / 0.664	54.9 / 0.195	66.3 / 0.096	63.0 / 0.338	67.6 / 0.323

and matches all outputs using semantics, position and region. A corresponding English block is recorded as **matched**; any absent, untranslated or unmatchable block is marked as **missing**. Missing blocks stay in the coverage denominator; only matched blocks enter conditional translation and style/layout judgements, while rendering and background fidelity are scored separately over the image. This gate keeps untranslated content visible in all subsequent conditional scores.

3.3 VISUAL JUDGE VALIDATION

We validate the visual judge on 50 stratified TRANSVIS-BENCH images rendered by 14 models, yielding 700 source–output pairs. Five annotators label coverage, font style, layout and background preservation as Yes / No / NA, and their majority vote forms the human reference. Our open-source TRANSVIS-JUDGER uses Qwen3.6-27B supervised-fine-tuned on dimension-specific judgements produced by GPT-5.5 (Qwen Team, 2026a; OpenAI, 2026a). This model-as-a-judge design follows work on language and multimodal evaluators (Liu et al., 2023; Xiong et al.). To avoid coupling matching, font, layout and background decisions inside one prompt, we train four specialised judges for match, style, layout and background, collectively called TRANSVIS-JUDGER. Their training images are strictly separate from both the TRANSVIS-BENCH release set and 700 validation pairs. Translation is evaluated by the CULTURE-MT translation judge; the character- and word-level rendering evaluator is described in Appendix G.

Comprehensiveness and usage principles. With source-block anchoring, TRANSVIS-JUDGER determines what to evaluate, text discovery, whether font and layout were preserved, and whether the background was damaged; independent modules assess translation and rendering, preventing a visual score from replacing IIMT evaluation. Missing blocks receive NA for style and layout, not forced scores. We therefore report scoring-eligibility agreement and Cohen's κ (Cohen, 1960) to test whether correct objects are scored and whether valid judgements agree with humans. Prompts appear in Appendix D; validation and human-agreement details appear in Appendix E.

Results and deployment choice. TRANSVIS-JUDGER reaches 76.6% scoring-eligibility agreement and 0.470 κ, close to the GPT-5.5 teacher's 77.4% and 0.500. Its background agreement is highest at 74.9%, and its 0.484 κ approaches the best, 0.489. These results show strong judgement consistency and support TRANSVIS-JUDGER as a competitive, locally deployable alternative to powerful proprietary VLM judges. All main experiments use it to fix the evaluation protocol, avoid score drift from service updates and remove dependence on external judging APIs. We release the model, prompts, training setup and validation materials for local deployment, reuse and audit.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

On 2,591 TRANSVIS-BENCH images we compare 14 strictly end-to-end foundation image models: 6 proprietary image-to-image models and 8 open-source unified / editing models. All baselines

Table 3: TRANSVIS-BENCH end-to-end main results (0–100, higher is better). Coloured columns group translation (blue), text rendering (purple), and visual fidelity (green); *Char* and *Word* are the proportions of correctly rendered characters and words. *Final* is highlighted in yellow. Models are ranked within each group, and bold marks the group best.

Model	Final	Translation		Text rendering		Visual fidelity		
		Match	Quality	Char	Word	Style	Layout	BG
Proprietary Models								
Qwen-Image-3-Pro	**84.43**	95.85	**75.45**	97.56	73.38	91.52	89.95	62.95
GPT-Image-2	83.08	**98.83**	56.71	**98.81**	**80.28**	**93.13**	**95.85**	**75.66**
Nano Banana Pro	81.64	94.73	70.34	97.92	73.69	80.09	89.88	61.81
MAI-Image-2.6	80.57	97.92	66.54	95.79	69.43	91.16	87.99	49.38
Seedream-5.0-Pro	78.81	84.31	72.13	96.75	70.45	77.99	80.91	66.48
Nano Banana 2	77.39	88.11	66.45	97.88	72.84	73.47	81.90	55.12
Open-Source Models								
SenseNova-U1.5-8B-MoT	**54.10**	**97.72**	**3.72**	**91.59**	**22.05**	**74.33**	**89.43**	30.19
Boogu-Image-0.1-Edit	45.39	92.08	0.46	81.37	15.05	54.64	79.01	16.05
SenseNova-U1-8B-MoT	41.54	79.63	0.82	87.51	15.15	36.35	67.89	24.74
Qwen-Image-Edit	33.11	62.62	0.23	79.31	11.18	25.89	48.13	20.10
FireRed-Image-Edit-1.1	27.88	52.34	0.40	76.22	13.76	15.49	30.85	14.50
FLUX.2	24.21	26.11	0.54	84.84	10.82	17.03	19.66	23.55
LongCat-Image	19.00	0.94	2.95	84.41	11.51	0.48	0.63	**59.96**
BAGEL	10.84	0.51	0.00	52.73	13.25	0.20	0.23	13.66

receive only the source image and the same translation instruction, without block-level OCR or the target English translation; this ensures Table 3 compares complete delivery capability from the source image to translated image. Appendix G lists all model versions and all settings used. The proprietary systems include Qwen-Image-3, GPT-Image-2, Nano Banana, MAI-Image-2.6 and Seedream-5.0-Pro (Qwen Team, 2026b; OpenAI, 2026b; Google, 2026b; 2025; Microsoft, 2026; ByteDance, 2026); the open-source systems include SenseNova, Boogu, Qwen-Image, FireRed, FLUX.2 and LongCat (Diao et al., 2026a;b; Chen et al., 2026; Wu et al., 2025; Team et al., 2026; Labs, 2025; Team et al., 2025).

4.2 MAIN RESULTS

Table 3 compares models using the source image and one instruction. We rank proprietary and open-source models separately by S_{final}, while language-assisted settings appear in Section 5. The table measures end-to-end coverage, translation, rendering and visual fidelity, excluding linguistic inputs. Figure 4 shows two cases from all 14 models, making their translation and visual-error profiles directly comparable.

Complementary strengths among the top models. Qwen-Image-3-Pro ranks first overall (84.43) and achieves the best translation score (75.45). GPT-Image-2 follows at 83.08 but leads match, character, word, style, layout and background, showing stronger text rendering and visual preservation.

A large proprietary–open-source gap. Proprietary models substantially outperform open-source editors in translation accuracy and complete text rendering. SenseNova, Boogu and Qwen-Image-Edit retain reasonable match and character scores but obtain translation scores of only 3.72, 0.46 and 0.23, indicating that the evaluated end-to-end open-source systems have almost no reliable text-translation capability. This gap likely reflects their tightly coupled single-pass generation, whereas proprietary models appear to use agentic pipelines with prompt rewriting and intermediate language processing before visual generation. Section 5 tests this explanation on fixed editing backbones.



Figure 4: **Generation cases from the** 14 **end-to-end models.** Two examples compare the originals with outputs from 6 proprietary and 8 open-source models, highlighting differences in translation completeness, text legibility and visual preservation.

Table 4: End-to-end SFT scaling comparison. Each checkpoint occupies a row so training progress is not compressed into intervals; **bold** marks the best value within the group of a given backbone.

