Text-Only vs. Image-Aware VLM Judges for Manga Translation Evaluation

Subaru Kimura **Jun Suzuki**
Tohoku University
is-failab-research@grp.tohoku.ac.jp

Abstract

In manga translation, visual context outside speech-bubble text, including speaker identities, referents, and scene tone, is depicted in page images, making it natural to refer to images during translation evaluation. However, most conventional manga translation evaluations rely solely on text-only automatic metrics, leaving the impact of image input insufficiently explored. We propose a practical validation framework that employs the same vision-language model (VLM) as both a text-only judge (TOJ) and an image-aware judge (IAJ) to investigate evaluation divergences associated with image-aware evaluation procedures. We use a Japanese-English manga dataset of official translations aligned at the page and speech-bubble levels. First, adding page images to TOJ improved the primary VLM's detection of deliberately injected, visually dependent degradations. Second, we analyzed evaluation divergences between TOJ and IAJ, finding different error counts and error-category distributions. Human verification did not reveal a consistent preference between TOJ and IAJ.

1 Introduction

Machine translation research has advanced with the central goal of accurately converting text between languages at the sentence or document level. In practice, however, what should be translated and which expressions are appropriate depend on the communicative situation, medium, and non-verbal information available to readers, which may not be captured by source and target text alone.

Recent work has advanced both automatic translation evaluation (Lavie et al., 2025) and multimodal translation research incorporating visual context (Feng et al., 2025). Manga translation is a prime example: page images depict not only dialogue text, but also speakers, referents, and characters' emotional states.

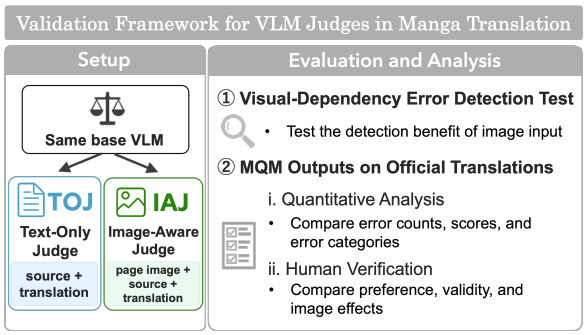

Figure 1: Overview of the proposed validation framework.

In manga translation quality evaluation (TQE), certain mistranslations, such as misidentifying speakers, pronouns, or object referents, or mismatching the tone of depicted scenes, can be difficult or impossible to identify reliably from source and target text alone and may require referencing page images. In addition, in literary and manga translation, trade-offs exist between literal fidelity, naturalness, character persona, and visual context adaptation, such that a single reference translation cannot always serve as the unique ground truth. Therefore, in manga TQE, including available page images in the input is natural to prevent information loss and evaluate appropriately.

Because VLMs can handle images and text simultaneously, an image-aware judge has the potential to detect visually dependent errors and incorporate context into evaluation that text alone cannot recover. However, providing images does not necessarily make a VLM judge universally more accurate. Errors in visual recognition or misinterpretations stemming from viewing images may also arise. The goal of this paper is not to assert that IAJ is uniformly superior to TOJ, but to investigate how VLM judge behavior changes under image-aware evaluation procedures.

Robust validation would ideally involve large-scale human correlation studies that account for di-

Proceedings of the Eleventh Conference on Machine Translation, pages 345–367
October 28-29, 2026 ©2026 Association for Computational Linguistics

versity across works, languages, and models. However, human evaluation of manga translation demands domain expertise, and suitable evaluation data are scarce, making such validation difficult to conduct at scale.

To address this difficulty, we propose a practical validation framework that compares the same underlying VLM under two conditions: **Text-Only Judge** (TOJ), which evaluates source and target text, and **Image-Aware Judge** (IAJ), which incorporates page images into the evaluation and, in our default setting, additional visual-grounding instructions. We use a Japanese-English manga dataset of official translations comprising 180 pages aligned at the speech-bubble level with page images. Using this framework, we investigate three research questions: whether image input improves the detection of degraded translations requiring visual context, how automatic evaluations of TOJ and IAJ differ, and whether human verification reveals a consistent preference between their fine-grained outputs.

The overview of the proposed validation framework is shown in Figure 1.

This paper makes four main contributions:

- **Validation framework for VLM judges:** We present a practical validation framework for manga translation evaluation using VLM judges that separately examines the effect of image input on detecting visually dependent errors and differences in evaluation characteristics between TOJ and IAJ on official translations.
- **Manga translation dataset based on official translations:** We transcribe Japanese source text and official English translations from manga pages publicly distributed by the rights holder, and construct an evaluation dataset of 180 pages aligned at the speech-bubble level with page images.
- **Benefit from image input:** In a degraded translation detection benchmark requiring visual context, simply adding page images to TOJ enables the primary VLM to detect more degraded translations.
- **Analysis of evaluation differences and human verification between TOJ and IAJ:** Comparing automatic evaluations of TOJ and IAJ on official translations, we show distinct trends in error counts, model-reported scores, and error-category distributions. Furthermore, human verification by 6 annotators on 40 sampled pages with TOJ/IAJ output differences reveals no consistent preference for either TOJ or IAJ. Post-task feedback suggests several directions for refining the human evaluation protocol, including prioritizing competing evaluation perspectives, defining boundary cases, and providing concrete examples of image effects.

2 Related Work

2.1 Machine Translation Evaluation and Large Language Model Judges

In automatic machine translation (MT) evaluation, following lexical overlap metrics with reference translations like BLEU (Papineni et al., 2002) and chrF (Popović, 2015), learned metrics such as COMET (Rei et al., 2020) and MetricX (Juraska et al., 2023) have been developed. Furthermore, large language model (LLM)-based metrics like G-Eval (Liu et al., 2023) and GEMBA (Kocmi and Federmann, 2023b) have enabled reference-free evaluation.

In human evaluation frameworks, Direct Assessment (DA) (Graham et al., 2013), Error Span Annotation (ESA) (Kocmi et al., 2024), and Multidimensional Quality Metrics (MQM) (Lommel et al., 2014; Freitag et al., 2021) have been widely adopted. Recently, LLM-as-a-Judge using MQM-based evaluation has gained traction (Fernandes et al., 2023; Kocmi and Federmann, 2023a; Agrawal et al., 2024), and has shown efficacy in document-level, context-aware translation evaluation (Yan et al., 2025; Semenov et al., 2025).

Building on the established efficacy of LLM judges in MQM evaluation, this study investigates the efficacy of VLM judges in MQM evaluation. We focus on differences between text-only and image-aware judges using the same underlying VLM. This enables realistic, albeit bounded, validation of image input effects in automatic manga translation evaluation, which was previously difficult to verify.

2.2 Multimodal Translation and Manga Translation

As multimodal machine translation research progresses, translation methods leveraging visual context in manga translation have been proposed (Hinami et al., 2021; Lippmann et al., 2025). In partic-

ular, Hinami et al. framed manga translation as a context-dependent multimodal task, showing that human evaluations viewing manga images rank translations differently than BLEU scores. This provides an early example highlighting the limitations of evaluating manga translation using text-only metrics alone. However, existing manga MT research focuses primarily on translation model development, leaving manga-specific TQE using page images, source text, and candidate translations underexplored.

Meanwhile, in the broader multimodal translation domain, studies leveraging VLM judges to analyze visual ambiguity in translation datasets have begun to emerge (Betala et al., 2025). However, because it remains unclear whether models supplied with images actually use visual information, or what image input improves and changes, VLM judges themselves require rigorous validation.

Additionally, MangaVQA (Baek et al., 2026) demonstrates that recent VLMs possess foundational visual understanding capabilities for manga pages. However, comprehending manga is necessary, but it is not equivalent to accurately evaluating translation quality. Furthermore, even a single misrecognition can alter translation evaluation decisions.

Accordingly, we use the proposed validation framework to examine the impact of image input on VLM judges' detection of visually dependent errors and output characteristics in manga translation evaluation.

3 Experimental Setup

3.1 Automatic Evaluation Design

The evaluation unit is page-level. A single page is represented as a triplet $x = (I, s_{1:n}, t_{1:n})$, consisting of Japanese page image I, Japanese source segment sequence $s_{1:n}$, and English translation segment sequence $t_{1:n}$. Here, n denotes the number of speech bubbles (segments) on a page, varying by page. TOJ takes $(s_{1:n}, t_{1:n})$ as input and uses a default MQM-style prompt. Our default IAJ (IAJ-full) takes $(I, s_{1:n}, t_{1:n})$ as input and adds a visual-grounding instruction and a `visual_caption` output requirement to the default prompt. Note that the input image I provided to IAJ is **only the Japanese page image**; the English target page image is not provided to the judge. Both conditions employ MQM-style evaluation with our custom 10-category taxonomy, outputting an auxiliary model-reported score $S_p \in [0, 100]$ per page and a structured list of errors for each mistranslation location. Here, S_p is the integer score output in the JSON field `quality_score`. We define the corresponding model-reported score deficit as $D_p^{\text{judge}} = 100 - S_p$. Separately, we compute fixed-weight penalties D_p^{MQM} from the three reported severity labels—`neutral`, `minor`, and `major`. Each error item consists of the corresponding segment index, error category, severity, and a natural language explanation.

