LESS IS AN ACT: PROBING SELECTIVE JUDGMENT IN VISION-LANGUAGE MODELS THROUGH IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In-image machine translation (IIMT) has advanced rapidly. Existing systems and benchmarks are largely concerned with two things: the visual quality of the rendered output and the fidelity of the translation itself. However, we argue these objectives sit closer to a design-and-application problem than to a core question about model capability: when a user simply asks to "translate this image," does the model understand *what* should be translated and what left untouched? Brand names, ingredient codes, and text physically part of an object must be preserved for translation to serve the user. We find that state-of-the-art vision-language models (VLMs), given no guidance, systematically *mistranslate* what should be left untouched, most often by *over-translating*: rewriting brand names and other text that competent translators would preserve. We introduce ***selective translation***, a new task that shifts the focus from translation quality to selective judgment and ***restraint***: deciding, for each piece of text, whether to translate or preserve it. We take a new approach to in-image translation, building **SPOT-Bench**, a benchmark that evaluates what a model chooses to translate rather than how well it translates, with *constructional* ground truth, derived from how each image is made rather than annotated by taste, across advertisements, news, and captioned video. Evaluating current proprietary and open VLMs, we find a large and consistent gap between what models *know* the rules to be and what they *do*: they fail not from missing knowledge but from failing to apply it when acting on the image. Our analysis shows that making the translate/preserve decision explicit substantially narrows this gap, pointing toward more selective, context-aware VLMs.

1 INTRODUCTION

Reading and translating the text embedded in an image is a task people perform effortlessly. A traveler glancing at a foreign advertisement, a shopper scanning a product label, or a viewer watching a subtitled clip all do something subtle without thinking about it: they decide, text by text, what to render into their own language and what to leave exactly as it is. The marketing copy on a poster should travel; the brand name printed on the jar should not. This decision is so automatic for humans that it is easy to overlook that it is a decision at all.

Recent progress in in-image machine translation (IIMT) has been substantial. Systems built on large language models and text-guided diffusion models can now detect text in an image, translate it, and re-render the result so that it blends into the original layout (Qian et al., 2024; Tian et al., 2025a; Kaing et al., 2025). The benchmarks that accompany these systems reflect their goals: they measure translation quality with metrics such as BLEU and COMET (Papineni et al., 2002; Rei et al., 2020), and visual quality with statistics such as FID (Heusel et al., 2017). These are the right targets for a deployable product. They are also closer to a *design-and-application* problem than to a core question about what a model understands. Rendering translated glyphs which harmonize with the surrounding image is a valuable engineering achievement, but it is a question of aesthetics and fusion. Here, we ask a question these benchmarks leave unmeasured, one that does not depend on anyone's particular sense of visual taste.

Gaps in current IIMT evaluation Two observations motivate our study. First, strong IIMT numbers do not obviously translate into user satisfaction: the gap between benchmark performance and

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.

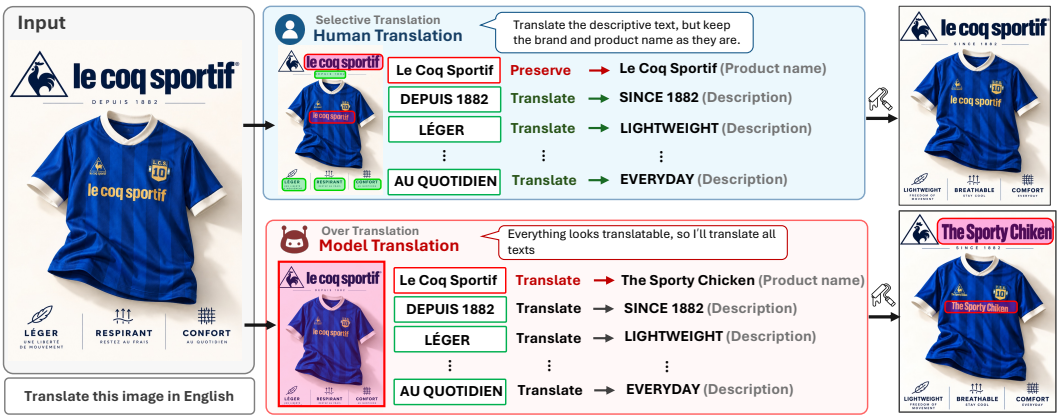


Figure 1: **Selective translation.** Asked to translate the image, a human preserves the brand *Le Coq Sportif* while VLMs over-translate it into "The Sporty Chicken," corrupting the product's identity. We benchmark whether models know *what* to translate.

real usefulness is less about the smoothness of the translation than about *what was translated in the first place*. Second, prior IIMT work assumes the target is given, translating the whole image or a user-specified region (Qian et al., 2024; Tian et al., 2025a), leaving the model's own decision of what to translate unexamined. Without such guidance, current VLMs often act without regard to context: they *over-translate*, re-translating text already in the target language and altering brand names and proper nouns that should be preserved, losses that standard metrics never penalize.

This is not a cosmetic error but a question of **factual integrity**: over-translation can silently distort facts, falsify safety-critical labels, and misrepresent people, places, and events. A translated ingredient list or dosage can turn a correct medical label into a dangerous one. A mistranslated sign inside a news photograph, for instance, rendering a landmark as if it stood in another country, can misattribute where an event took place. As image translation is deployed at scale in cameras, e-commerce, and media, these are no longer edge cases: a fluent output can be a **factual and even physical corruption** of the original.