Model	Step	Seen	Final	Translate	Match	Char	Word	Style	Layout	BG
Boogu-Image-0.1-Edit	base	0	45.39	0.46	92.08	81.37	15.05	54.64	**79.01**	16.05
Boogu-Image-0.1-Edit + SFT	900	33k	48.52	0.98	**96.40**	83.36	17.95	69.69	73.70	23.53
Boogu-Image-0.1-Edit + SFT	1,800	66k	**49.52**	1.15	96.23	84.32	19.18	**73.52**	74.97	24.20
Boogu-Image-0.1-Edit + SFT	2,854	100k	49.34	**1.25**	93.61	**84.51**	**20.04**	73.21	74.55	**26.19**
SenseNova-U1.5-8B-MoT	base	0	**54.10**	3.72	97.72	**91.59**	22.05	74.33	**89.43**	**30.19**
SenseNova-U1.5-8B-MoT + SFT	4,000	33k	52.34	3.55	97.83	90.55	20.50	78.16	78.26	22.59
SenseNova-U1.5-8B-MoT + SFT	8,000	66k	52.47	4.42	97.94	90.28	**22.07**	77.13	77.02	22.05
SenseNova-U1.5-8B-MoT + SFT	11,416	100k	52.72	**4.64**	**98.14**	89.98	21.61	**80.29**	76.86	20.66

5 Diagnostic Analysis

5.1 Scale Effects of End-to-End Supervision

We fix the backbone, objective and end-to-end prompt and compare the base model with roughly 33k, 66k and 100k SFT checkpoints (Brooks et al., 2023; Henighan et al., 2020), testing whether more supervision under the same coupled objective removes the language bottleneck. All checkpoints use the same source image, instruction and target-image format, isolating scale from architecture or input changes. The conclusion is limited to this coupled objective rather than SFT with arbitrary data or designs. Because the data format and inference prompt remain fixed, improvements can be attributed to repeated exposure to paired examples rather than additional OCR, translations or region annotations.

End-to-end SFT has little effect: at roughly 100k samples, translation rises only from 0.46 to 1.25 for Boogu-Image-0.1-Edit and from 3.72 to 4.64 for SenseNova-U1.5-8B-MoT, while final scores improve little or decline. Boogu's peak comes mainly from style and background, while SenseNova's small translation gains are offset by lower visual scores. More supervision therefore redistributes performance without resolving the language bottleneck.

Table 5: Visual execution given linguistic information. *One-stage* supplies the source image and the unified instruction; *two-stage* supplies block-level OCR and English translations on the same backbone. Each condition occupies a row; **bold** marks the higher value within a backbone.

Backbone	Mode	Final	Translate	Match	Char	Word	Style	Layout	BG
SenseNova-U1.5-8B-MoT	One-stage	54.10	3.72	97.72	91.59	22.05	**74.33**	**89.43**	30.19
	Two-stage	**78.86**	**73.77**	**98.85**	**95.89**	**70.87**	62.02	89.17	**40.53**
Boogu-Image-0.1-Edit	One-stage	45.39	0.46	92.08	81.37	15.05	54.64	**79.01**	16.05
	Two-stage	**71.24**	**56.01**	**94.37**	**93.78**	**67.94**	**59.16**	78.58	**38.32**
Qwen-Image-Edit	One-stage	33.11	0.23	**62.62**	79.31	11.18	**25.89**	**48.13**	20.10
	Two-stage	**40.99**	**29.81**	40.45	**86.76**	**46.68**	22.62	27.07	**24.88**

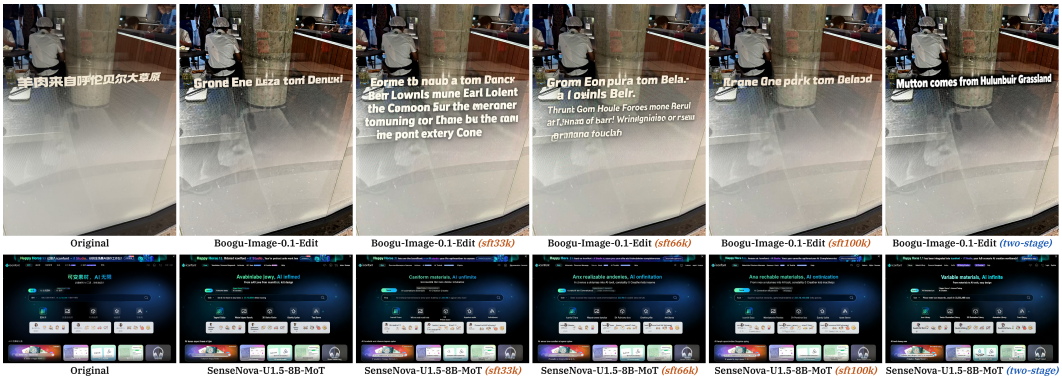


Figure 5: **SFT-scale and two-stage cases.** Rows for Boogu-Image-0.1-Edit and SenseNova-U1.5-8B-MoT compare the base with 33k/66k/100k SFT and language-assisted two-stage outputs.

5.2 Visual Execution under Language Conditioning

We compare the same backbone under two inputs: *one-stage* receives only the image and instruction, whereas *two-stage* moves translation upstream and supplies block-level OCR and English text from Gemini-3.1-Pro (Google, 2026a; Huang et al., 2024; Mou et al., 2024). Image constraints and evaluation remain identical; Table 5 and Figure 5 report the results. This separates missing linguistic information from failure to render known target text. Each Chinese block is paired with an automatically obtained English replacement, while preservation requirements remain unchanged; the setting is therefore practical rather than a human-gold upper bound. The comparison therefore tests whether the editor's main failure occurs before generation, when English content must be inferred, or during generation, when already known text must be placed and rendered.

Providing target text raises translation from 3.72 to 73.77 for SenseNova-U1.5-8B-MoT and from 0.46 to 56.01 for Boogu-Image-0.1-Edit, with final-score gains of 24.76 and 25.85. Their word scores also rise by 48.82 and 52.89 points, showing that explicit target text helps the models render complete English words. Translation should therefore be performed upstream and supplied to the image generator rather than learned implicitly in the generation pass. The two-stage design substantially improves generation, although correspondence, typesetting, layout and background preservation remain visual challenges.

6 Conclusion

TransVis-Bench provides a Chinese-to-English IIMT benchmark of 2,591 images spanning five text–visual relations and 18 sub-scenarios. Its block-anchored protocol separates coverage, translation, rendering and visual fidelity. Evaluating 14 models reveals capability profiles despite similar final scores. Explicit language input improves translation, but cannot fully resolve rendering, layout or background errors. The benchmark supports auditable comparison and targeted IIMT improvement.

AI USE STATEMENT

We used generative AI tools to support code development, figure preparation and language editing during this project. The authors inspected all generated text, code, plots and experimental outputs, verified factual claims and numerical results against the underlying records, and remain fully responsible for the manuscript and released artifacts.

ETHICS STATEMENT

TRANSVIS-BENCH combines self-collected images, screenshots, public research datasets, public online content and controlled synthesis. Public-source samples are included only when compatible with research use and redistribution. During human review, we remove identifiable personal information, sensitive content, obvious watermarks and candidates presenting privacy or copyright risk. Restricted raw material and samples that fail these checks are excluded from the release.

REPRODUCIBILITY STATEMENT

The main paper and appendices document benchmark construction, source-block annotation, layered metrics, model settings, SFT schedules, judge validation and both one- and two-stage prompts. Our release includes the $2{,}591$ benchmark images with metadata and $19{,}453$ block-level annotations, auxiliary OCR and translations used in the controlled study, TRANSVIS-JUDGER with its prompts and training setup, validation materials, evaluation records, and code for computing all reported metrics.

REFERENCES

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. Seedream 5.0 pro. Online, 2026. URL `https://seed.bytedance.com/en/seedream5_0_pro`. Accessed: 2026-09-23.