Our MQM-style evaluation prompt adapts MQM-Chat (Li et al., 2025) (which defines 7 error types for dialogue translation, including ambiguity, slang/loanwords, and dialogue consistency), redefining and expanding categories to match manga translation characteristics (segment-level bubbles, personal reference ambiguity, tone/emotion, and visual scene alignment) into our custom 10-category taxonomy (full prompt in Appendix A).

3.2 Dataset

As the evaluation target, we use Volume 1 of *Give My Regards to Black Jack*, by Shuho Sato. In this study, we transcribed Japanese source text and English translation text from publicly available manga images, establishing 180 pages with unique 1:1 segment alignment at the bubble level verified via optical character recognition (OCR) and human inspection (dataset statistics in Appendix B).

The evaluation data are publicly available at https://huggingface.co/datasets/joddy/gmr2bj_manga_mt_ja_en for reproducibility. Details of released dataset contents are provided in Appendix C.

3.3 Models

The primary evaluation model is `gemini-3-flash-preview` (Google, 2025). This selection is based on pilot study findings showing that this model balances manga page image comprehension and structured output stability effectively. Furthermore, MangaVQA reports strong performance of Gemini-family models on manga page understanding (Baek et al., 2026), which motivated our use of a Gemini-family VLM in this study. We employ the same model under both TOJ and IAJ conditions to analyze differences between text-only and image-aware evaluation.

In this paper, we primarily report results for the main evaluation model. To verify that gains

Degradation Type	Description
speaker_or_addressee	Confuses speaker or addressee, shifting subjects or honorifics
referent_or_object	Translates referenced objects/entities in image as something else
tone_emotion_or_style	Translates with tone or style contradicting facial expression or scene
action_or_state	Translates with actions or physical states contradicting depicted artwork

Table 1: Degradation operation types used in Experiment 1 to create translations contradicting manga page image content.

from image-aware procedures are not unique to the primary model, we also perform automatic MQM evaluation with `gpt-5.4-mini` (OpenAI, 2026a) and `qwen3-vl-32b-instruct` (Qwen Team, 2025) as a robustness check. Experiment 1 evaluates 75 degraded translations, and Experiment 2 evaluates 180 official translation pages.

4 Experiment 1: Degraded Translation Detection

4.1 Creation of Degraded Translations

To intentionally generate "translation errors that cannot be decisively identified without viewing page images," we designed four types of degradation operations shown in Table 1. These four degradation types are benchmark-construction categories and are distinct from the ten MQM error categories used by the judges. Using `gemini-2.5-pro` (Gemini Team, 2025) and `gemini-3-flash-preview`, we generated degraded translation candidates contradicting image content across these four types, selecting a total of 75 benchmark instances through human vetting. Across the entire process, 75 final instances were adopted from over 900 generated candidates, yielding an overall adoption rate of approximately 8%. We note that the number and variety of highly visually dependent degradations that can be constructed from a single 180-page volume are inherently limited. Prompts used for candidate creation are detailed in Appendix A.

4.2 Metrics

In Experiment 1, we post-hoc evaluate whether error explanations output by TOJ/IAJ as MQM-style judges using our custom 10-category taxonomy semantically capture the pre-designed degradation intent.

For each degraded translation, `gpt-5.6-luna` (OpenAI, 2026b) classifies the relationship between the judge's error explanation and the degradation intent into {`same`, `related`, `different`}, counting `same` as a semantic match. We term this proportion the **semantic match rate** (prompts in Appendix A). This is an instance-level detection rate: the proportion of the 75 degraded translations for which at least one reported error is classified as `same`. It is not conventional overall accuracy, error-level recall, or MQM accuracy. This metric does not assess the precision of the reported errors. Reported errors that do not match the injected degradation may still identify genuine translation issues and therefore cannot automatically be treated as false positives. Without independent annotations of these additional errors, we cannot reliably compute precision or an F1 score. We also report the mean number of error flags per degraded translation, computed from the 75 fixed MQM outputs independently of the five semantic match runs. During exploratory stages, we experimented with rule-based metrics, but selected semantic match as an LLM-based evaluation method capable of accounting for paraphrastic variation. We formulate semantic matching as a natural language inference (NLI)-like classification task that can accommodate paraphrased explanations. Its validity remains limited by the lack of independently annotated gold labels.

4.3 Compared Evaluation Conditions

The prompt used for our default IAJ (IAJ-full) adds a visual-grounding instruction and a `visual_caption` requirement to TOJ's base prompt. The difference between TOJ and IAJ-full encompasses not only image presence, but also the combined effect of input modality and visual grounding procedures. To disentangle this confounding factor, Experiment 1 introduces **IAJ-minimal** (adding only page images to TOJ) and **IAJ-blank** (replacing IAJ-full images with a blank image).

4.4 Results

Because semantic match judgments can vary across runs, we repeated semantic match classification five times using the same fixed MQM outputs, without re-running the underlying VLM judges. We report the mean and range in Table 2.

Condition	Page Image	Grounding	Mean (%)	5-run Range (%)	Mean flags
TOJ	None	None	57.9%	56.0–58.7%	1.11
IAJ-blank	Blank image	Yes	52.0%	50.7–53.3%	1.03
IAJ-minimal	Actual page	None	81.6%	80.0–82.7%	1.40
IAJ-full	Actual page	Yes	80.5%	78.7–82.7%	1.20

Table 2: Experiment 1 results. Semantic match rates are summarized by their mean and range across five runs. Mean flags denotes the average number of reported errors per instance.

IAJ-minimal achieved a mean rate of 81.6%, and IAJ-full achieved 80.5%; both had higher semantic match rates than TOJ's 57.9%. Conversely, IAJ-blank achieved a mean rate of 52.0%, dropping below both IAJ-full and TOJ. The pure contrast of image presence (TOJ vs. IAJ-minimal) yielded a 5-run mean difference of +23.7 percentage points. Comparing identical degraded translations between TOJ and IAJ-minimal and testing aggregated differences grouped by source page revealed that IAJ-minimal's gain was statistically significant across all 5 runs. Significance test details are shown in Appendix D. This ablation result shows, on this constructed benchmark, that providing page images increases the semantic match rate compared to TOJ, whereas substituting actual images with blank images degrades the semantic match rate. This provides evidence that page images help detect degraded translations requiring visual context on this constructed benchmark.

The slightly higher semantic match rate of IAJ-minimal over IAJ-full does not necessarily imply that IAJ-minimal is a more accurate judge. Semantic match measures whether an error explanation matching the degradation intent is contained in the MQM output, not overall MQM precision. IAJ-minimal produced more error flags overall than IAJ-full, which may have driven up the semantic match rate. Furthermore, we note that the additional prompt in IAJ-full was added to the base prompt because preliminary qualitative checks showed it aided MQM precision prior to scaling experiments.

We illustrate one representative degraded translation contradicting manga image content that was detected by IAJ-full but missed by TOJ. This example concerns a visually dependent degradation involving a referent. Additional examples of visually dependent degradations are provided in Figure 3 in Appendix E.

4.5 Robustness Across Other VLMs

To verify whether divergences manifest across models, we evaluated TOJ and IAJ-full conditions using `gpt-5.4-mini` and `qwen3-vl-32b-instruct` (Table 3).

In both models, IAJ-full achieved a higher semantic match rate than TOJ, though the rates varied considerably by model. These results show that the directional divergence between TOJ and IAJ-full is not unique to the primary model, while indicating that absolute semantic match rates and gain magnitudes depend on the underlying VLM.

5 Experiment 2: TOJ/IAJ Evaluation Characteristics and Human Verification

In Experiment 2, we deliberately evaluate original official translations using automatic MQM-style evaluation to quantitatively and qualitatively analyze output differences between TOJ and IAJ-full. Hereafter, references to IAJ imply IAJ-full unless noted otherwise. When treating official translations as a provisional reference baseline, errors reported by judges can be interpreted as provisional false positives relative to the official translation or as overflagging relative to that baseline. However, official translations are not absolute ground truth; they may contain actual translation errors or room for improvement, and alternative valid translations exist. Furthermore, discrepancies between choices made in official translations and evaluation aspects prioritized by TOJ/IAJ can cause judges to flag errors in official translations. Thus, evaluating official translations allows us to analyze the evaluation characteristics of TOJ and IAJ.