Knowing when not to translate. To probe a model's judgment, it is not enough to ask whether it *can* translate, but whether it knows *when not to*. There is a distinction between knowing a rule and applying it: a model may recognize that brand names should be preserved, yet still translate them when tasked with translating an image in a specific context. This second ability, deciding what to translate and what to leave untouched, has not been systematically evaluated for in-image translation.

Why translation, and why images. Translation is a particularly clean test for selective decision-making: choosing not to translate certain elements, such as a brand name, is not a limitation but professional practice, and translation studies have long treated the handling of proper nouns as a deliberate decision rather than a matter of style (Newmark, 1988; Venuti, 2012). Regulatory standards for cosmetic ingredient names (European Parliament and Council of the European Union, 2009) further fix which terms must remain unchanged, giving us objective ground truth.

Images add a critical layer: the same word may demand different choices depending on its visual placement, such as inside a logo versus in a sentence, or printed on an object versus overlaid as a caption. Thus, correct judgment depends on reading the visual context, not just the text.

Our key contributions are summarized as follows:

- We introduce ***selective translation***, a new task that asks whether a model can decide what should be translated and what should remain unchanged when translating an image, shifting the focus from translation quality to selective judgment and restraint.
- We introduce a **benchmark for evaluating selective translation** in current VLMs, with constructional ground truth across advertisements, news, and captioned video, enabling systematic evaluation of the gap between what models know, decide, and ultimately do.
- We provide a **systematic analysis of proprietary and open VLMs**, revealing consistent over-translation and a substantial knowing–doing gap that shrinks when translate/preserve decisions are made explicit.

2 Related Work

In-image and document machine translation. Researchers have developed methods to translate text in images while preserving the image's visual appearance, starting with early end-to-end methods (Mansimov et al., 2020) and progressing to more recent systems. Qian et al. (2024) propose AnyTrans, which combines optical character recognition (OCR), an LLM translator, and diffusion-based text fusion, and this system is evaluated using the MTIT6 dataset. Tian et al. (2025a) introduce PRIM, a real-world benchmark that measures translation quality (using BLEU and COMET) and image fidelity (using FID). Other studies have extended IIMT to cases involving complex backgrounds (Tian et al., 2025b; Lan et al., 2023), real-time images and video (Futeral et al., 2025), document-level translation, and have created multilingual benchmarks in order to improve data quality and to identify failures (Wang et al., 2025; Zouhar et al., 2026; Lyu et al., 2026; Li et al., 2025). End-to-end systems such as Translatotron-V (Lan et al., 2024) generate the translated images directly and use location-aware metrics. In all of this work, we assume the text to be translated is already specified: systems translate all detected text or text in user-specified areas, and they are assessed on both translation and visual quality. We instead investigate whether a model can decide for itself *what* to translate when given an open-ended prompt, thereby separating its understanding of the rule from the way it behaves when carrying out the translation.

Rendering translated text relies on scene-text understanding (Singh et al., 2021; Liu et al., 2020) and visual text editing (Tuo et al., 2024; Chen et al., 2023). To isolate model performance, our main metric holds image pixels constant and varies only the localizable layer, ensuring that decision errors reflect model accuracy rather than rendering artifacts. Rendering quality is reported only as a secondary diagnostic.

Selective prediction and abstention. A separate line studies when a model should decline to act. Selective prediction, where a model may abstain rather than commit to a low-confidence output, has a long history in classification (El-Yaniv et al., 2010; Geifman & El-Yaniv, 2017) and question answering (Kamath et al., 2020), and VQA systems are rewarded for abstaining over guessing wrong (Whitehead et al., 2022). These frame abstention as a reaction to *uncertainty*. Our case differs: a VLM is fully able to translate a brand name, and the question is whether it exercises *restraint* when doing so is wrong. Selective translation asks not "do you know enough to answer?" but "do you know you should hold back, and do you act on it?", which current models do poorly even when they can state the rule.

The role of the image and the distinction between knowing and doing. Although multimodal machine translation (MT) often incorporates images to resolve ambiguity (Specia et al., 2016), recent studies indicate that images are frequently unnecessary for translation (Caglayan et al., 2019; Wu et al., 2021; Futeral et al., 2025). However, for in-image text, visual context can determine whether translation is required, as identical strings may require different actions depending on their location. Neglecting visual cues causes models to over-translate, irrespective of output quality. This work also addresses the gap between a model's knowledge and its actions. Prompting a model to reason before responding can be beneficial (Wei et al., 2022; Zhou et al., 2022; Press et al., 2023), but the reasoning provided is not always consistent with the model's actual behavior (Lanham et al., 2023). A similar discrepancy is observed in embodied settings, where models may fail to abstain or act on recognized information. (Yeke et al., 2026; Wang et al., 2026). The same knowing-doing gap is documented in a purely linguistic setting: models can articulate the translate/preserve rule and label spans accurately, yet fail to implement this rule during translation. This reframes over-translation as a failure to connect knowledge with action.