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

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.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M 6 doc: a large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15138–15147. IEEE, 2023.

Liwei Cheng, Shibo Feng, Lunjie Zhou, Yixuan Guan, and Dayan Guan. Multitextedit: Benchmarking cross-lingual degradation in text-in-image editing. *arXiv preprint arXiv:2605.08163*, 2026.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Cheng Cui, Yubo Zhang, Ting Sun, Xueqing Wang, Hongen Liu, Manhui Lin, Yue Zhang, Tingquan Gao, Changda Zhou, Jiaxuan Liu, et al. Pp-ocrv5: A specialized 5m-parameter model rivaling billion-parameter vision-language models on ocr tasks. *arXiv preprint arXiv:2603.24373*, 2026.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, et al. Sensenova-u1. 5: Towards native unified visual intelligence. *arXiv preprint arXiv:2609.11929*, 2026a.

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*, 2026b.

Matthieu Futeral, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. Towards zero-shot multimodal machine translation, 2024. URL https://arxiv.org/abs/2407.13579.

Google. Introducing nano banana pro. Online, 2025. URL https://blog.google/innovation-and-ai/products/nano-banana-pro/. Accessed: 2026-09-23.

Google. Gemini 3.1 pro: A smarter model for your most complex tasks. Online, 2026a. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/. Accessed: 2026-09-23.

Google. Nano banana 2: Combining pro capabilities with lightning-fast speed. Online, 2026b. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/. Accessed: 2026-09-23.

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.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6018–6026, 2023.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371. IEEE, 2024.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pp. 150–168. Springer, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

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

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, Min Zhang, and Jinsong Su. Translatotron-v (ison): An end-to-end model for in-image machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5472–5485, 2024.

Gengluo Li, Chengquan Zhang, Yupu Liang, Huawen Shen, Yaping Zhang, Pengyuan Lyu, Weinong Wang, Xingyu Wan, Gangyan Zeng, Han Hu, et al. Mmtit-bench: A multilingual and multi-scenario benchmark with cognition-perception-reasoning guided text-image machine translation. *arXiv preprint arXiv:2603.23896*, 2026.

Hanxi Li. Chaystack: Benchmarking chinese document retrieval and vqa. *arXiv preprint arXiv:2607.24760*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Google LLC. Open images dataset. `https://storage.googleapis.com/openimages/web/index.html`, 2020. Accessed: 2020-09-01.

Jiahao Lyu, Pei Fu, Zhenhang Li, Weichao Zeng, Shaojie Zhang, Jiahui Yang, Can Ma, Yu Zhou, Zhenbo Luo, and Jian Luan. Imtbench: A multi-scenario cross-modal collaborative evaluation benchmark for in-image machine translation. *arXiv preprint arXiv:2603.10495*, 2026.

Elman Mansimov, Mitchell Stern, Mia Xu Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. Towards end-to-end in-image neural machine translation. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pp. 70–74, 2020.

Microsoft. Mai image 2.6 generator: Features, access and arena rank. Online, 2026. URL `https://mai-image2.com/mai-image-2-6`. Accessed: 2026-09-23.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8488–8497. IEEE, 2024.

Liqiang Niu, Fandong Meng, and Jie Zhou. Umtit: Unifying recognition, translation, and generation for multimodal text image translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16953–16972, 2024.

OpenAI. Introducing gpt-5.5. Online, 2026a. URL `https://openai.com/index/introducing-gpt-5-5/`. Accessed: 2026-09-23.

OpenAI. Introducing chatgpt images 2.0. Online, 2026b. URL `https://openai.com/index/introducing-chatgpt-images-2-0`. Accessed: 2026-09-23.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24838–24848. IEEE, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Zhipeng Qian, Pei Zhang, Baosong Yang, Kai Fan, Yiwei Ma, Derek F Wong, Xiaoshuai Sun, and Rongrong Ji. Anytrans: Translate anytext in the image with large scale models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2432–2444, 2024.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026a. URL `https://qwen.ai/blog?id=qwen3.6-27b`.

Qwen Team. Qwen-image-3.0: Rich content, authentic details, deep knowledge. Online, 2026b. URL `https://qwen.ai/blog?id=qwen-image-3.0`. Accessed: 2026-09-23.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pp. 2685–2702, 2020.

Ying Tai, Nikai Du, Rui Xie, Zhennan Chen, Qian Wang, Zhengkai Jiang, Kai Zhang, and Jian Yang. Investigating text insulation and attention mechanisms for complex visual text generation. *arXiv preprint arXiv:2503.23461*, 2025.

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.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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.

Yanzhi Tian, Zeming Liu, Zhengyang Liu, Chong Feng, Xin Li, He-Yan Huang, and Yuhang Guo. Prim: Towards practical in-image multilingual machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13693–13708, 2025a.

Yanzhi Tian, Zeming Liu, Zhengyang Liu, and Yuhang Guo. Exploring in-image machine translation with real-world background. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 124–137, 2025b.

Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Linjuan Wu, Ruiqi Zhang, Xinze Lyu, Ye Guo, Daoxin Zhang, Zhe Xu, Yao Hu, Yixin Cao, Yongliang Shen, and Weiming Lu. Beyond literal translation: Evaluating cultural effectiveness in social media ugc. *arXiv preprint arXiv:2605.25626*, 2026a.

Qingyu Wu, Yuxuan Han, Haijun Li, Zhao Xu, Jianshan Zhao, Xu Jin, Longyue Wang, and Weihua Luo. Vectra: A new metric, dataset, and model for visual quality assessment in e-commerce in-image machine translation. *arXiv preprint arXiv:2602.07014*, 2026b.

Yiping Xie, Song Chen, Jingxuan Xing, Wei Jiang, Zekun Zhu, Yingyao Wang, Pi Bu, Jun Song, Yuning Jiang, and Bo Zheng. Secagent: Efficient mobile gui agent with semantic context. *arXiv preprint arXiv:2603.08533*, 2026.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models, 2025. *URL https://arxiv.org/abs/2410.02712*.

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.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Composition-aware graphic layout gan for visual-textual presentation designs. *arXiv preprint arXiv:2205.00303*, 2022.

Hanshen Zhu, Yuliang Liu, Xuecheng Wu, An-Lan Wang, Hao Feng, Dingkang Yang, Chao Feng, Can Huang, Jingqun Tang, and Xiang Bai. Textpecker: Rewarding structural anomaly quantification for enhancing visual text rendering. *arXiv preprint arXiv:2602.20903*, 2026.

Wanru Zhuang, Wenbo Li, Zhibin Lan, Xu Han, Peng Li, and Jinsong Su. Patimt-bench: A multi-scenario benchmark for position-aware text image machine translation in large vision-language models. *arXiv preprint arXiv:2509.12278*, 2025.

A Related Work

A.1 In-Image Machine Translation

IIMT requires a model to recognise source text, translate it, and render the translation back into the image. Early work studied end-to-end and context-rich multimodal translation (Mansimov et al., 2020; Futeral et al., 2024); UMTIT unified recognition, translation, and generation (Niu et al., 2024), while Translatotron-V introduced discrete visual tokens and a location-aware metric for multi-line and oriented text (Lan et al., 2024). Later work improved multilinguality or visual realism: AnyTrans released MTIT6 (Qian et al., 2024), DebackX placed text over real-world backgrounds (Tian et al., 2025b), and PRIM collected real-world one-line text images (Tian et al., 2025a). These datasets nevertheless remain dominated by sparse or localised text. TransVis-Bench instead covers multiple text blocks and category-specific constraints across natural surfaces, overlays, documents, designs, and interfaces.