5.1 Quantitative Analysis Design

By analyzing differences in MQM-style outputs between TOJ and IAJ-full, we quantitatively elucidate differences in their evaluation characteristics. Errors in MQM-style outputs are categorized into TOJ-only, IAJ-only, and common, analyzing output divergences. We match TOJ and IAJ flags

MQM Judge	Semantic match rate			Mean flags	
	TOJ	IAJ-full	Difference	TOJ	IAJ-full
`gemini-3-flash-preview`	57.9% (56.0–58.7)	80.5% (78.7–82.7)	+22.7 pp	1.11	1.20
`gpt-5.4-mini`	24.8% (24.0–25.3)	30.7% (30.7–30.7)	+5.9 pp	1.05	0.89
`qwen3-vl-32b-instruct`	45.6% (44.0–48.0)	59.2% (57.3–61.3)	+13.6 pp	1.43	1.53

Table 3: Experiment 1 cross-model comparison. Semantic match rates are five-run means with ranges in parentheses. Differences are IAJ-full minus TOJ, computed before rounding. Mean flags is defined as in Table 2.

bj_vol1_113 (JA)

bj_vol1_113 (EN)

Degraded segment
type: `referent_or_object`
JP "ヒゲが……"
EN (original) "HIS BEARD–"
EN (degraded) "MY BEARD–"

Degradation intent
"The possessive 'HIS' was changed to 'MY', shifting the referent of the beard from the patient to the speaker (the doctor)."

IAJ-full error explanation
"The doctor is observing the patient's beard, not his own. The use of the first-person possessive 'MY' is a reference error that misidentifies whose beard is being discussed."

Figure 2: Representative example (`bj_vol1_113`) of degraded translation contradicting manga image content missed by TOJ but detected by IAJ-full. Left: JA/EN page images; Right: degraded segment, degradation intent, and IAJ-full error explanation. The EN page image is shown for illustration only and was not provided to IAJ. The patient's beard and the doctor's face clarify the referent error changing "HIS" to "MY". *Give My Regards to Black Jack / SHUHO SATO*

at the segment level using a rule-based procedure, classifying corresponding flags as common and non-corresponding flags as TOJ-only or IAJ-only. This heuristic matching combines explanation token set Jaccard similarity and string similarity with `error_category` and `severity` alignment. The objective is to mechanically link obviously identical flags to aggregate output differences quantitatively. Therefore, this classification does not guarantee strict semantic identity, and results should be interpreted as diagnostic statistics describing TOJ and IAJ output divergences. Matching methodology details are provided in Section G of the Appendix.

5.2 Qualitative Analysis Design

To qualitatively analyze TOJ and IAJ-full evaluations, we conduct human verification involving 6 Japanese annotators from the same laboratory (non-authors). Annotators were native Japanese speakers familiar with reading manga and possessing backgrounds in language processing research. All annotators were master's or doctoral students at Tohoku University and had no prior experience in professional translation or MQM annotation. Standardized English proficiency test scores were not collected. While using professional Japanese-to-English manga translators would be ideal, recruiting and executing verification with multiple professional translators presents substantial practical hurdles.

For human verification targets, from the candidate set of 116 pages among 180 pages having at least one unique flag in TOJ or IAJ, we selected 40 pages via simple random sampling. Sampling from pages with output differences allows for a clearer analysis of evaluation characteristic divergences between TOJ and IAJ. The selected 40 pages comprised 12 pages with unique flags in TOJ only, 4 pages in IAJ only, and 24 pages with unique flags in both. For these 40 pages, evaluators compared TOJ and IAJ flags for the same page.

The main judgment asked evaluators which judge's flags were superior, with three response options: `prefer_IAJ` / `prefer_TOJ` / `tie`. Additionally, evaluators reported confidence, page-level output validity ratings for TOJ/IAJ, image effect on IAJ flags, and usefulness of image information in translation evaluation. Each validity

rating covered the complete flag set produced by one judge for one page; thus, the denominator was 40 pages $\times$ 6 annotators $=$ 240 ratings per judge. Evaluators received descriptions for each option but were not instructed on a rigid priority hierarchy when source fidelity, naturalness, and visual consistency conflicted. When 4 or more of 6 evaluators selected the same main judgment, it was classified as strict majority; otherwise, it was classified as `no_majority`. For operational protocol details, annotation user interface (UI), and instructions, see Appendix F.

5.3 Quantitative Analysis of MQM-style Evaluation

5.3.1 Overview of Automatic Evaluation across 180 Pages

Table 4 shows automatic evaluation results across all 180 official translation pages for TOJ and IAJ-full, including error-flag counts and model-reported score statistics. For gemini-3-flash-preview, IAJ-full yielded fewer mean error flags per page than TOJ ($1.77 \rightarrow 1.46$), and a lower mean model-reported score deficit ($100 - S_p$) ($10.24 \rightarrow 8.08$). Each reported error is assigned one of three severities: `major`, `minor`, or `neutral`. Prompts define `major` as errors substantially distorting meaning, `minor` as errors with limited impact, and `neutral` as reported issues with constrained effect. In this paper, all three severities are included in error flag counts. Major flags were fewer than minor flags, with major-free pages accounting for 147/180 (81.7%) in TOJ and 158/180 (87.8%) in IAJ-full. Conversely, minor errors occurred at 1–2 items per page across most pages for both judges. While this does not directly show that IAJ-full produced more accurate evaluations, it shows that image-aware procedures alter judge output volume and reported scores.

As a fixed-weight scoring analysis based on MQM severity labels, we assigned weights of 0, 1, and 5 to neutral, minor, and major, respectively, and calculated penalty points per 100 words of the official English translation. For the primary evaluation model, TOJ yielded 5.80 penalty points, compared with 4.57 for IAJ-full; thus, the tendency for IAJ-full to incur fewer penalties was reproduced in the fixed-weight MQM aggregation. This result is consistent with the comparison based on the judge-reported `quality_score`; the page-level Spearman correlations between the two measures were 0.762 for TOJ and 0.717 for IAJ-full. The auxiliary models showed differences in the same direction, while fixed-weight scoring and sensitivity analyses are provided in Appendix G.

Thus, automatic comparison across 180 pages showed that IAJ-full produced fewer flags and smaller auxiliary score deficits than TOJ. Auxiliary models exhibited a similar directional trend: IAJ-full produced fewer total error flags and smaller auxiliary mean score deficits than TOJ. Auxiliary model results for Experiment 2 are detailed in Appendix H.

Metric	**TOJ**	**IAJ-full**
All 180 Pages		
Total error flags	318	262
Mean / page	1.77	1.46
Major error flags	36	24
Mean major / page	0.20	0.13
Minor error flags	268	233
Mean minor / page	1.49	1.29
Neutral error flags	14	5
Mean neutral / page	0.08	0.03
Mean quality_score	89.8	91.9
Mean score deficit ($100 - S_p$)	10.24	8.08
Page-Level Error Distribution		
0-error pages	20 (11.1%)	31 (17.2%)
1+ error pages	160 (88.9%)	149 (82.8%)
0-major pages	147 (81.7%)	158 (87.8%)
0-error in both judges	19 (10.6%)	

Table 4: Automatic evaluation results reported by judges on official translations across 180 pages (based on `gemini-3-flash-preview`). The table includes total error flags, severity counts, mean model-reported `quality_score`, and mean score deficit ($100 - S_p$). Total error flags include `major`, `minor`, and `neutral`.

5.3.2 Structural Analysis of Judge Output Divergence

To examine TOJ and IAJ-full output differences in detail, we categorized errors across all 180 pages into TOJ-only, IAJ-only, and common. Overall, TOJ-only = 152, IAJ-only = 96, and common = 166, showing TOJ produced more unique flags. We define TOJ-only share as TOJ-only/(TOJ-only + IAJ-only). In `tone_and_emotion` (65.5%), `omission_or_addition` (64.6%), `terminology_and_named_entities` (62.5%), and `naturalness_and_register` (61.2%), TOJ-only flags were more frequent than IAJ-only flags, suggesting that TOJ often flags fine-grained differences in naturalness, style, and fidelity. In contrast, `reference_and_disambiguation` and

Category	TOJ-only	IAJ-only	TOJ-only Share
Overall	152	96	61.3%
meaning_accuracy	42	32	56.8%
naturalness_and_register	30	19	61.2%
omission_or_addition	31	17	64.6%
tone_and_emotion	19	10	65.5%
terminology_and_named_entities	20	12	62.5%
reference_and_disambiguation	4	4	50.0%
action_and_state	1	1	50.0%
speaker_and_dialogue_consistency	0	0	–
colloquial_and_cultural_expression	3	0	100.0%
other	2	1	66.7%

Table 5: Breakdown of TOJ/IAJ unique error flags by MQM category across 180 pages.

`action_and_state` showed nearly equal TOJ-only and IAJ-only shares, indicating IAJ-full also produced unique flags regarding visual reference resolution and action understanding. Category breakdowns for TOJ-only / IAJ-only are presented in Table 5.

Furthermore, comparing 54 segment locations where both judges flagged the same page segment but could not be matched as common, 37 instances (68.5%) differed in `error_category`. This indicates that even when targeting the same segment, TOJ and IAJ frequently conceptualize issues from different perspectives.

5.3.3 Retention of TOJ Flags in IAJ-full

We define coverage as the proportion of error candidates flagged by TOJ that are retained as identical flags in IAJ-full. Across 160 pages with 1+ TOJ flag (totaling 318 TOJ flags), 166 flags (52.2%) were retained in IAJ-full. On 45 pages (28.1%), IAJ-full retained zero TOJ flags. This indicates that while IAJ-full incorporates flags overlapping with TOJ, it also does not retain some flags reported by TOJ. Detailed coverage distribution statistics are provided in Table 13 in Appendix I.