3 SPOT-Bench: A Benchmark for Selective Translation

We introduce SPOT-Bench (Selective Preserve-Or-Translate), a benchmark that assesses whether a vision-language model, when asked to translate an image, decides correctly *what* to translate and what to leave unchanged. Existing in-image translation resources assume the translation target is given and evaluate only rendering quality; to our knowledge none assess the decision itself. Our benchmark isolates this decision and traces it through several stages, from stating the rule to preserving identity-bearing text while translating. Throughout, the model receives *only the image* and a minimal instruction; regions are not highlighted and rules are not provided at test time. We describe the tasks and metrics (§3.1), how we obtain ground truth without subjective per-image annotation (§3.2), and the three media and why we chose them (§3.3).

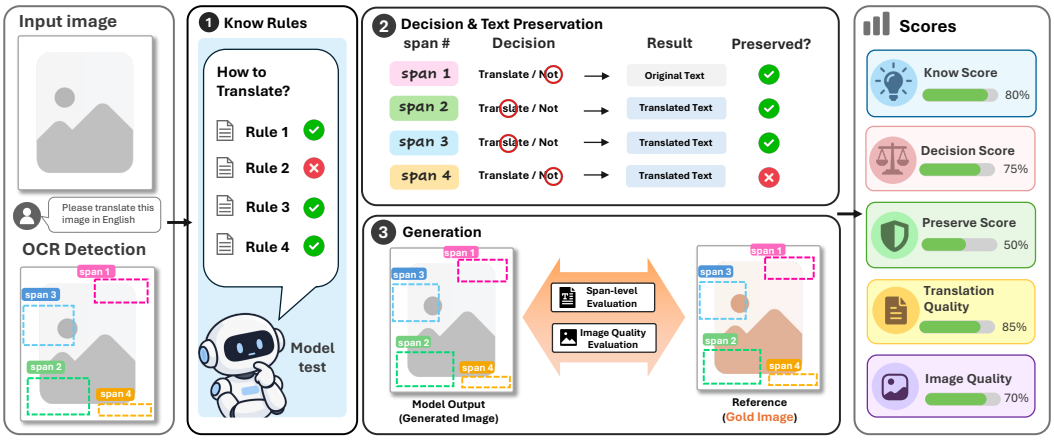


Figure 2: **Evaluation pipeline.** We detect text spans and score whether the model knows the translate/preserve rules (KNOW), decides correctly (DECISION), preserves identity text when translating (PRESERVE), and keeps it intact in the generated image.

3.1 TASKS AND METRICS

An image contains visible text *spans* $S = \{s_1, \ldots, s_n\}$. For a target language, each span has a gold decision $d(s_i) \in \{\text{TRANSLATE}, \text{PRESERVE}\}$: marketing copy, headlines, and body text are translated so the message reaches the reader, while brand names, ingredient codes, proper nouns, and text physically part of an object or scene are preserved. The core input is always the image alone; the model must detect the text itself, decide per span, and act. The central design principle is to separate what a model *knows* from what it *does*, and to trace the decision into both the produced text and image. We define four tasks, each with its own metric.

KNOW asks whether the rule is available to the model at all, without the page image, through two probes. KNOW-*rules* presents a fixed battery of nine constructional rule statements (Appendix A), each with a known answer, on a blank image and scores yes/no agreement, testing verbal policy. KNOW-*spans* presents the gold span list without the image and asks for a translate/preserve label per span, testing whether the model can apply the rule to concrete strings without seeing them. Measuring agreement at both steps establishes whether the rule is genuinely available to the model, so that any later failure to preserve reflects a breakdown in acting on it rather than in holding it.

DECISION gives only the image and asks the model to detect each text region and label it TRANSLATE or PRESERVE. Since the model must both read and judge, we align its detected regions to the gold spans and score the decision on matched spans; any gold span it fails to detect counts as a failure, so a model cannot appear accurate by simply skipping hard text. We report per-span accuracy and a strict per-item score (all spans correct).

ACT gives the image and a translate instruction; the model must detect the text, translate the TRANSLATE spans, and leave the PRESERVE spans unchanged. This is the end-to-end behavior a user would get. We run it *unguided* (image and instruction only) and *guided* (additionally told, per detected span, whether it is translate or preserve, never the gold target strings). Contrasting the two isolates *binding* failures, where the model knows the rule but does not apply it to the image, from *execution* failures, where it applies the rule but translates poorly. The per-item score is all-or-nothing; per-span rates and over/under-translation counts are diagnostics.

PRESERVE (image) extends this to systems that render a translated image: we ask them to translate the image with no boxes, OCR the output, and align it to the gold spans to check whether PRESERVE spans survive in pixels, not only in an intermediate text response.

A model may score high on KNOW yet fail at ACT, stating the rule but over-translating when it processes the image. This discrepancy, not translation quality, is the phenomenon the benchmark is designed to reveal. Before asking whether models *apply* the rules, we verify they *hold* them: Table 1 reports per-model agreement under the KNOW-rules probe. Agreement is high across models, which is what makes the later ACT failures meaningful, they are not disagreement with our labeling scheme but an inability to bind a rule the model demonstrably holds to the visual input.

Table 1: Rule availability (KNOW-rules, verbal yes/no, no page image). High values indicate the rule is available independent of vision, so downstream failures are binding failures, not disagreement with the gold. Rule statements are in Appendix A.