A.2 IIMT Benchmarks and Evaluation

Recent benchmarks extend IIMT along complementary axes. PATIMT-Bench (Zhuang et al., 2025) evaluates translation and grounding but outputs text and boxes rather than translated images, whereas MMTIT-Bench (Li et al., 2026) broadens image-to-text evaluation across languages and scenarios. For image-output models, IMTBench (Lyu et al., 2026) combines translation, background, perceptual-quality, and cross-modal metrics over four scenarios; Vectra (Wu et al., 2026b) instead diagnoses visual defects in e-commerce IIMT outputs. These resources advance grounded translation, multilingual breadth, or image-level visual assessment. TransVis-Bench is distinguished by source-block anchoring: it records whether each required block is translated before separately scoring translation, text rendering, style, layout, and background.

A.3 Text-Centric Image Editing

Text-centric editing benchmarks assess specified manipulations while requiring the remaining image to be preserved. MultiTextEdit (Cheng et al., 2026) studies cross-lingual degradation, TextEditBench (Gui et al., 2025) adds geometric and semantic reasoning, and WeEdit (Zhang et al., 2026) covers eight operations with separate instruction, clarity, and preservation scores. Glyph-controlled and diffusion-based methods further illustrate the importance of explicit text placement and rendering (Yang et al., 2023; Chen et al., 2023). Unlike these instruction-driven tasks, IIMT must infer the complete set of source-language text from the image alone, translate and re-render it, and preserve all non-target content. This makes omission an independent failure rather than merely imperfect instruction following.

B Extended Results and Error Records

This appendix describes the *recomputable fields and exception-handling rules* that accompany the end-to-end main results. Every model is released with an image-level summary and block-level records: the former reproduces the eight metrics of Table 3, the latter explains whether a given score was caused by a missed translation, illegible text, a semantic deviation or visual destruction. We describe below, in turn, how coverage, translation, rendering, visual preservation and abnormal states are recorded and aggregated.

Coverage and correspondence. For each model we record the total number of to-be-translated blocks, the number of valid matched blocks and the number of missing blocks, and retain for each block the source text, source position, target correspondence and actual transcription. Together these determine the denominator of match and localise failures caused by missed translation, unreplaced Chinese or unestablished correspondence.

Translation and rendering. For each valid block we store the 0–3 translation-quality judgement and the grade distribution; for the whole image and for text blocks we simultaneously store the character-level and word-level matching records derived from OCR. The former recomputes trans-

late and separates semantic errors, partial translations and acceptable translations; the latter recomputes char / word and localises garbling, malformed characters and unstable rendering.

Aggregation and exception handling. Coverage uses all to-be-translated blocks as its denominator; translation is computed on matched blocks only, but the number of valid blocks is disclosed alongside it; style and layout record NA for missing blocks and never count NA as a visual pass; background preservation is computed per whole image. Empty outputs, unparsable outputs and outputs with no scorable text are not deleted but enter the corresponding denominator with an explicit status or are counted separately. As a result, any model score in Table 3 can be traced back to the same set of block-level and image-level judgements, and no failure is hidden by conditional filtering.

Figure 6 provides additional qualitative results for all 14 evaluated models across a broader range of text densities, layouts and visual styles.

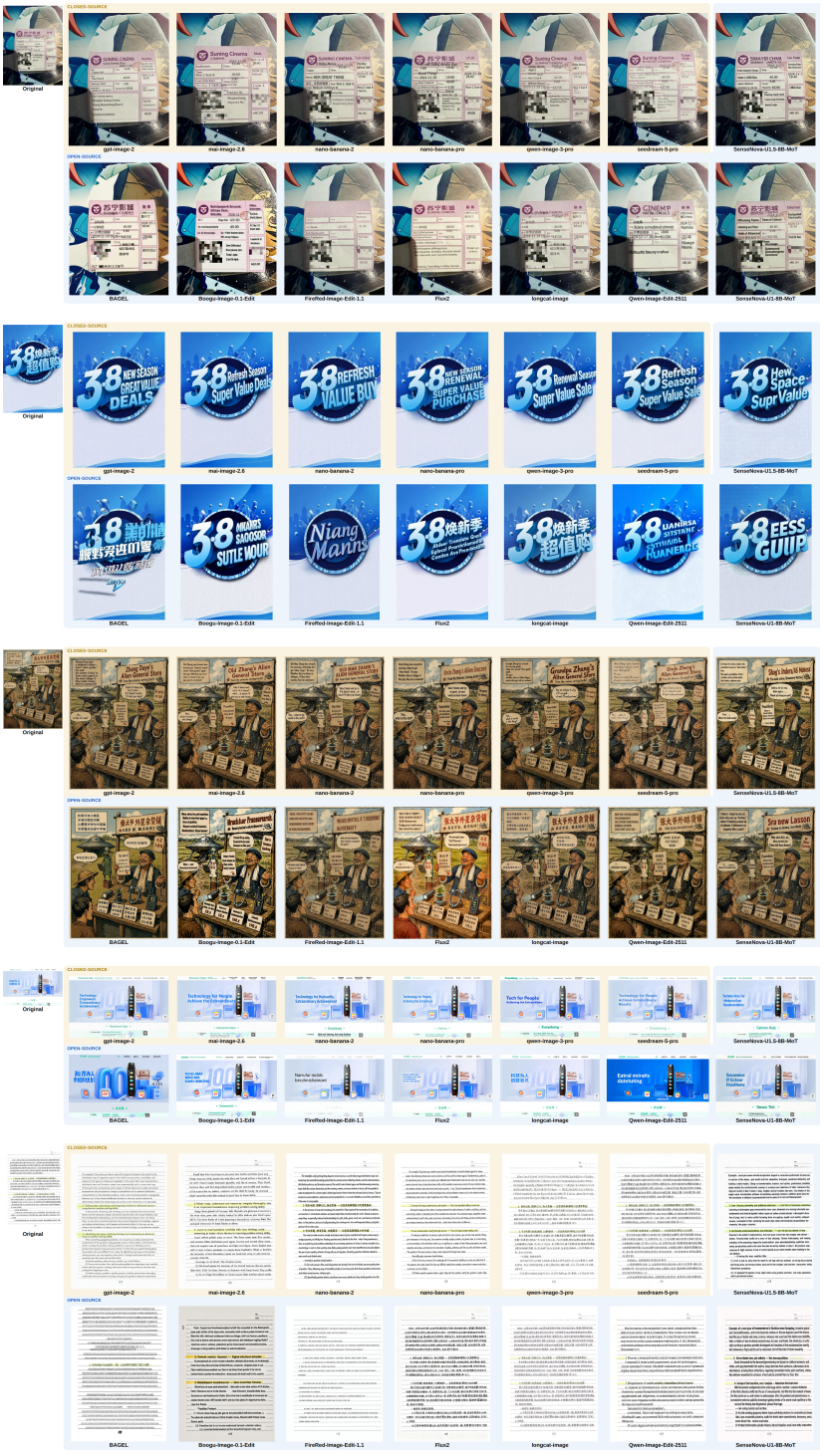


Figure 6: **Additional qualitative comparison of the** 14 **end-to-end models.** Five representative examples compare the original images with outputs from closed- and open-source systems across tickets, promotional graphics, illustrated designs, interfaces and dense documents.