5.4 Qualitative Analysis via Human Verification

5.4.1 Main Judgment and Agreement among 6 Annotators

Table 6 summarizes the 240 judgments from 6 annotators over 40 pages. Preferences were dispersed, and agreement was limited despite predominantly medium or high confidence. Among the 14 pages without a majority, 12 had a 3–2–1 split and two a 3–3 split; most disagreements were therefore not clean binary divisions.

Annotators also showed distinct tendencies: one chose `tie` on 27 pages, another preferred TOJ on 24, and another preferred IAJ on 17. Low agreement thus cannot be explained merely by hesitation. Possible factors include small differences between outputs, differing criteria or tie thresholds, and characteristics of the sampled pages, but these results do not isolate their contributions.

Metric	Value
Annotators / Verified Pages	6 / 40
Individual votes ($n = 240$)	
Prefer IAJ / TOJ / Tie	73 / 87 / 80
Percentage	30.4 / 36.3 / 33.3%
Medium or high confidence	228 (95.0%)
Mean Pairwise Agreement	0.512
Fleiss' κ	0.266
Strict Majority - unanimous	5 / 40
Strict Majority - non-unanimous	21 / 40
Any majority	26 / 40 (65.0%)
`prefer_IAJ` Consensus	7
`prefer_TOJ` Consensus	12
`tie` Consensus	7
`no_majority`	14 / 40 (35.0%)

Table 6: Experiment 2: Main judgment agreement and page-level consensus among 6 annotators over 40 TOJ/IAJ-full MQM outputs.

5.4.2 Page-Level Output Validity and Reasons for Tie

Table 7 summarizes ratings of each judge's complete page-level flag set, not individual errors. Both judges were often rated valid, with similar valid rates, whereas IAJ received more unclear ratings. The dominant tie reason was that both sets were valid, rather than both invalid. Rating definitions are provided in Appendix F.

5.4.3 Usefulness of Image Information and Effect on IAJ

Perceived image usefulness and observed effects on IAJ flags address different questions (Table 7).

Output validity ($n = 240$ **each**)	**TOJ Count (%)**	**IAJ Count (%)**
Valid	168 (70.0)	172 (71.7)
Partially valid	46 (19.2)	28 (11.7)
Invalid	15 (6.3)	12 (5.0)
Unclear	11 (4.6)	28 (11.7)
Valid or partially valid	214 (89.2)	200 (83.3)
Tie reason ($n = 80$)		**Count (%)**
Both valid		58 (72.5)
Both partially valid		10 (12.5)
Difference too small		10 (12.5)
Both invalid		1 (1.3)
Other		1 (1.3)

Image usefulness ($n = 240$)	**Count (%)**
Helpful or essential	124 (51.7)
Not needed	109 (45.4)
Unclear	7 (2.9)
Image effect ($n = 240$)	**Count (%)**
No observable effect	137 (57.1)
Any positive effect	55 (22.9)
Any negative effect	26 (10.8)
Unclear	24 (10.0)

Table 7: Experiment 2 human ratings. Left: output validity and reasons for ties. Right: perceived image usefulness and observed effects on IAJ outputs. Values are counts with percentages in parentheses. IAJ denotes IAJ-full. Positive effects include valid detection, overflagging prevention, or rationale strengthening; negative effects include visual misrecognition/overinterpretation or missed textual issues. Effects are counted per response, with two responses reporting both positive and negative effects.

Judgments of potential usefulness were nearly evenly divided, whereas most responses reported no observable effect on IAJ flags. Positive and negative effects were both reported, indicating instance-dependent benefits and occasional adverse effects rather than consistent improvement. Image effects allowed multiple selections; full option definitions appear in Appendix F.

5.4.4 Post-Task Feedback

As post-task feedback upon completing annotations, evaluators answered 5-point scale items regarding overall task difficulty, rubric sufficiency, priority clarity when criteria conflict (source fidelity vs. target naturalness vs. visual consistency), main label clarity, and usefulness of concrete examples. Additionally, evaluators picked the single annotation field that caused the most hesitation across 40 pages, and selected multiple-choice options for explanations or examples that would aid judgment in future iterations (e.g., priority rules, boundary case examples, image effect examples, severity rules, concrete label examples). Across 6 annotators' post-task feedback, median task difficulty was 4.0/5, while rubric sufficiency and main label clarity medians were both 4.0. In contrast, priority clarity when criteria conflict received a median of only 2.5. Usefulness of concrete label examples or boundary cases had a median of 4.0, and 4 of 6 annotators picked image effect classification as the hardest field. Additional guidance covering priority rules, boundary examples, image effect examples, severity examples, and label examples was requested by 3 annotators each. Thus, future evaluation designs should consider providing explicit priority rules and concrete benchmark examples when evaluation perspectives conflict. However, causal improvements in agreement or accuracy from adding guidelines remain unverified.

6 Discussion

We briefly summarize key insights derived from experimental results.

Does Adding Page Images Aid Visually Dependent Error Detection? In Experiment 1, for 75 degraded translations designed to be difficult to judge without images, IAJ-minimal (adding only page images to TOJ) achieved a higher mean semantic match rate than TOJ, with the difference being statistically significant across all 5 runs. These results provide evidence that, on this constructed benchmark, inputting page images aids the detection of visually dependent degradations.

How Does Image-Aware Procedure Alter Judge Evaluation Characteristics? In Experiment 2, IAJ-full produced fewer flags and smaller auxiliary score deficits derived from model-reported scores than TOJ, exhibiting divergences in unique flag counts and error category assignments for identical segments. This indicates that IAJ-full does not merely add flags: some flags reported by TOJ are not retained, and errors are reported from distinct angles.

Can Human Verification Stably Arbitrate TOJ vs. IAJ Differences? Human verification

showed no consistent preference trend favoring either TOJ or IAJ. Page-level output validity ratings for both conditions did not differ markedly, and `both_valid` was the dominant reason for `tie` judgments. This pattern is consistent with the difficulty of ranking two flag sets both judged valid overall, although the present results do not isolate the cause of disagreement. Evaluator perceptions of potential image usefulness split evenly, while observed image effects on actual IAJ flags showed no effect in 57.1%, positive effects in 22.9%, and negative effects in 10.8%. Thus, human verification found no uniform preference for IAJ-full, indicating that the image-aware evaluation procedure does not yield uniformly preferred evaluations across all pages. Post-task feedback showed that annotators sometimes had difficulty choosing between judge outputs when different evaluation criteria favored different outputs. The limited agreement may partly reflect differences in how annotators prioritize source fidelity, target-language naturalness, and visual consistency when comparing judge outputs. Our protocol did not prescribe a strict priority among these criteria, leaving room for differing judgments. However, the present results cannot determine whether disagreement primarily reflects small differences between judge outputs, differing priorities among evaluation criteria, or characteristics of the manga title and sampled pages.

What Is Required for Practical VLM Judge Evaluation? Future work should calibrate evaluators after sharing priority rules for conflicting evaluation perspectives, common criteria, boundary cases, and concrete examples of image effects. However, whether adding such explanations or calibration improves agreement or judgment stability remains unverified. Practical use also requires additional validation across multiple works, other language pairs, and models.

7 Conclusion

This study presents a validation framework for manga translation VLM judges that separately evaluates detection gains for visually dependent errors and characteristic output changes induced by image-aware procedures. To accomplish this, we built an open evaluation dataset of 180 pages from transcribed official Japanese and English manga, aligned at the speech-bubble level with page images. In Experiment 1, we found that adding page images aids in detecting visually dependent errors on this constructed benchmark, providing evidence for a baseline benefit of utilizing VLM judges. In Experiment 2, while IAJ-full exhibited different error volume and evaluation perspectives from TOJ, human preference judgments across 6 annotators showed no consistent superiority. These findings suggest that IAJ merits further investigation for manga translation evaluation, but do not support replacing TOJ with IAJ universally. Through our proposed validation framework, we present a practical validation approach for settings where large-scale human correlation studies are difficult. Our framework provides a starting point for extending VLM judge evaluation to other visually dependent translation domains, such as subtitles, advertisements, and instruction manuals.