Model	A1 brand	A2 product	A3 mktg	A4 pack	A5 INCI	C1 sub	C3 dieg.	N1 head	N2 photo	Acc.
Gold	P (NO)	P (NO)	T (YES)	P (NO)	P (NO)	T (YES)	P (NO)	T (YES)	P (NO)	
InternVL	NO	NO	YES	**YES**	**YES**	YES	NO	YES	NO	7/9
Qwen3-VL	NO	NO	YES	**YES**	NO	YES	NO	YES	NO	8/9
Qwen3.8	NO	NO	YES	NO	NO	YES	NO	YES	NO	9/9
Gemma-12B	NO	NO	YES	NO	NO	YES	NO	YES	NO	9/9
Gemma-27B	NO	NO	YES	NO	NO	YES	NO	YES	NO	9/9
Pixtral	NO	NO	YES	NO	NO	YES	NO	YES	**YES**	8/9
Astra	NO	NO	YES	NO	NO	YES	NO	YES	NO	9/9
Claude Opus 5	NO	NO	**NO**	NO	NO	YES	NO	YES	NO	8/9

3.2 CONSTRUCTIONAL GROUND TRUTH

Whether a given text should be translated can appear subjective. We avoid per-image subjective labeling and instead derive the gold decision from structure. Each stimulus is two layers: a *localizable* layer (copy, headlines, captions) and an *identity* layer (brands, ingredient codes, proper nouns, in-scene signage). The localizable layer is TRANSLATE; the identity layer is PRESERVE. The decision thus follows from how the image is constructed or aligned, not from annotator preference. This rule is grounded in professional practice: localization does not translate brand names, and regulatory standards fix certain surfaces, such as international cosmetic ingredient names, across languages (Newmark, 1988; Venuti, 2008). To confirm the rule matches expert judgment, three professional translators independently annotated **600** images; their agreement validates the rule, which is then applied to label the remaining images consistently.[1] The ground truth is therefore both defensible and scalable.

3.3 CATEGORIES AND WHY WE CHOSE THEM

Selective translation is only testable where text that must be translated and text that must be preserved co-occur densely in one image, exactly the setting existing IIMT benchmarks omit by assuming the target is given. We build three media that stress different reasons for preservation, moving it beyond a single easy cue (capitalized proper nouns) to a spectrum requiring knowledge and visual context. *Advertisements* are the densest case: a headline that must travel sits beside a brand wordmark, an ingredient list, and a net-weight figure that must not. We compose a text-free product image with an authored text layer over three sub-domains with distinct pressures: cosmetics (INCI ingredient names, legally fixed across languages), medical/health (regulated claims and dosages where mistranslation is unsafe), and fashion (global brand names that read like ordinary words); because we place the text, layout is fixed across languages and errors cannot be blamed on rendering. *News* adds text physically *in a photograph*: a correct localization rewrites the headline and body but leaves a ministry sign untouched. We crop real articles from parallel official editions and read the gold off the aligned editions; each article has four spans (headline, body, masthead, in-photo sign), making the sign a controlled probe. *Captioned video* is the purest test of grounding, since the same string is translate or preserve by position alone: an overlaid subtitle is localizable, while identical words on a shop sign within the scene are diegetic and must be preserved.

Scale and construction. In total, the benchmark comprises **5,000** images spanning three media and their sub-domains (Figure 3): advertisements (2,000: cosmetics, food, clothing, and others), news (2,000), and captioned video (1,000: music video, movie, YouTube, and documentary). Images are rendered across **six languages** (English, Korean, Japanese, French, Spanish, and Chinese); not every image exists in all six, giving over **20,000** image-language evaluation instances. Because a single image is evaluated across multiple target languages, the benchmark also probes a revealing failure mode: text that is *already* in the requested target language, which should simply be left unchanged, yet models often translate it again. Every image is synthetically composed under our

[1] Inter-annotator agreement and adjudication are in Appendix B.