C LIMITATIONS

Scope of languages and writing systems. TRANSVIS-BENCH focuses on Chinese to English in order to fix the language direction, invest the data budget in real scenes, text density and visual constraints, and reduce the confound of differing resource levels across language pairs. This choice does not cover other writing systems or linguistic phenomena—for example the complex typesetting of Arabic or Devanagari, or the reverse length compression of English to Chinese. The language breaking points and typographic pressures observed between Chinese and English therefore cannot be extrapolated directly into general laws of cross-lingual IIMT. Future extensions should incorporate language pairs with different writing directions, morphological variation and word order, while retaining the same layered evaluation chain.

Error bounds of the evaluation and of visually hard cases. Very small font sizes, artistic type, strong perspective, low-contrast text and complex occlusion simultaneously increase the difficulty of OCR, cross-image block correspondence and visual judgement. TRANSVIS-JUDGER reduces the risk of mistaking a missed translation for visual success through per-dimension judgement and human validation, but automatic judgement is still not an error-free substitute for human preference. In particular, for long-text reflow or layouts with ambiguous boundaries the consistency of human annotation itself is low. We therefore retain the layered results over coverage, translation, rendering and visual dimensions, and treat automatic scores as auditable diagnostic signals rather than as the sole arbiter of quality.

Extrapolation range of the controlled diagnosis. The two-stage experiment supplies obtainable OCR and English translations in order to distinguish missing linguistic information from subsequent visual-execution failure; these auxiliary inputs are not a human gold standard, so the result is not an ideal upper bound for the image-editing backbones. The SFT comparison likewise covers only two open-source diffusion editing backbones, a fixed training set and the present compute budget. It supports the scoped claim that "under this setting, adding more end-to-end supervision of the same kind has not yet stably crossed the language bottleneck", but it cannot establish that SFT at any scale, with any data recipe or on any architecture is ineffective. The direction of decoupling language understanding, region localisation and rendering still needs further examination at larger training scale and across more architectures.

D VISUAL JUDGING PROTOCOL

The visual evaluation of TRANSVIS-BENCH splits the complete judgement into four items: text matching, font style, layout and background preservation. The split does not change the object of evaluation—every dimension compares the same pair of source and translated image; it merely decomposes a multi-step visual inference that would otherwise have to be performed at once into shorter, independently auditable decisions. Text matching first determines whether each to-be-translated Chinese block has a unique English counterpart in the output image; font style and layout are judged on matched blocks only; and background preservation is judged at the whole-image level, asking whether non-target content was accidentally altered.

Inputs, gating and training isolation. The judge receives the source image and the translated image. It first identifies the Chinese natural-language text blocks of the source in reading order, excluding existing English, numbers, formulas, URLs and brand marks; it then establishes a cross-image correspondence through semantics, position, carrying region and visible style. If no reliable English counterpart exists in the output, or the original Chinese is still untranslated, the block is recorded as *missing*. Only matched blocks enter the font-style and layout judgements, which prevents "translation not completed" from being misread as visual success. The training images of the visual judges are strictly separated from the TRANSVIS-BENCH release set and from the human-validation pairs; the human annotations are used only to assess agreement with human judgement.

Correspondence with the scores of Section 3.1. The visual judgements in this appendix produce s_{match}, s_{style}, s_{layout} and s_{bg}. s_{match} is the proportion of all to-be-translated source blocks that are uniquely corresponded to an English output block; missing, still-Chinese and unreliably corresponded blocks all count as failures. s_{style} and s_{layout} are aggregated over matched blocks only:

the former checks whether font, colour, outline/stroke and clearly discernible decorative effects are continued, the latter whether position, alignment, orientation, the boundary of the carrying region and the relation to neighbouring elements remain reasonable. s_{bg} checks at the whole-image level whether non-target background and content were accidentally altered. The complete evaluation additionally includes s_{trans}, s_{char} and s_{word} from independent modules: s_{trans} is aggregated from 0–3 translation-quality judgements on matched blocks, and s_{char} and s_{word} from character- and word-level matches between the OCR transcription and the target translation. All scores are on a 0–100 scale, higher is better; together they form the seven scored components of Section 3.1, but they do not replace the joint reading of the four layers.

D.1 Judging Flow and Full Prompts

TransVis-Judger completes its decisions in the conditional order **match** → **style / layout** → **background**. Match first discovers the to-be-translated Chinese blocks and establishes the cross-image correspondence; style and layout receive that correspondence and judge matched blocks only; background judges at the whole-image level whether non-target content is preserved. Below we give the full prompts and output schemas actually used, in the fixed format "judging module | input | output", so that this conditional dependency can be reproduced without splitting each prompt into an extra subsection. The cards reproduce the prompts verbatim; only the Markdown emphasis of the originals is realised as typographic emphasis, and the output schemas are set in a monospaced block.

Match | input: source and translated image | output: to-be-translated blocks, cross-image correspondence, visible transcription and matched / missing.

Match judge — full prompt

Role and Task

You are a professional in-image translation quality assessment expert, skilled in scene text recognition, cross-lingual text correspondence, text visual analysis, layout analysis, and image-editing quality assessment. Given Image 1 (the original image) and Image 2 (the translated image), you must identify the text blocks to be translated in Image 1, and establish the correspondence between the source text blocks and the text regions in Image 2.

Assessment Setup

Translation task: This task is whole-image Chinese→English translation. The unit of translation is the text block: any text block containing natural-language Chinese content is a text block to be translated, and it is translated and re-rendered as a whole; text blocks containing no Chinese at all (text in other languages, pure numbers, formulas, URLs, and brand logos / registered trademarks) are preserved content and must be kept unchanged.

Assessment procedure: source text identification → text matching.

1. Source Text Identification

Goal: Identify all text blocks containing natural-language Chinese content in Image 1. Treat each text block as an independent assessment unit, assign it a unique `block_id` in normal reading order, and output `source_text` and `source_bbox` (`[x1, y1, x2, y2]`). `source_text` strictly transcribes the actual visible text of that block, without correction, completion, or rewriting.

Text-block segmentation principle: Group into one text block the text regions that are visually contiguous and form a unified whole in position and layout, judging the boundary together with content completeness; text regions that are independent of one another are segmented into separate text blocks.

2. Text Matching

Goal: For each identified source text block, find in Image 2 the text region produced by translating or editing it, and determine the match status as `matched` or `missing`. This dimension only establishes the correspondence between the source text block and the Image 2 text region; text visual style and layout are assessed separately by `style_consistency` and `layout_consistency`.

Judged as matched: Image 2 contains a text region that (a) can be reliably and uniquely corresponded to the source text block, AND (b) is the product of actually translating/re-rendering the source text into

the target language (English). Both conditions must hold. The correspondence in (a) may be established from any of the following cues, and not all need be satisfied simultaneously:

- The candidate text and the source text have a recognizable semantic association; this association may include partial translation errors, spelling errors, or expression deviations;
- The candidate text and the source text are similar enough in text color, outline/stroke, overall font style, and other features to confirm the correspondence;
- The candidate text's position, its carrying region, and its relationship to surrounding text or graphic elements clearly correspond to the source text.

When the candidate text's position or visual style has changed, it is still judged as matched as long as a reliable correspondence can be established from the other cues.