Limitations

- The evaluation target is restricted to a single manga work and language pair (ja→en). Generalizability to other works, genres, or language pairs remains unverified; we do not claim identical effect sizes or output characteristics hold across manga broadly. Copyright constraints surrounding redistributable official parallel manga data exacerbate this limitation.
- Degraded translations in Experiment 1 were intentionally created and selected to require visual information, and do not reflect the natural distribution of manga translation errors. Therefore, detection rates in Experiment 1 should not be interpreted as operational error detection rates or representative values for manga translation errors as a whole.
- Experiment 2 compares TOJ with IAJ-full; therefore, observed differences cannot be attributed solely to image presence because IAJ-full also modifies the evaluation instructions.
- The 5 runs in Experiment 1 repeat semantic-match classification of fixed MQM outputs, rather than MQM judge execution under identical conditions. MQM judge execution used `temperature=0.1`, but lowering the temperature does not guarantee identical outputs under repeated execution. Therefore, observed differences among TOJ, IAJ-minimal, IAJ-full, and IAJ-blank may partly reflect single-run MQM-output variability, in addition to differences in image input and evaluation instructions.
- Semantic match evaluation relies on LLM judge precision. While authors performed spot-checks, alignment against independent multi-annotator gold labels and systematic verification of meta-evaluator precision were not conducted. Consequently, reported semantic match rates may contain meta-evaluator judgment noise.
- We did not validate the judges' error localization, categories, severity labels, or explanations against independent MQM annotations by experienced translation experts. Human verification in Experiment 2 assessed judge outputs rather than independently annotating translation errors. Thus, our findings concern detection of injected degradations and differences in judge outputs, and do not establish overall MQM accuracy.
- Common flag matching between TOJ and IAJ-full employs a rule-based heuristic based on explanation similarity, category, and severity, without human alignment validation. Therefore, unique flag counts and coverage rates may reflect heuristic matching errors.

Ethical Considerations

Content Safety. The evaluated work is a general-audience medical drama containing medical depictions (surgeries, death), but review of the 180 evaluated pages revealed no explicit hate speech, discrimination, or explicit violence.

Acknowledgments

We thank Shuho Sato and Sato Manga Works Ltd. for making *Give My Regards to Black Jack* available under a generous reuse policy, which enabled this study and the release of the accompanying dataset. This work was partly supported by JST Moonshot R&D Grant Number JPMJMS2011-35 (fundamental research), by JSPS KAKENHI Grant Number JP26H02511, and by JST BOOST Japan Grant Number JPMJBS2421.

References

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. 2024. Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267.

Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. 2026. MangaVQA and MangaLMM: A benchmark and specialized model for multimodal manga understanding. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5349–5370, Rabat, Morocco. Association for Computational Linguistics.

Siddharth Betala, Kushan Raj, Vipul Betala, and Rohan Saswade. 2025. A picture is worth a thousand (correct) captions: A vision-guided judge-corrector system for multimodal machine translation. In *Proceedings of the 12th Workshop on Asian Translation*, pages 124–137, Mumbai, India. Association for Computational Linguistics.

Yi Feng, Chuanyi Li, Jiatong He, Zhenyu Hou, and Vincent Ng. 2025. Multimodal neural machine translation: A survey of the state of the art. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22130–22147, Suzhou, China. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Google. 2025. Gemini 3 flash preview. `https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview`. Accessed August 5, 2026.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12998–13008.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2025. MQM-Chat: Multidimensional Quality Metrics for Chat Translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3283–3299, Abu Dhabi, UAE. Association for Computational Linguistics.

Philip Lippmann, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. 2025. Context-informed machine translation of manga using multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3444–3464, Abu Dhabi, UAE. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 12:455–463.

OpenAI. 2026a. GPT-5.4 mini model | OpenAI API. `https://developers.openai.com/api/docs/models/gpt-5.4-mini`. Accessed August 5, 2026.

OpenAI. 2026b. GPT-5.6 system card. OpenAI Deployment Safety Hub. July 9, 2026. Accessed August 5, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3-VL technical report. *CoRR*, abs/2511.21631.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.

Brian Yan, Shuoyang Ding, Kuang-Da Wang, Siqi Ouyang, Oleksii Hrinchuk, Vitaly Lavrukhin, and Boris Ginsburg. 2025. NVIDIA-NeMo's WMT 2025 metrics shared task submission. In *Proceedings of the Tenth Conference on Machine Translation*, pages 920–925, Suzhou, China. Association for Computational Linguistics.

A Prompts

Complete templates for primary prompts used in this study are shown below. In actual execution, placeholders `{...}` were populated at runtime.

Degraded Translation Candidate Generation Prompt To generate degraded translation candidates, page images, source text, and official translations were input with the following prompt. Degradation types were substituted at runtime as `{degradation_types}`. At runtime, the full-page source and official target text were substituted into `{source_full}` and `{translation_full}`, respectively. `{pairs}` contained bubble-level source-target pairs.

> You are generating natural but visually contradictory degradation candidates for manga translation evaluation.
> The source language is Japanese and the target language is English.
> For each of these degradation types, return up to 3 candidates in descending order of plausibility: `{degradation_types}`.
> Each candidate must change exactly one segment if possible and should only be clearly wrong when considering the image.
> Do not create candidates that are clearly wrong from the source text alone. Avoid obvious text-only mismatches such as changing explicit names, countries, numbers, occupations, or key terms that are directly stated in the source text. Also avoid candidates that become clearly wrong after reading neighboring text segments on the same page.
> Prefer subtle changes where the source text and translation can still look plausible in isolation, but the page image makes the error clear.
> Aim for candidates that remain plausible to a text-only judge, while still being strong enough to force a vision-aware judge to name a concrete translation error.
> Accepted patterns: omitted Japanese subjects resolved differently by the image; gender-neutral references resolved by the depicted character; short replies whose tone depends on expression; object mismatch should target ambiguous references or generic words whose concrete referent is only visible in the image.
> Rejected patterns: replacing an explicit source word with another term; adding a new action or proposition not licensed by the source or neighboring text; changing speaker/addressee when the page text already fixes it; making object/action changes that are already wrong without the image.
> Type-specific constraints: avoid speaker/addressee changes when neighboring text already makes the addressee clear; avoid strong interjections whose tone is obvious from text alone; avoid count changes or lexical narrowing that already create a clear text-only meaning difference; avoid action changes when the source text already states the action or physical state.
> The degraded_text must be the rewritten translation segment only, in the same target language as the original translation. Do not output Japanese source text.
> Return metadata:
> - edit_description
> - visual_evidence_note
>
> Only return JSON.
>
> <input>
> <page_source_full>
> {source_full}
> </page_source_full>
> <page_translation_full>
> {translation_full}
> </page_translation_full>
> {pairs}
> </input>

Common Prompt for MQM Judges In TOJ, `{vision_instruction}` and `{evaluation_procedure}` are left empty. In IAJ-minimal, only page images are added. In IAJ-full, additional instructions described below are embedded in `{vision_instruction}` and

`{evaluation_procedure}`. IAJ-blank uses the IAJ-full prompt with page images replaced by a blank image. MQM judge execution used `temperature=0.1`.

You are an annotator for the quality of manga translation.
Your task is to identify translation errors on a single manga page and assess the overall quality.

{vision_instruction}
{evaluation_procedure}
<input>
<source_language>{src_lang}
</source_language>
<target_language>{tgt_lang}</target_language>
<page_source_full>
{source_full}
</page_source_full>
<page_translation_full>
{translation_full}
</page_translation_full>
{pairs}
</input>

First read the page as connected discourse, then evaluate individual segments.
Evaluate each source/translation segment pair carefully.
Also consider consistency, coherence, and fluency across neighboring segments on the same page.

Use the following error categories:
- meaning_accuracy: Errors where the translation conveys a meaning that is substantially different from the source. Use this category only when no more specific category better explains the error.
- omission_or_addition: Errors where the translation omits source information that should be conveyed, or adds meaning that is not supported by the source.
- reference_and_disambiguation: Errors where the translation incorrectly resolves, over-specifies, or fails to appropriately preserve underspecified reference in the source.
- speaker_and_dialogue_consistency: Errors where the translation is inconsistent with who is speaking, who is being addressed, or how the utterance fits into the local dialogue context on the page.
- tone_and_emotion: Errors where the translation fails to match the emotional tone, intensity, interpersonal stance, or expressive force conveyed by the source and page context.
- action_and_state: Errors where the translation misrepresents an action, posture, physical state, ongoing event, or target of action.
- terminology_and_named_entities: Errors involving incorrect, inconsistent, or contextually inappropriate translation of specialized terms, role labels, names, titles, organizations, locations, or other named entities.
- naturalness_and_register: Errors where the translation is grammatically awkward, stylistically unnatural, or mismatched in register, formality, or social tone for the local context. Do not penalize a translation merely for being colloquial, slangy, or non-literal if it naturally conveys the source nuance in context.
- colloquial_and_cultural_expression: Errors where the translation fails to appropriately render colloquial, stylized, era-specific, socially marked, or culturally grounded expression. Natural slang, casual phrasing, and expressive wording are acceptable when they help preserve the source nuance rather than flatten it into overly formal or literal English.
- other: Errors that materially affect translation quality but cannot be adequately explained by any category above. Use sparingly.

For each error: identify the primary segment_index, optionally context_segment_indices, error_span, explanation, exactly one error_category, severity (major/minor/neutral), and improvement string.

Severity guideline: major = meaning substantially distorted; minor = meaning mostly preserved but suboptimal; neutral = limited impact.

Scoring guideline: Start from 100; subtract more for major than minor/neutral; keep scoring internally consistent across pages; do not double-count cross-segment issues (count a cross-segment issue as one error, not multiple duplicated errors).