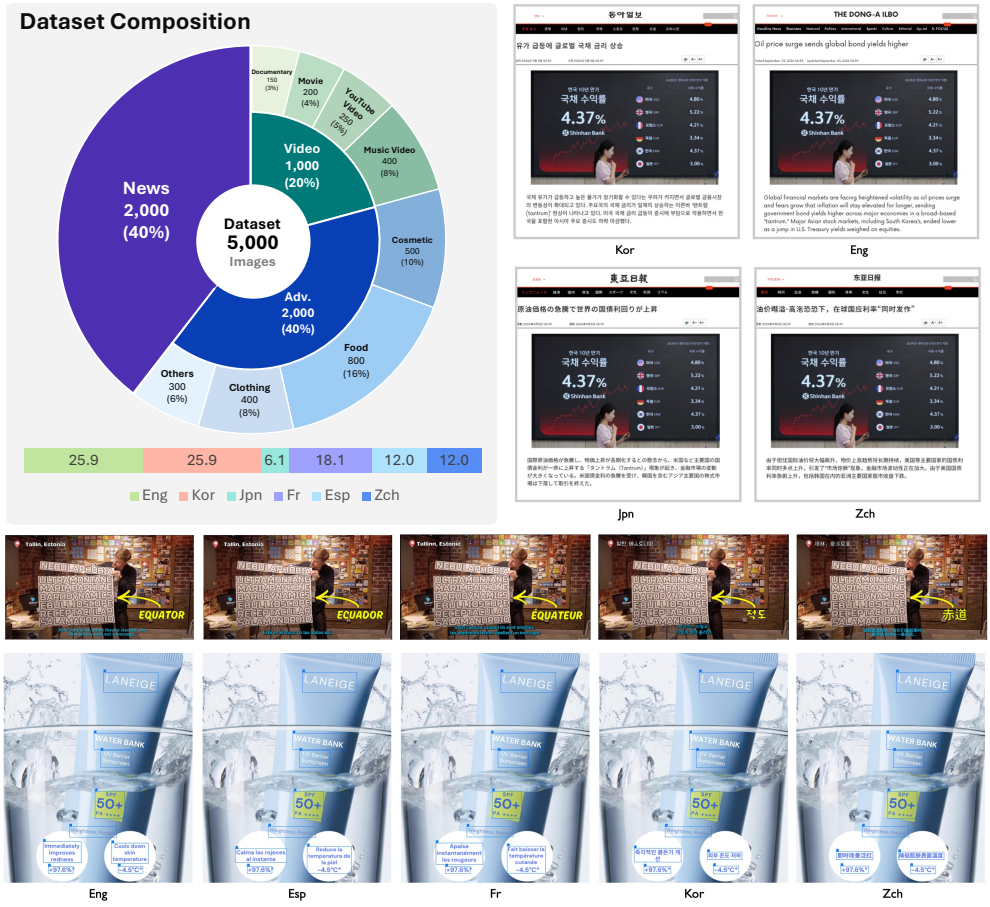


Figure 3: **Dataset composition** and multilingual examples. Our dataset comprises 5,000 images across advertisements, news, and video, covering six languages.

control, so the gold decision follows from construction rather than post-hoc labeling; every one is then manually inspected span by span, verifying the text, its layer, and its translate/preserve label by hand and correcting any error before inclusion. Selective translation is thus not a niche concern but a pervasive one: wherever an image carries both a message and an identity, translating everything corrupts the identity, precisely what a translation-quality benchmark cannot measure and this task is built to expose.

4 Experimental Setup

Models. We evaluate open and proprietary vision-language models spanning a range of scales, plus two systems from related multimodal MT work. Open VLMs include InternVL3.5-8B, Qwen3-VL-8B, Pixtral-12B, Gemma-3-12B, Qwen3.8-27B, and Gemma-3-27B. As related baselines we include ZeroMMT (Futeral et al., 2025) (image-conditioned text MT) and Translatotron-V (Lan et al., 2024) (end-to-end in-image MT). Proprietary models include recent generations of GPT, Claude, and Gemini.[2] Unless noted, VLMs are run with deterministic decoding (temperature 0) and reasoning traces disabled.

Tasks and scoring. Each task uses a single fixed prompt (Appendix C), and the model receives only the image and a minimal instruction. For DECISION and ACT we align the model's detected text to the gold spans by normalized string match; a PRESERVE span passes if left unchanged, a TRANSLATE span if its surface changes and matches the gold target under a lenient similarity threshold. In the guided ACT condition the model additionally receives, per span, whether it is

[2]Exact versions, prompts, and compute are in Appendix C.

Table 2: Main result across domains (full dataset). For each domain we report KNOW (P/T without image), DECISION (P/T with image), and PRESERVE (gold-preserve accuracy at ACT). Red values are the drop from DECISION to PRESERVE: models label a span preserve yet fail to keep it unchanged when translating.

	News			Ads			Video			Avg		
Model	Know	Dec.	Pres.	Know	Dec.	Pres.	Know	Dec.	Pres.	Know	Dec.	Pres.
InternVL3.5-8B	86.0	81.8	52.7 (−29.1)	75.5	60.3	45.6 (−14.7)	70.8	84.1	56.4 (−27.7)	77.4	75.4	51.6 (−23.8)
Qwen3-VL-8B	88.3	83.9	54.2 (−29.7)	76.4	69.1	54.0 (−15.1)	74.5	84.4	55.2 (−29.2)	79.7	79.1	54.5 (−24.6)
Qwen3.8-27B	87.8	85.0	60.5 (−24.5)	79.1	83.1	59.3 (−23.8)	78.9	85.6	57.8 (−27.8)	81.9	84.6	59.2 (−25.4)
Gemma-3-12B	79.9	75.2	40.6 (−34.6)	80.4	75.6	55.2 (−20.4)	72.2	80.1	51.1 (−29.0)	77.5	77.0	49.0 (−28.0)
Gemma-3-27B	81.7	80.4	42.5 (−37.9)	86.8	79.5	59.4 (−20.1)	74.0	85.1	56.4 (−28.7)	80.8	81.7	52.8 (−28.9)
Pixtral-12B	76.1	71.7	47.6 (−24.1)	79.8	69.7	48.5 (−21.2)	70.0	79.0	50.5 (−28.5)	75.3	73.5	48.9 (−24.6)
AnyTrans	–	66.7	41.2 (−25.5)	–	62.3	44.8 (−17.5)	–	83.8	52.3 (−31.5)	–	70.9	46.1 (−24.8)

Table 3: Proprietary systems on the API subset (200 ads / 200 news / 200 video). We report DECISION, translation quality on translate spans (BLEU/COMET), and image-level PRESERVE (identity text intact in the rendered output). Final images are produced by a shared renderer (GPT Image 2).