Untranslated text is NOT matched — it is missing: The translation task requires the Chinese source text block to be replaced by target-language (English) text. If the region in Image 2 that corresponds to the source block still shows the original (unchanged) Chinese content — i.e. no translation has actually occurred — this MUST be judged as **missing**, even though a region clearly corresponds by position, style, and content. The rationale: matched means the source block was successfully translated/edited; an untranslated block means the required translation never happened, so it counts as missing (not matched). Specifically:

- If the corresponding region in Image 2 still displays the same Chinese natural-language text as the source (e.g. source ‘‘你好” and Image 2 still shows ‘‘你好”), judge as **missing**, regardless of how well position/style/content correspond.
- If the corresponding region in Image 2 still shows the original Chinese text merely restyled, repositioned, resized, or partially re-rendered, but the natural-language content remains Chinese and was not converted into the target language, judge as **missing**.
- Only judge as matched when the corresponding region has actually been converted into target-language (English) text (even if that English contains translation errors, spelling errors, or expression deviations). A region that mixes some translated English with leftover original Chinese from the same source block is a defective/incomplete translation: still judge the block as matched (so that `style_consistency`, `layout_consistency`, and the transcription of what is actually rendered can be assessed), and transcribe the leftover Chinese faithfully in `recognized_text`.

Judged as missing: No text region in Image 2 can be reliably corresponded to the source text block, OR the corresponding region still shows the untranslated original Chinese content as described above.

Match output: For a matched block, output the position of the corresponding text region in Image 2 as `target_bbox` (`[x1, y1, x2, y2]`), and transcribe the actual visible text of that region as `recognized_text` (multi-line text is merged into a single string in normal reading order); for a missing block, set `target_bbox` to `null` and `recognized_text` to an empty string. The `recognized_text` of a matched block is transcribed under the following rules:

- **Transcribe faithfully, without correction:** Transcribe strictly according to the text actually rendered in Image 2, without correcting, completing, or rewriting it based on `source_text` or the expected translation; even if there are spelling errors or non-existent words, keep them exactly as seen.
- **Replace illegible text with placeholders:** For text that is illegible due to severely broken glyph structure (extra strokes, missing strokes, distortion, etc.), use a placeholder — use <#> to replace a single illegible Chinese character or single English letter, and use <###> to replace a single illegible English word (<###> is used only for a single word, not for phrases, whole lines, or sentences); legible text is still transcribed faithfully, and placeholders only replace the parts that are genuinely illegible.

Output Format

Output valid JSON only, without Markdown code fences, without any content outside the JSON, and without adding or removing fields. `text_blocks` lists all text blocks in `block_id` order; both matched and missing blocks output the same field structure. Each `analysis` first states the specific comparison evidence, then gives a clear conclusion in one sentence, and the conclusion must be consistent with the corresponding `status` value.

Example output

```
{
  "text_blocks": [
    {
      "block_id": 1,
```

```
      "source_text": "the actual text of a to-be-translated block identified in Image
1",
      "source_bbox": [x1, y1, x2, y2],
      "match": {
        "status": "matched",
        "recognized_text": "the corresponding text actually recognized in Image 2",
        "target_bbox": [x1, y1, x2, y2],
        "analysis": "state the basis for establishing the correspondence, including the
actually observed cues among content association, text features, position, carrying
region, and surrounding-element relationships; end with the conclusion: judged as
matched"
      }
    },
    {
      "block_id": 2,
      "source_text": "the actual text of another to-be-translated block identified in
Image 1",
      "source_bbox": [x1, y1, x2, y2],
      "match": {
        "status": "missing",
        "recognized_text": "",
        "target_bbox": null,
        "analysis": "state the basis for judging missing: either (i) no text region in
Image 2 can be reliably corresponded to the source block (describe whether any candidate
 text exists and why it is insufficient in content, text features, position, carrying
region, or surrounding-element relationships), or (ii) the corresponding region still
shows the untranslated original Chinese content (state that the region remains the
original Chinese and was not converted into the target language); end with the
conclusion: judged as missing"
      }
    }
  ]
}
```

Style | input: source image, translated image and match result | output: font-style consistency of matched blocks.

Style judge — full prompt

Role and Task

You are a professional in-image translation quality assessment expert, skilled in text visual analysis. Given Image 1 (the original image), Image 2 (the translated image), and the match result between the source text blocks in Image 1 and the corresponding text regions in Image 2, you must assess text visual style.

Assessment Setup

Translation task: This task is whole-image Chinese→English translation. The unit of translation is the text block: any text block containing natural-language Chinese content is a text block to be translated, and it is translated and re-rendered as a whole; text blocks containing no Chinese at all (text in other languages, pure numbers, formulas, URLs, and brand logos / registered trademarks) are preserved content and must be kept unchanged.

Given match results: {match_blocks}
Each text block above includes `block_id`, `source_text`, and `source_bbox` (`[x1, y1, x2, y2]`, its position in Image 1); `match.status` is `matched` or `missing`, and for matched blocks `match.recognized_text` and `match.target_bbox` give the actual text and position of the corresponding region in Image 2. Treat this match result as given and correct; do not re-derive or override it.

Assessment procedure: block-level visual assessment (`style_consistency`).

Block-level Visual Assessment

For matched blocks, assess `style_consistency`.

`style_consistency`

Assessment target: Compare the text visual style of the source text block in Image 1 with the corresponding translated text block in Image 2, including overall font style, text fill color, presence and color

of outline/stroke, and clearly discernible decorative effects such as shadows. The visual size adapted to the target-language length is assessed by `layout_consistency`, not by this dimension.

Judged as true: The translated text's overall font style, fill color, outline/stroke, and decorative effects correspond to the source text.

Judged as false includes:

- The overall font style does not correspond to the source text;
- The text fill color has changed;
- An existing outline/stroke disappears, or text that originally had no outline/stroke gains one;
- The outline/stroke color is inconsistent with the source text;
- An existing shadow or other clearly discernible decorative effect is lost, or a decorative effect not present in the source text is added.

Output Format

Output valid JSON only, without Markdown code fences, without any content outside the JSON, and without adding or removing fields. `text_blocks` must keep the same `block_id` and order as the given match results above; both matched and missing blocks output the same field structure. `style_consistency.pass` uses boolean `true` or `false`; missing blocks use `null`. Each `analysis` first states the specific comparison evidence, then gives a clear conclusion in one sentence, and the conclusion must be consistent with the corresponding `pass` value.

Example output

```
{
  "text_blocks": [
    {
      "block_id": 1,
      "style_consistency": {
        "analysis": "compare the text visual size, overall font style, fill color,
outline/stroke presence and color, and shadows or other clearly discernible decorative
effects between the source text in Image 1 and the corresponding text in Image 2, state
the attributes that are actually kept consistent or changed, and end with a clear
conclusion consistent with pass",
        "pass": true
      }
    },
    {
      "block_id": 2,
      "style_consistency": {
        "analysis": "N/A: this block is not matched",
        "pass": null
      }
    }
  ]
}
```

Layout | input: source image, translated image and match result | output: layout consistency of matched blocks.

Layout judge — full prompt

Role and Task

You are a professional in-image translation quality assessment expert, skilled in layout analysis. Given Image 1 (the original image), Image 2 (the translated image), and the match result between the source text blocks in Image 1 and the corresponding text regions in Image 2, you must assess layout consistency.

Assessment Setup

Translation task: This task is whole-image Chinese→English translation. The unit of translation is the text block: any text block containing natural-language Chinese content is a text block to be translated, and it is translated and re-rendered as a whole; text blocks containing no Chinese at all (text in other languages, pure numbers, formulas, URLs, and brand logos / registered trademarks) are preserved content and must be kept unchanged.