Important constraints:
- Only annotate errors that are supported by the source text and translation.
- Do not invent missing context that is not justified by the input.
- Do not over-penalize stylistic variation unless it is clearly unnatural, misleading, or contextually inappropriate.
- Prefer translations that preserve source nuance, voice, and interpersonal texture, not just grammar or dictionary-level literalness.
- If the error is an omission, set “error_span”: “” and describe the missing content in “explanation”. For omission errors, set improvement to the missing phrase or sentence.
- If an issue mainly appears when reading neighboring segments together, anchor it to the most affected segment_index, include the relevant neighbors in context_segment_indices, and explain the cross-segment context in explanation.
- Do not report the same underlying issue multiple times. Merge it into one error anchored to the most affected segment.
- If there is no error, return an empty errors array.

Output requirements:
- Respond with valid JSON only.
- Do not include any text before or after the JSON.
- Use plain text strings without Markdown.

MQM Judge JSON Schema The following JSON schema was inserted at runtime.

```
{
  "$defs": {
    "MQMError": {
      "properties": {
        "segment_index": {
          "minimum": 0,
          "title": "Segment Index",
          "type": "integer"
        },
```

```
        "context_segment_indices": {
          "items": {
            "minimum": 0,
            "type": "integer"
          },
          "title": "Context Segment Indices",
          "type": "array"
        },
        "error_span": {
          "title": "Error Span",
          "type": "string"
        },
        "explanation": {
          "title": "Explanation",
          "type": "string"
        },
        "error_category": {
          "enum": [
            "meaning_accuracy",
            "omission_or_addition",
            "reference_and_disambiguation",
            "speaker_and_dialogue_consistency",
            "tone_and_emotion",
            "action_and_state",
            "terminology_and_named_entities",
            "naturalness_and_register",
            "colloquial_and_cultural_expression",
            "other"
          ],
          "title": "Error Category",
          "type": "string"
        },
        "severity": {
          "enum": [
            "major",
            "minor",
            "neutral"
          ],
          "title": "Severity",
          "type": "string"
        },
        "improvement": {
          "title": "Improvement",
          "type": "string"
        }
      },
      "required": [
        "segment_index",
        "error_span",
        "explanation",
        "error_category",
        "severity",
        "improvement"
      ],
      "title": "MQMError",
      "type": "object"
    }
  },
  "properties": {
    "errors": {
      "items": {
        "$ref": "#/$defs/MQMError"
      },
      "title": "Errors",
      "type": "array"
    },
    "quality_score": {
      "maximum": 100,
      "minimum": 0,
      "title": "Quality Score",
      "type": "integer"
    },
    "visual_caption": {
      "default": "",
      "title": "Visual Caption",
      "type": "string"
    }
  },
  "required": [
    "errors",
    "quality_score"
  ],
  "title": "MQMStructuredResponse",
  "type": "object"
}
```

Additional Instructions for IAJ-full and IAJ-blank In IAJ-full and IAJ-blank, `{vision_instruction}` and `{evaluation_procedure}` in the common prompt are populated as follows:

> [vision_instruction]
> Evaluate the translation with reference to the original manga image. Understand the visual content of the page carefully and use it to assess whether the translation is consistent with the depicted scene and context. Before judging translation quality, first form a concise internal caption of the page. Then identify the visible facts that matter for translation evaluation (who is present, who is likely speaking or being addressed, what object or person a referring expression points to, what action_or_state/emotional stance is depicted). Include the concise caption in the JSON field visual_caption.
>
> [evaluation_procedure]
> - Step 1 (visual grounding): Form a concise internal caption and extract only translation-relevant visual facts.
> - Step 2 (MQM evaluation): Evaluate translation quality using source text, translation text, page-level discourse, and the grounded facts from Step 1.

Semantic Match Classification Prompt Judge output error lists and degradation intents were input to an LLM (`gpt-5.6-luna`) to classify semantic match as `verdict` $\in$ {`same`, `related`, `different`}. An instance was counted as detected when at least one reported error received `verdict=same`; the proportion of detected instances was aggregated as the semantic match rate. `gpt-5.6-luna` was executed with `reasoning_effort=medium`.

> You are judging whether MQM errors semantically catch an intended degradation. Decide if MQM has caught the intended degradation.
>
> Degradation:
> degradation_type: {degradation_type}
> edit_description: {edit_description}
> visual_evidence_note: {visual_evidence_note}
>
> MQM Output:
> mqm_errors: {mqm_errors}
>
> Choose verdict and return concise reason:
> - same: MQM catches essentially the same issue.
> - related: MQM points to a related issue but not the intended issue strongly enough.
> - different: MQM misses the intended issue.

B Dataset Statistics

Dataset details constructed and used in this study are shown in Table 8.

Item	Value
Base Dataset	
Target Manga Work	*Give My Regards to Black Jack* Vol. 1
Page Images (JA)	196
Page Images (EN)	196
Pages with Established Segment Alignment	180
Total Segment Count (180 pages)	974
Mean Segments per Page	5.4
Experiment 1 (Degraded Translation Detection Benchmark)	
Degraded Translation Instances	75
Base Unique Pages	54
Instance Breakdown by Category	
`speaker_or_addressee`	16
`referent_or_object`	27
`tone_emotion_or_style`	26
`action_or_state`	6

Table 8: Dataset statistics.

C Data Availability

The following materials are publicly available in the Hugging Face dataset repository at https://huggingface.co/datasets/joddy/gmr2bj_manga_mt_ja_en:

- Page IDs, page numbers, speech bubble segments, and Japanese-English parallel segment alignment tables for 180 pages.
- 75 degraded translation instances contradicting manga image content used in Experiment 1 (degraded target text, degradation type, intent, target segment, visual cue metadata).

Data Copyright and Licensing. All images in this study originate from data for *Give My Regards to Black Jack* officially distributed for download by Sato Manga Works Ltd., rather than from scans of physical books. Published terms of use allow copying, public transmission, adaptation, and secondary use under specified conditions. Figures and released datasets explicitly state "Give My Regards to Black Jack / SHUHO SATO." Because this permission is specific to this work, it cannot be directly applied to other manga titles. Usage will be reported to Sato Manga Works Ltd. within one month of publication. After verifying license compliance, the page images are released as part of the dataset. Official distribution page: https://densho810.com/free/.

D Experiment 1 Paired Cluster Test

In Experiment 1, the same 75 degraded translation instances were evaluated by both TOJ (no image) and IAJ-minimal (with page image). Because these 75 instances originate from 54 base pages, we aggregated differences by source page to avoid counting instances from the same page as independent data points. We computed cluster sign-flip tests and cluster bootstrap 95% confidence intervals across these 54 page-level differences.

Because the 5 runs represent re-executions over identical inputs, they are not pooled as 375 independent observations. Rather than repeating mean and run-to-run range values from main text Table 2, test results per run are reported in Table 9.

For reference, exact McNemar test p-values based on instance-level paired disagreement are also listed. Because McNemar's test does not account for intra-page correlation, cluster sign-flip tests on base pages serve as our primary inference method.

E Experiment 1: Additional Representative Examples

Figure 3 presents 2 additional representative degraded translations contradicting manga image content: changing to forceful speech contradicting character state (`bj_vol1_114`) and changing to hesitant speech contradicting resolute determination (`bj_vol1_116`).

F Annotator Information

Execution Summary Six annotators evaluated the same set of 40 pages and completed post-task feedback. The 6 annotators were researchers belonging to the same laboratory as the primary author (non-collaborators). Compensation was paid equivalent to an hourly rate of 2,000 JPY. Completing 40 pages required approximately 1.5 to 2 hours per annotator.

Human Verification Task Items Rigid criteria were not enforced, allowing each annotator to judge independently. A screenshot of the annotation UI is shown in Figure 4. Using identical UIs, annotators entered the following items per page:

- **Main Judgment**: Select which judge's output felt superior overall.
  - `prefer_TOJ`: TOJ flags were superior.
  - `prefer_IAJ`: IAJ flags were superior.

Run	TOJ (%)	IAJ-min. (%)	Diff (pp)	TOJ-only	IAJ-min.-only	McNemar p	Cluster p	Cluster 95% CI
1	58.7%	81.3%	+22.7	2	19	0.000221	0.000740	[+11.0, +36.1]
2	58.7%	80.0%	+21.3	2	18	0.000403	0.001700	[+8.3, +35.1]
3	56.0%	81.3%	+25.3	1	20	< 0.0001	< 0.0001	[+13.3, +38.6]
4	57.3%	82.7%	+25.3	1	20	< 0.0001	< 0.0001	[+13.9, +37.8]
5	58.7%	82.7%	+24.0	1	19	< 0.0001	0.000390	[+11.5, +37.7]

Table 9: Significance test results for TOJ vs. IAJ-minimal in Experiment 1. Differences per run are percentage-point differences (IAJ-minimal − TOJ). "TOJ-only detected" indicates instances detected only by TOJ; "IAJ-min.-only detected" indicates instances detected only by IAJ-minimal, aggregated across 75 paired degraded translation instances. McNemar p provides exact instance-level paired reference values. Cluster p represents sign-flip test p-values over 54 page-level clusters, and Cluster 95% confidence interval (CI) reflects bootstrap confidence intervals resampling source pages.

bj_vol1_114 (JA)

bj_vol1_114 (EN)

bj_vol1_116 (JA)

bj_vol1_116 (EN)

Degraded segment
type: `tone_emotion_or_style`
JP "助けてくれよ先生……先生……"
EN (original) "Help me, Doctor... Doctor..."
EN (degraded) "Help me, Doctor! Doctor!"