	Text			Image (rendered)		
Model	KNOW	BLEU	COMET	BLEU	Decision	Preserve
Astra (GPT 6.0)	92.5	**40.4**	**76.0**	**35.5**	80.2	**75.8**
Claude Opus 5	**94.2**	29.6	74.4	27.0	79.3	70.6
Gemini Pro 3.1	88.9	28.9	65.3	49.3	**84.3**	69.9
Qwen3.8-27B	84.4	25.4	71.8	25.7	82.0	62.6
Gemma-3-12B	78.0	15.1	70.3	13.3	68.0	57.5
Gemma-3-27B	81.3	19.5	70.3	22.1	81.8	59.3
Pixtral-12B	76.7	11.9	68.5	10.8	79.4	45.1

translate or preserve (never the gold target strings). Unmatched gold spans count as failures, so partial coverage cannot inflate the score.

Image-level preservation (API subset). Unlike DECISION and ACT, which run on the full open-model set, image-level preservation is evaluated only on a stratified subset of **600** items (200 ads / 200 news / 200 video), since rendering is API-costly. There we call proprietary APIs and an image-to-image renderer (GPT image 2, shared across conditions) to produce translated images, then OCR and align the output to the gold spans to check whether PRESERVE spans remain unchanged in pixels.

Languages. Advertisements and captioned video are evaluated across six languages (English, Spanish, French, Korean, Chinese, Japanese). News coverage follows each source: Korean–English for Chosun, and Korean, English, Japanese, and Chinese for Dong-A. Open models run locally on six H100 80GB GPUs; proprietary models are accessed through vendor APIs on the 600-item subset.

5 RESULTS AND ANALYSIS

We organize this section around one question: when a model is simply asked to translate an image, does it preserve what should be preserved, and if not, why? We first establish the core phenomenon and confirm it holds end to end (§5.1), then isolate its cause (§5.2), and finally show that the gap can be closed by making the decision explicit (§5.3).

5.1 MODELS DECIDE WELL BUT FAIL TO PRESERVE

Table 2 reports KNOW, DECISION, and PRESERVE per domain. Two patterns stand out. First, DECISION accuracy is high: given the image, models label translate versus preserve well above 70% almost everywhere. Second, PRESERVE is much lower. Even when a model has correctly decided that a span should be kept, its output frequently translates it anyway, with drops of twenty

Table 4: Vision ablation on ads Decision (n=600). Scoring with OCR text only (txt) versus with the image (img); Δ is the gain from adding the image. Vision improves every model.

Model	Dec_{txt}	Dec_{img}	ΔDec	Pres_{txt}	Pres_{img}	ΔPres
InternVL3.5-8B	69.5	71.2	+1.7	38.8	53.3	+14.5
Qwen3-VL-8B	66.8	73.0	+6.2	50.3	54.9	+4.6
Gemma-3-12B	65.7	85.5	+19.8	46.2	57.6	+11.4
Pixtral-12B	60.8	78.9	+18.1	42.1	51.3	+9.2
Gemma-3-27B	61.8	86.3	+24.5	53.1	61.4	+8.3
Qwen3.8-27B	77.6	88.8	+11.2	54.0	62.7	+8.7

Table 5: Error decomposition of Decision failures (n=600, 7104 spans). Over-translation counts a gold-preserve span that was translated; under-translation a gold-translate span left unchanged.

Model	#fail	Over%	Under%
InternVL3.5-8B	102	74.5	25.5
Qwen3-VL-8B	86	98.8	1.2
Qwen3.8-27B	108	80.6	13.9
Gemma-3-12B	129	91.5	8.5
Gemma-3-27B	129	83.7	16.3
Pixtral-12B	176	71.0	24.4

Table 6: Guided versus unguided ACT. Guided supplies the translate/preserve inventory (never gold strings); preservation rises sharply, isolating a binding failure.

	Decision		Preserve	
Model	Unguid.	Guided	Unguid.	Guided
InternVL3.5-8B	75.4	93.3	51.6	93.6
Qwen3-VL-8B	79.1	86.6	56.2	83.6
Qwen3.8-27B	83.7	87.1	56.2	92.5
Gemma-3-12B	76.6	92.8	48.8	93.4
Gemma-3-27B	81.7	91.5	52.8	90.9
Pixtral-12B	73.5	84.1	48.9	80.4

to forty points from DECISION to PRESERVE across models and domains. The failure is not in deciding, but in acting on the decision.

This is not specific to open models or to the text interface. On the stratified API subset (§4) we evaluate proprietary systems, and two open models for comparison, end to end, including preservation in the *rendered* image; final images use a shared renderer (GPT Image 2) so differences reflect the decision and translation fed into it, not the renderer. Table 3 shows the same pattern: identity-bearing text is frequently rewritten, in the intermediate text and in the output pixels alike. Over-translation is thus a property of how these systems translate images, not an artifact of any single model or interface.

5.2 WHY MODELS FAIL

The failure is a failure to use vision. If models can decide correctly, why do they not preserve? A first hypothesis is that the decision is shallow, made from text strings alone rather than from the image. Table 4 tests this by scoring DECISION and PRESERVE with the image (img) and with OCR text only (txt). Removing the image hurts every model ($\Delta > 0$ throughout, up to $+24.5$ on Decision and $+14.5$): the task genuinely requires seeing where text sits in the scene. The failure is thus a failure to carry visual grounding into action, not to have it available.