Given match results: {match_blocks}
Each text block above includes block_id, source_text, and source_bbox ([x1, y1, x2, y2], its position in Image 1); match.status is matched or missing, and for matched blocks match.recognized_text and match.target_bbox give the actual text and position of the corresponding region in Image 2. Treat this match result as given and correct; do not re-derive or override it.

Assessment procedure: block-level visual assessment (layout_consistency).

Block-level Visual Assessment

For matched blocks, assess layout_consistency.

layout_consistency

Assessment target: Compare whether the position and layout of the translated text preserve the spatial organization of the corresponding text in Image 1, including the text's region, alignment, text orientation, its boundary relationship with text-carrying regions such as buttons, labels, color blocks, text boxes, signboards, and cards, and whether it overlaps or occludes neighboring text, graphics, or icons. The translated text may undergo necessary layout adjustments according to the expression length of the target language; this dimension assesses whether the adjusted layout still preserves the above spatial correspondence.

Judged as true: The translated text still lies in the region corresponding to the source text, preserves the corresponding alignment and text orientation, stays within the reasonable bounds of its text-carrying region, and has no abnormal overlap or occlusion with neighboring elements.

Judged as false includes:

- The translated text is placed in a position that does not correspond to the source text;
- The alignment has changed noticeably, e.g. originally centered text is clearly shifted to one side;
- The text orientation is inconsistent with the corresponding region of the source text;
- The text exceeds a clear text-carrying region such as a button, label, color block, text box, signboard, or card;
- The text overlaps or occludes neighboring text, graphics, or icons.

Output Format

Output valid JSON only, without Markdown code fences, without any content outside the JSON, and without adding or removing fields. text_blocks must keep the same block_id and order as the given match results above; both matched and missing blocks output the same field structure. layout_consistency.pass uses boolean true or false; missing blocks use null. Each analysis first states the specific comparison evidence, then gives a clear conclusion in one sentence, and the conclusion must be consistent with the corresponding pass value.

Example output

```
{
  "text_blocks": [
    {
      "block_id": 1,
      "layout_consistency": {
        "analysis": "compare the region, alignment, text orientation, boundary
relationship with the text-carrying region, and whether there is overlap or occlusion
with neighboring text/graphics/icons between the source text in Image 1 and the
corresponding text in Image 2, state the actually observed layout, and end with a clear
conclusion consistent with pass",
        "pass": false
      }
    },
    {
      "block_id": 2,
      "layout_consistency": {
        "analysis": "N/A: this block is not matched",
        "pass": null
      }
    }
  ]
}
```

Background | input: source and translated image | output: whole-image background preservation.

Background judge — full prompt

Role and Task

You are a professional in-image translation quality assessment expert, skilled in image-editing quality assessment. Given Image 1 (the original image) and Image 2 (the translated image), you must assess background content preservation.

Assessment Setup

Translation task: This task is whole-image Chinese→English translation. The unit of translation is the text block: any text block containing natural-language Chinese content is a text block to be translated, and it is translated and re-rendered as a whole; text blocks containing no Chinese at all (text in other languages, pure numbers, formulas, URLs, and brand logos / registered trademarks) are preserved content and must be kept unchanged.

Assessment procedure: whole-image background assessment (`background_preservation`).

Whole-image Background Assessment

`background_preservation`

Assessment target: Compare whether the image content in Image 1 and Image 2 other than the text content to be translated itself is preserved, including preserved content (text in other languages, pure numbers, formulas, URLs, brand logos / registered trademarks) and people, objects, products, graphics, icons, decorative elements, lines, color blocks, and other background structures. Assess the whole image once.

Judged as true: The above content is preserved in Image 2, with its content, position, shape, and structure corresponding to Image 1, and no extra modifications introduced by the text translation.

Judged as false includes:

- **Deletion or omission:** Text, objects, graphics, icons, decorative elements, or other background content that should be preserved in Image 1 is missing in Image 2;
- **Addition or replacement:** Image 2 contains extra content not present in Image 1, or existing background content is replaced by different text, objects, graphics, or other visual content;
- **Position, scale, or shape change:** A background element that should be preserved is noticeably moved, scaled, or reshaped, becoming inconsistent with its counterpart in Image 1;
- **Preserved-text distortion or garbling:** The glyphs of preserved background text — including text inside brand logos / registered trademarks, text in other languages, pure numbers, formulas, and URLs, and other non-translated content — are distorted, deformed, smeared together, garbled, or rendered illegible, even if the text's position has not moved;
- **Local editing artifacts:** After the text to be translated is removed and re-rendered, its surrounding area shows leftover original text, smearing, broken edges, wrong inpainting, repeated content, or other local anomalies inconsistent with Image 1.

Pixel differences caused by regeneration, compression, or scaling do not affect this dimension as long as they do not change the above background content, element position, shape, or structure.

Output Format

Output valid JSON only, without Markdown code fences, without any content outside the JSON, and without adding or removing fields. `background_preservation` is output once as a top-level field for the whole image. The `analysis` first states the specific comparison evidence, then gives a clear conclusion in one sentence, and the conclusion must be consistent with the `pass` value.

Example output

```
{
  "background_preservation": {
    "analysis": "compare the image content in Image 1 and Image 2 other than the to-be-
translated text blocks, state whether there is deletion/omission, addition/replacement,
position/scale/shape change, or local editing artifacts, point out the actually observed
 changes and their locations, and end with a clear conclusion consistent with pass",
    "pass": false
  }
}
```

Conditional input of style / layout | `match_blocks`. Style and layout take the following structured match result as their condition. It retains the block index, source text, source position and correspondence outcome, and requires the downstream judges to treat it as a given fact rather than re-deriving it.

Conditional input: `match_blocks`


```
[
  {
    "block_id": 1,
    "source_text": "上海环球港 Global Harbor",
    "source_bbox": [196, 73, 446, 296],
    "match": {
      "status": "matched",
      "recognized_text": "Shanghai Global Harbor",
      "target_bbox": [198, 80, 445, 292]
    }
  },
  {
    "block_id": 2,
    "source_text": "如何开启会员自动积分抵扣？",
    "source_bbox": [307, 323, 646, 351],
    "match": {
      "status": "missing",
      "recognized_text": "",
      "target_bbox": null
    }
  }
]
```


E Visual Judge Validation

E.1 Validation Design and Metrics

This appendix supplements only the sampling and human-annotation details not expanded in the main text. We draw a stratified sample of 50 TransVis-Bench images across the five text–visual relations, generate outputs with the 14 foundation image-generation models, and obtain 700 source–output image pairs; five annotators label independently and a majority vote is taken. For style and layout, unmatched blocks are labelled NA, so in addition to Cohen's κ we compute a scoring-eligibility agreement rate, which checks whether the automatic evaluator found the right objects to score. Table 2 in the main text already reports the complete candidate comparison and the deployment choice.

E.2 Human Agreement and Error Bounds

Over the 700 pairs, TransVis-Judger attains a mean κ of 0.470 and a mean scoring-eligibility agreement rate of 76.6%. These two quantities reflect, respectively, judgement consistency on valid scoring objects and agreement on the valid / NA boundary. Layout is the dimension with the lowest human agreement among all candidates—the highest κ is only 0.272—which is consistent with the intrinsic ambiguity introduced by cross-lingual length change, text-container boundaries and neighbour occlusion. We therefore retain per-dimension scores and cases in the main results rather than treating automatic judgement as an error-free human substitute; the complete candidate comparison is given in Table 2.

F Translator Instructions

F.1 One-Stage End-to-End Generation Prompt

All one-stage models use the following unified instruction.