Degradation intent
"Converted the desperate, trailing-off internal thoughts (indicated by ellipses and fuzzy bubble borders) into assertive, forceful shouts by replacing ellipses with exclamation marks and using direct, clear phrasing."

IAJ error explanation
"The translation uses exclamation marks, conveying a forceful, shouting tone.
However, the source uses ellipses ('……') and the visual context shows an intubated, dying patient, suggesting a weak, desperate plea or internal monologue.
Additionally, 'I AM NOT READY TO DIE' is a more composed phrasing than the raw 'I don't want to die yet' (オレはまだ死にたくねえ), losing the directness and the nuance of 'yet' (まだ)."

Degraded segment
type: `tone_emotion_or_style`
JP "僕は医者だ ・・・・・・!! "
EN (original) "I am! I'm a Doctor!"
EN (degraded) "I... I'm a doctor...!!"

Degradation intent
"Changed the tone from a fierce, defiant declaration of identity to a hesitant, fearful-sounding stutter."

IAJ error explanation
"The repetition of 'I' followed by 'I'm' creates a stuttering effect that suggests hesitation or uncertainty.
This contradicts the character's extremely determined and resolute facial expression.
The Japanese text uses a space for a dramatic pause between the subject and predicate, which is better rendered without repeating the pronoun to maintain the intensity of the declaration."

Figure 3: Representative examples (`bj_vol1_114`, `bj_vol1_116`) of degraded translations contradicting manga image content in Experiment 1. Format matches Figure 2 (JA/EN page images + degraded segment + degradation intent + IAJ-full error explanation). The EN page images are shown for illustration only and were not provided to IAJ. *Give My Regards to Black Jack / SHUHO SATO*

- tie: Difference was small or ranking was difficult. Must answer tie_reason.

- **Confidence**: Select confidence level in main judgment.
  - high: High confidence in main judgment.
  - medium: Medium confidence in main judgment.
  - low: Low confidence; judgment felt unstable.

- **Page-Level Output Validity**: Rate the overall validity of the complete TOJ and IAJ flag sets for the page independently on a 4-point scale.
  - valid: The complete flag set is valid overall.
  - partially_valid: The flag set is directionally valid but needs correction.
  - invalid: The complete flag set is invalid.
  - unclear: Cannot judge due to insufficient information.

- **Image Effect**: image_effect_on_evaluation (multiple-choice). Select how viewing images affected IAJ flags compared to TOJ.
  - enabled_valid_detection: Discovered valid flag via image.
  - prevented_overflagging: Prevented overflagging via image.
  - strengthened_rationale: Strengthened flag rationale via image.
  - visual_misrecognition_or_overinterpretation: Visual misrecognition or overinterpretation.
  - missed_textual_issue: Missed text issue due to viewing image.
  - no_observable_effect: No observable image effect (mutually exclusive).
  - unclear: Cannot judge image effect (mutually exclusive).

- **Image Information Usefulness**: expected_image_helpfulness (single-choice). Rate how useful manga images were for evaluating translation quality on this page.
  - essential: Difficult to judge proper wording/meaning/speaker/tone without image.
  - helpful: Text alone suffices, but image aids translation choices.
  - not_needed: Text alone is sufficient.
  - unclear: Cannot judge image usefulness.

  This item represents potential image usefulness as perceived by evaluators.

- **Tie Reason**: Single-choice reason when selecting tie.
  - both_valid: Both TOJ and IAJ flags are valid.
  - both_partially_valid: Both flags are partially valid.
  - both_invalid: Both flags are invalid.
  - difference_too_small: Output difference is too small to rank.
  - rubric_dependent: Judgment depends on which evaluation criterion is prioritized.
  - other: Other reason (requires text comment).

Post-Task Feedback Items After completing 40 pages, annotators answered post-task feedback items on 5-point Likert scales (1 = strongly disagree, 5 = strongly agree):

- **5-Point Likert Items**:
  - overall_difficulty: Comparing TOJ and IAJ flags to judge which was better was difficult.
  - rubric_sufficiency: Procedure explanations and item descriptions were sufficient to judge consistently.
  - priority_clarity: Priority rules were clear when source fidelity, naturalness, and visual fit conflicted.
  - preference_label_clarity: Criteria for selecting prefer_TOJ, prefer_IAJ, and tie were clear.
  - example_helpfulness: Concrete label examples or boundary cases would have aided decision-making.

- **Hardest Annotation Field** (hardest_annotation_field, single-choice):

- overall_preference: Main judgment.
- preference_confidence: Judgment confidence.
- TOJ_validity: TOJ page-level output validity rating.
- IAJ_validity: IAJ page-level output validity rating.
- tie_reason: Reason for selecting tie.
- image_effect_on_evaluation: Image effect on IAJ flags.
- expected_image_helpfulness: Image usefulness for translation evaluation.
- none: No item was particularly difficult.

- **Reasons for Difficulty** (difficulty_reasons, multiple-choice):
  - small_difference: Output difference between TOJ and IAJ was small.
  - both_valid: Both flags were valid.
  - both_problematic: Both flags had flaws.
  - fidelity_vs_naturalness: Conflict between source fidelity and target naturalness.
  - text_vs_visual_fit: Difficulty evaluating text vs. visual fit.
  - severity_threshold: Severity thresholds were difficult to calibrate.
  - insufficient_context: Context was insufficient for judgment.
  - complex_judge_outputs: Judge output content was complex.
  - unclear_label_boundaries: Label boundaries were unclear.
  - image_quality: Image quality hindered judgment.
  - ui_issue: UI display or operation issues.
  - other: Other reasons.

- **Most Prioritized Criterion** (prioritized_criterion, single-choice):
  - source_fidelity: Source fidelity.
  - translation_naturalness: English translation naturalness.
  - visual_consistency: Consistency with manga image.
  - error_severity: Flagged error severity.
  - overall_balance: Overall balance across criteria.
  - case_by_case: Evaluated on a case-by-case basis.
  - unclear: Could not set consistent criteria.

- **Desired Guidance Types** (requested_guidance, multiple-choice):
  - label_examples: Concrete examples for each label.
  - boundary_examples: Examples of boundary cases.
  - tie_examples: Concrete examples of tie cases.
  - priority_rules: Priority rules when criteria conflict.
  - severity_rules: Guidelines for severity assessment.
  - image_effect_examples: Concrete examples of image effects.
  - full_walkthrough: End-to-end task walkthrough.
  - more_context: Additional context.
  - none: No additional guidance needed.

- **Free Comments** (free_comment, optional): Free-text input for hesitation details, UI feedback, or requested explanations.

G Automatic Matching and Score Aggregation

To detail the automatic output differences reported in §5, we aggregated TOJ / IAJ-full flag counts on official translations across 180 pages (main text Table 4).

Common Flag Matching Procedure Common flags in this paper are established by **rule-based automatic matching rather than human annotation**. Specifically, candidate TOJ/IAJ error pairs within the same segment_index are created, taking the maximum of explanation token Jaccard similarity and SequenceMatcher string similarity as the base score. We add $+0.15$ for matching error_category and $+0.05$ for matching severity, capping the score at 1.0. For the SequenceMatcher component, explanations are

Figure 4: Annotation UI screenshot. Left panel: JA/EN page images and comparison of source and translation segments. Middle panel: TOJ/IAJ flags. Right panel: annotation fields including main judgment (`prefer_IAJ`/`prefer_TOJ`/`tie`), confidence, page-level output validity, image effect, image usefulness, and comments. Images expand on click.

normalized with NFKC normalization, lowercasing, and removal of non-word characters; therefore, exact matches after this normalization already have a base similarity of 1.0. Candidates are sorted by score descending, executing 1-to-1 greedy matching while excluding previously matched errors. Pairs with `score < 0.45` are rejected without an exact-match exception. Exact normalized explanations still require 1-to-1 pairing within the same segment and do not guarantee human semantic identity. Accepted pairs are stored as `same_issue` in `common_errors`. Unmatched flags are output as `text_only_errors` or `vision_only_errors`. Restricting matching candidates to identical segments and combining string similarity with `error_category` and `severity` allows this matching to serve as a practical classification separating common and unique flags. However, common classification does not guarantee strict semantic identity or human ground truth, and inter-annotator agreement (IAA) for common matching was not separately calculated. Accordingly, TOJ-only / IAJ-only / common aggregations in §5 should be interpreted as diagnostic statistics comparing judge output trends rather than human gold labels.