Mistranslation is dominated by over-translation, not omission. Models do not fail by being conservative. Table 5 decomposes DECISION failures into over-translation (a gold-preserve span translated) and under-translation (a gold-translate span left unchanged). Across every model, 71 to 99% of failures are over-translation: the models are not hesitant, they are over-eager, translating too much and erasing exactly the identity-bearing text that should have survived. This is what makes the phenomenon a matter of restraint rather than capability. A striking special case is text that is *already* in the target language: even then, Astra often schedules it for translation, with a redundant "translate again" rate reaching 40–51% on Latin-script directions (Appendix Table 10). Text already in the requested language cannot need translating, so this is over-translation in its purest form, a failure of restraint rather than of understanding.

Scale does not close the gap. A natural expectation is that larger models preserve better; they do not. Placing KNOW next to unguided preserve (Appendix Table 9), KNOW is high everywhere and

Table 7: Effect of the mapping intervention. Base: unguided. +Map: explicit translate/preserve mapping before acting (no gold labels); Δ vs. Base (gain/drop). Oracle: gold guidance (upper bound).

	Decision			Preserve		
Model	Base	+Map (Δ)	Oracle	Base	+Map (Δ)	Oracle
InternVL3.5-8B	70.7	90.3 (+19.6)	95.9	56.5	86.4 (+29.9)	93.9
Qwen3-VL-8B	84.2	92.0 (+7.8)	95.8	61.3	87.9 (+26.6)	93.7
Qwen3.8-27B	77.0	93.9 (+16.8)	97.8	65.3	91.5 (+26.2)	97.3
Gemma-3-12B	76.3	87.1 (+10.8)	95.9	55.9	83.7 (+27.8)	90.7
Gemma-3-27B	66.0	92.1 (+26.1)	96.4	53.9	91.6 (+37.6)	95.0
Pixtral-12B	83.1	91.3 (+8.3)	95.9	52.2	85.5 (+33.3)	92.4

does not track size while unguided preserve stays low, so the knowing-doing gap persists, and in some cases widens, at 27B. Knowing the rule better does not translate into applying it, mirroring findings in embodied settings where added capacity does not produce more restraint.

5.3 Closing the Gap

The analyses above point to a binding failure: the model holds the rule and can read the image but does not connect the two when it translates. Table 6 tests this directly by contrasting the unguided setting with a guided one in which the model is told which spans to translate and preserve. Preservation rises sharply once the inventory is handed over. The inventory, however, is not privileged: it is exactly what the model itself produces under KNOW and DECISION. The model already has it, so the gain isolates the binding step.

If the missing piece is a mapping the model could produce anyway, prompting it to construct that mapping before translating should help, with no training. Table 7 shows this: a lightweight intervention (Bind-then-Act) that makes the model bind each span to its type and visual context before deciding raises preservation over the unguided baseline; we detail the method in Appendix E. We present this not as a finished method but as evidence for the diagnosis and a direction (§6).

Language analysis. Table 11 breaks preservation down by language pair. The decision-to-preserve gap is not uniform: it varies with script and with how close the source and target surfaces are, indicating that the failure is modulated by how visually distinct the preserve targets are from ordinary translatable text.

6 Conclusion

We introduced *selective translation* and SPOT-Bench, which reframes in-image translation as a decision about *what* to translate, spanning advertisements, news, and captioned video. Our central finding is a consistent knowing-doing gap: models can state the translate/preserve rule and decide correctly, yet when asked to translate they systematically over-translate identity-bearing text that should be preserved. The failure is not missing knowledge but a failure to bind it to action, and it does not close with scale; making the decision explicit before translating raises preservation without any training, which we offer as evidence for the diagnosis and a direction rather than a finished method. More broadly, this mistranslation, predominantly over-translation, may be one instance of a general pattern: a model that knows a constraint but does not apply it when acting, which could recur in selective editing, redaction, or embodied settings where an agent must act on some cues and abstain on others; whether it generalizes, and whether one mechanism can close it, is worth pursuing.

AI USE STATEMENT

In this work, generative AI tools were used in three ways, all of which we disclose here. First, we used AI tools for language editing, including grammar correction, sentence refinement, and improving the clarity and readability of the manuscript. Second, generative vision-language models are themselves the *object of study*: we query them to measure whether they hold and follow localization rules, and their outputs constitute our experimental data rather than authored content. Third, in constructing the benchmark we used AI tools to assist simple, mechanical steps of data generation, such as drafting candidate text and flagging text regions, prior to human verification; every item was then manually inspected and corrected by the authors before inclusion.

We did not use generative AI tools to generate research ideas, design the methodology or experiments, analyze experimental results, or produce scientific claims. All AI-assisted language edits were manually reviewed by the authors to ensure they preserved the intended technical meaning, and all AI-assisted data was human-verified as described above. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 4159–4170, 2019.

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.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

European Parliament and Council of the European Union. Regulation (EC) no 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. Official Journal of the European Union, L 342, pp. 59–209, 2009. URL `https://eur-lex.europa.eu/eli/reg/2009/1223/oj`. Article 19 requires ingredient labelling using the common ingredient names glossary based on the International Nomenclature of Cosmetic Ingredients (INCI).