One-stage end-to-end instruction (shared by all models)

Translate all Chinese text in the image into natural and accurate English, replacing each original text element in its corresponding position.

Requirements

1. Translate all Chinese text completely and accurately. Do not omit, duplicate, or add content.
2. Match the original typography and visual style as closely as possible, including font style, weight, size, color, outline, shadow, gradient, texture, and other effects.
3. Preserve the original layout, including text position, alignment, line breaks, spacing, rotation, perspective, and visual hierarchy. Minor font-size or line-break adjustments are allowed when needed for English text.
4. Keep existing English text, numbers, prices, dates, brand names, and other non-Chinese content unchanged.
5. Do not modify any non-text elements, including people, products, logos, icons, illustrations, background, textures, lighting, colors, or overall composition.
6. Remove the original Chinese text cleanly and reconstruct the underlying background naturally, without remnants, blur, seams, or editing artifacts.

Core principle: complete and accurate translation, faithful typography and layout, and minimal changes outside the text regions.

F.2 Two-Stage Block-Level Replacement Template

The two-stage condition instantiates the upstream OCR and translations into the following block-level replacement template; the ellipsis stands for the remaining to-be-replaced blocks of the same image.

Two-stage block-level replacement template

Replace the Chinese text “TESLA” with the English text “TESLA”;
Replace the Chinese text ‘‘上海就开特斯拉” with the English text “Drive a Tesla in Shanghai”;
Replace the Chinese text ‘‘全球超 900 万车主的智能安心之选” with the English text “The smart and safe choice of over 9 million owners worldwide”;
Replace the Chinese text “Model Y 蝉联全球乘用车销量冠军” with the English text “Model Y continues to be the world’s best-selling passenger car”;
. . . *(more block-level replacements)*

All new text should preserve the original font style, color and background of the source Chinese text; keep everything else in the image unchanged except the specified text regions.

G Implementation and Reproduction Settings

G.1 Training and Automatic Evaluation

We perform full-parameter supervised fine-tuning on two open-source image-editing backbones. Training uses 30k (source image, task instruction, target image) triples for one epoch; both runs are executed on a single node with $8\times$ NVIDIA H800 (80 GB) using DiffSynth-Studio, Accelerate, DeepSpeed, PyTorch 2.6 and CUDA 12.4 (Rajbhandari et al., 2020; Paszke et al., 2019). The shared configuration is bf16, gradient checkpointing, per-GPU batch size 1, gradient accumulation 1, an input cap of 2,359,296 pixels (1536×1536) and a learning rate of 1×10^{-6}, with a checkpoint saved every 1,000 steps. Boogu-Image-0.1-Edit uses ZeRO-2, a 200-step warmup and max gradient norm 1.0, and results are reported at 900, 1,800 and 2,854 steps; SenseNova-U1.5 uses ZeRO-3 without warmup or gradient clipping, and results are reported at 4,000, 8,000 and 11,416 steps. These correspond to roughly 33k, 66k and 100k cumulative samples, respectively.

The primary evaluation chain consists of TransVis-Judger, TextPecker-8B-Qwen3VL (Zhu et al., 2026), the translation judge and a fixed aggregation rule. The four Qwen3.6-27B specialised

judges of TRANSVIS-JUDGER are all fine-tuned to step 304 (Qwen Team, 2026a) and run in the conditional order match → style / layout → background; vLLM inference (Kwon et al., 2023) uses tensor parallelism 1, GPU memory utilisation 0.88, 7 replicas per dimension and 140-way concurrency. TextPecker performs OCR over the whole image with a dedicated GPU and 128-way concurrency; GPT-5.5 makes 0–3 translation judgements on matched blocks only, with 400-way concurrency. All block-level and image-level records are then aggregated into the final score and the eight layered metrics under the fixed rules of Section 3.1.

The complete prompts, output schemas and `match_blocks` conditional input of TRANSVIS-JUDGER are given in Appendix D. The evaluation protocol is fixed and does not vary with the evaluated model or scenario; for parameters such as temperature and top-p that are not publicly exposed by a model service, we do not force an override but adopt each model's default decoding behaviour.

G.2 EVALUATED MODELS AND THE TWO-STAGE CONDITION

All end-to-end models receive only the source image and the unified instruction of Appendix F.1; proprietary models preserve the original resolution where possible, and open-source models constrained by GPU memory use an input cap of 2,073,600 pixels. The unified instruction requires preservation of existing English, numbers, brand marks and all non-text elements, so that the task constraints are comparable across models.

The two-stage condition adds, under the same image constraints, block-level OCR and English translations produced by Gemini-3.1-Pro (Google, 2026a), and restricts the editable text blocks through the replacement template of Appendix F.2. It receives no additional human-corrected text, no visual annotation and no per-sample prompt tuning; the difference from the end-to-end condition is therefore attributable to visual-execution capability under changed availability of linguistic information, rather than to evaluation criteria, sampling parameters or manual optimisation.

H DATA SOURCES AND RELEASE PROCESSING

TRANSVIS-BENCH combines real collection, public benchmarks and controlled synthesis across five text–visual relations. Table 6 jointly reports category and sub-scenario counts, principal sources and pre-release processing, avoiding a separate provenance table with duplicated categories and totals. Samples from public sources follow their research-use terms; self-collected and screenshot samples have identifiable personal information, sensitive content and obvious watermarks removed during review. Synthetic overlays supply controlled pressures from post-hoc text, social-media visual language and subtitles rather than replacing naturally collected samples.

Table 6: Composition, data sources and pre-release processing of TRANSVIS-BENCH. The benchmark contains 2,591 images across 5 text–visual relations and 18 sub-scenarios. Categories are defined by how text and visual content constrain each other rather than by application name.

Text–visual relation	**Sub-scenarios (#images)**	**Total**	**Main sources**	**Pre-release processing**
Real-world captured text	Indoor displays (194); product packaging (148); public signage (158)	500	Self-collected scenes and interfaces	Removal of identifiable personal information and sensitive elements
Synthetic overlay text	Memes (124); social-media posts (155); synthetic typography (121); video subtitles (118)	518	Open Images base images (LLC, 2020); public memes / posts; video screenshots	Unified text overlay and visual-quality review
Dense document text	Books (102); comics (97); handwriting (100); reports (109); slides (97)	505	Chaystack (Li, 2026), OmniDocBench (Ouyang et al., 2025), M6Doc (Cheng et al., 2023)	Filtering of watermarks and illegible or potentially private samples
Graphic design composition	E-commerce posters (189); event posters (195); movie posters (162)	546	PKU-PosterLayout (Hsu et al., 2023), CGL-Dataset (Zhou et al., 2022), Movie-Poster	Human review of text legibility and copyright marks
Screen-rendered interface	Game HUDs (164); mobile apps (180); web pages (178)	522	Self-collected screenshots; CMGUI (Xie et al., 2026)	Removal of accounts, chat records and personal identifiers
Total	18 **sub-scenarios**	2,591	—	—

Release content and boundaries. Each image is released with metadata such as its text–visual relation, sub-scenario, image size and block-level annotation; the OCR and translations used for the two-stage condition are released as auxiliary inputs separate from the benchmark images. We do not include restricted raw material, identifiable personal information, or intermediate candidates that do not meet redistribution conditions in the release set. The "real-world" claim of TRANSVIS-BENCH therefore refers to the text–visual constraints and usage contexts that the samples carry, and does not imply that every released image is unprocessed raw web content.