G.1 Model-Reported Scores and Auxiliary Score Deficits

The `quality_score` in this study is an auxiliary model-reported score directly output by judges alongside error lists, rather than a standard MQM quality score calculated from fixed severity weights. Prompts instructed judges to start from 100, penalize `major` more than `minor` or `neutral`, and avoid double-counting, but did not fix numerical penalty amounts per severity or aggregate formulas. Therefore, this score should not be interpreted as a raw MQM penalty sum or a metric mechanically recomputed from error counts.

For comparison, we define the auxiliary score deficit from model-reported scores as $D_p^{\text{judge}} = 100 - S_p$ per page. Mean score deficits reported in Table 4 and derived from the mean scores in Table 11 represent D_p^{judge} averaged across 180 pages.

G.2 Fixed-Weight Scoring Based on MQM Severity and Normalized Penalties

In addition to `quality_score`, we conducted an auxiliary analysis aggregating reported severity labels with fixed weights. Letting $n_{p,\sigma}$ be the count of flags with severity σ on page p, we assign weights of 0, 1, and 5 to neutral, minor, and major severities, respectively, defining raw MQM penalty as:

$$D_p^{\text{MQM}} = 0n_{p,\text{neutral}} + 1n_{p,\text{minor}} + 5n_{p,\text{major}} \quad (1)$$

To account for text length variation across pages, we normalize penalty by official target word count W_p:

$$E_p^{\text{word}} = 100\frac{D_p^{\text{MQM}}}{W_p} \quad (2)$$

representing penalty points per 100 target words. The primary aggregation is computed as a dataset micro-average across all 180 pages:

$$E_{\text{micro}}^{\text{word}} = 100\frac{\sum_p D_p^{\text{MQM}}}{\sum_p W_p} \quad (3)$$

where lower values indicate smaller penalties.

Across all 3 models (Gemini, GPT-5.4-mini, Qwen3-VL), IAJ-full yielded lower normalized penalty-point values than TOJ. The difference from TOJ to IAJ-full (IAJ-full − TOJ) was -1.23, -1.05, and -3.93 penalty points per 100 target words, respectively. Page counts where IAJ-full penalty was lower, TOJ penalty was lower, or tied were 62/21/97 for Gemini, 70/50/60 for GPT-5.4-mini, and 92/42/46 for Qwen3-VL. Sensitivity checks across 4 different denominator choices (per 100 target words, per 10 segments, per 100 source characters, and per 100 target words excluding short target pages) consistently preserved directional gains across all 3 models. However, this does not directly validate individual flag accuracy or human correlation superiority.

G.3 Correlation Between Severity-Weighted Auxiliary Penalties and `quality_score`

To examine the alignment between auxiliary penalties and `quality_score`, we calculated page-level correlations between `quality_score` and sign-flipped normalized MQM penalties ($-E_p^{\text{word}}$). Results are shown in Table 10.

Pooled across all conditions, Pearson $r = 0.492$ and Spearman $\rho = 0.817$, showing general ranking alignment between model-reported `quality_score` and severity-weighted auxiliary penalties. However, Pearson correlations varied by model and condition (0.348–0.752), indicating they cannot be treated as identical scale metrics. Removing text volume effects via normalization also weakened Pearson correlation relative to raw penalty sums.

Model	Condition	Pearson r	Spearman ρ
`gemini-3-flash-preview` (Main)	TOJ	0.402	0.762
`gemini-3-flash-preview` (Main)	IAJ-full	0.348	0.717
`gpt-5.4-mini` (Auxiliary)	TOJ	0.557	0.873
`gpt-5.4-mini` (Auxiliary)	IAJ-full	0.752	0.881
`qwen3-vl-32b-instruct` (Auxiliary)	TOJ	0.467	0.803
`qwen3-vl-32b-instruct` (Auxiliary)	IAJ-full	0.391	0.693
3 Models, 2 Conditions (Pooled)	All Conditions	0.492	0.817

Table 10: Page-level correlation between `quality_score` and sign-flipped normalized MQM penalty ($-E_p^{\text{word}}$). Each model and condition has $n = 180$; pooled values have $n = 1080$.

Comparing improvement magnitudes from TOJ to IAJ-full between $\Delta S_p = S_{p,\text{IAJ}} - S_{p,\text{TOJ}}$ and $G_p = E_{p,\text{TOJ}}^{\text{word}} - E_{p,\text{IAJ}}^{\text{word}}$, Pearson/Spearman correlations were $0.779/0.783$ for `gemini-3-flash-preview`, $0.586/0.731$ for `gpt-5.4-mini`, and $0.726/0.820$ for `qwen3-vl-32b-instruct` (pooled 540 pairs: $0.664/0.773$). Thus, both metrics demonstrate moderate-to-strong alignment in improvement direction, though this correlation reflects convergent metric behavior rather than human ground-truth alignment.

H Experiment 2 Model-Reported Score Results

In Experiment 2, we conducted automatic MQM evaluation with the primary model, `gpt-5.4-mini`, and `qwen3-vl-32b-instruct` on 180 official translation pages. Table 11 compares reported page scores (`quality_score`). Across all three models, IAJ-full yielded higher mean model-reported scores and lower auxiliary score deficits than TOJ. The number of pages where IAJ-full scored higher than TOJ also exceeded the reverse. However, this is an auxiliary diagnostic showing judge output distribution shifts on official translations, not proof of page-level output validity or universal superiority.

Error volume aggregations for auxiliary models in the format of main text Table 4 are shown in Table 12. In `gpt-5.4-mini`, IAJ-full produced fewer total error flags than TOJ, and `qwen3-vl-32b-instruct` similarly yielded fewer total and major flags. However, in Qwen, minor flags for IAJ-full exceeded TOJ, showing severity-level shifts vary by architecture.

I Detailed TOJ Coverage in IAJ-full

TOJ Flag Coverage by IAJ-full Regarding coverage mentioned in §5, mathematical definition, target set, micro-average, and page-level distributions are detailed in Table 13. For each page p, let common_p and TOJ-only_p denote the numbers of retained and non-retained TOJ flags, respectively. Page-level coverage is defined as:

$$c_p = \frac{\text{common}_p}{\text{TOJ-only}_p + \text{common}_p}. \tag{4}$$

The dataset-level micro-average is defined as:

$$c_{\text{micro}} = \frac{\sum_p \text{common}_p}{\sum_p (\text{TOJ-only}_p + \text{common}_p)}. \tag{5}$$

We additionally report the page-level macro-average and median of c_p. Aggregated over 160 pages with at least 1 TOJ error flag (excluding 20 pages with 0 TOJ flags), nearly 30% (45 pages; 28.1%) exhibited 0% coverage (IAJ retained zero TOJ flags). This indicates that evaluation perspectives between TOJ and IAJ diverge significantly on select pages.

Model	IAJ Higher	TOJ Higher	Tied	Mean IAJ	Mean TOJ	Mean Diff
`gemini-3-flash-preview`	81 (45.0%)	32 (17.8%)	67 (37.2%)	91.9	89.8	+2.2
`gpt-5.4-mini`	90 (50.0%)	43 (23.9%)	47 (26.1%)	87.9	86.4	+1.5
`qwen3-vl-32b-instruct`	102 (56.7%)	24 (13.3%)	54 (30.0%)	85.5	80.1	+5.4

Table 11: Experiment 2 score summary across the primary and auxiliary models. For 180 official translation pages, the table shows page counts where the IAJ-full model-reported score exceeded TOJ, vice versa, ties, and mean model-reported scores (`quality_score`). Mean difference reflects average unrounded score difference (IAJ − TOJ).

Metric	gpt TOJ	gpt IAJ-full	qwen TOJ	qwen IAJ-full
All 180 Pages				
Total error flags	463	444	568	547
Mean / page	2.57	2.47	3.16	3.04
Major error flags	136	123	212	146
Mean major / page	0.76	0.68	1.18	0.81
Minor error flags	301	285	348	375
Mean minor / page	1.67	1.58	1.93	2.08
Neutral error flags	26	36	8	26
Mean neutral / page	0.14	0.20	0.04	0.14
Page-Level Error Distribution				
0-error pages	24 (13.3%)	36 (20.0%)	11 (6.1%)	4 (2.2%)
1+ error pages	156 (86.7%)	144 (80.0%)	169 (93.9%)	176 (97.8%)
0-major pages	84 (46.7%)	92 (51.1%)	52 (28.9%)	85 (47.2%)
0-error in both judges	24 (13.3%)		3 (1.7%)	

Table 12: Error volume reported by auxiliary models on 180 official translation pages. Total error flags include `major`, `minor`, and `neutral`. gpt represents `gpt-5.4-mini`; qwen represents `qwen3-vl-32b-instruct`.

Metric / Page Group	Value / Pages
Micro-Average Coverage (166 / 318 flags)	52.2%
Page-Level Macro-Average Coverage (n=160)	53.1%
Page-Level Median Coverage	50.0%
100% Coverage (Fully Retained) Pages	54 pages (33.8%)
Partial Coverage ($0 < c_p < 1$) Pages	61 pages (38.1%)
0% Coverage (Fully Suppressed) Pages	45 pages (28.1%)

Table 13: Coverage of TOJ error flags by IAJ-full and page-level distributions.