Matthieu Futeral, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. Towards zero-shot multimodal machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 761–778, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

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.

Hour Kaing, Jiannan Mao, Haiyue Song, Chenchen Ding, Hideki Tanaka, and Masao Utiyama. Imagetra: Real-time translation for texts in image and video. In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 1–8, 2025.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5684–5696, 2020.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. Exploring better text image translation with multimodal codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3479–3491, 2023.

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.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning, july 2023. *URL http://arxiv. org/abs/2307.13702*, 2023.

Haijun Li, Tianqi Shi, Zifu Shang, Yuxuan Han, Xueyu Zhao, Hao Wang, Yu Qian, Zhiqiang Qian, Linlong Xu, Minghao Wu, Chenyang Lyu, Longyue Wang, Gongbo Tang, Weihua Luo, Zhao Xu, and Kaifu Zhang. Transbench: Benchmarking machine translation for industrial-scale applications, 2025. URL https://arxiv.org/abs/2505.14244.

Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9806–9815. Ieee, 2020.

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, 2026. URL https://arxiv.org/abs/2603.10495.

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.

Peter Newmark. *A textbook of translation*, volume 66. Prentice hall New York, 1988.

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.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

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.

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.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. *arXiv preprint arXiv:2105.05486*, 2021.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 543–553, 2016.

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, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *International Conference on Learning Representations*, volume 2024, pp. 56783–56799, 2024.

Lawrence Venuti. *The Translator's Invisibility: A History of Translation*. Routledge, London, 2 edition, 2008.

Lawrence Venuti. *The translator's invisibility: A history of translation*. Routledge, 2012.

Xintong Wang, Jingheng Pan, Yixiao Liu, Xiaohu Zhao, Chenyang Lyu, Minghao Wu, Chris Biemann, Longyue Wang, Linlong Xu, Weihua Luo, and Kaifu Zhang. Rethinking multilingual vision-language translation: Dataset, evaluation, and adaptation, 2025. URL https://arxiv.org/abs/2506.11820.

Yihao Wang, Zijian He, Jie Ren, and Keze Wang. Rose: Benchmarking the perception-to-action gap in multimodal models. *arXiv preprint arXiv:2606.19965*, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European conference on computer vision*, pp. 148–166. Springer, 2022.

Zhiyong Wu, Lingpeng Kong, Xiang Li, Ben Kao, et al. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6153–6166, 2021.

Doguhan Yeke, Elif Su Temirel, Ananth Shreekumar, Brandon Lee, Dongyan Xu, and Z Berkay Celik. The yes-man syndrome: Benchmarking abstention in embodied robotic agents. *arXiv preprint arXiv:2605.20544*, 2026.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Vilém Zouhar, Niyati Bafna, Mukund Choudhary, et al. Last translation benchmark, 2026. URL https://arxiv.org/abs/2609.04173.

APPENDIX

A CONSTRUCTIONAL RULE BATTERY

Table 8 lists the nine constructional rule statements used for the KNOW-rules probe, grouped by medium. Each is presented on a blank image as "Is this a correct localization rule? Answer YES or NO." One video rule that requires a source/target caption pair (unreadable from a single frame) is excluded from the battery.

Table 8: The nine constructional rules for the KNOW-rules probe.

ID	Gold	Rule
Advertisements		
A1	preserve	Brand names must stay in their original form.
A2	preserve	Product names printed on the pack must stay in their original form.
A3	translate	Marketing copy (headline, slogan, CTA, claims, hashtags, step labels) must be translated.
A4	preserve	Text printed on the physical pack must stay in its original form.
A5	preserve	INCI/ingredient lists, model codes, net weight, SPF, and units printed on the pack must stay in their original form.
Captioned video		
C1	translate	Text from the timed-text caption burned on the frame is the localization layer and must be translated.
C3	preserve	Text belonging to the video world itself (logos, signs, watermarks, painted lyrics) must stay in its original form.
News		
N1	translate	The headline, standfirst, and body must be translated.
N2	preserve	Text inside the photograph (building signs, diegetic lettering, logo) must stay in its original form.

B RULES AND GROUND-TRUTH FROM HUMAN EVALUATIONS AND ANNOTATIONS

Our ground truth is derived by construction (§3.2): each stimulus separates a localizable layer from an identity layer, and the translate/preserve decision follows from that separation rather than from case-by-case annotation. To confirm that this rule reflects how professional translators actually decide, rather than our own preference, we ran a human validation study.

Setup. We hired three professional translators (denoted A, B, and C). Each independently annotated a shared set of **600** images, labeling every visible text span as TRANSLATE or PRESERVE with no access to our gold labels or to the constructional rule. Annotators were shown the image and the target language only, and were instructed to mark, for each span, whether a competent localizer would translate it or leave it in its original form.

Agreement with the constructional rule. Figure 4 reports, per annotator, the agreement between their manual labels and the constructional gold, together with the mean across annotators. Agreement is high throughout (above 90% for every annotator, 96.0% on average), indicating that the rule is not an arbitrary choice but closely matches independent expert judgment. Figure 4 breaks this down by rule (A1–A5 for advertisements, C1/C3 for captioned video, N1/N2 for news; see Appendix A). Agreement is uniformly high across rules, with the largest disagreements on ingredient- and pack-related preservation (e.g., A2, A4), where a small number of borderline spans (for example, a product line name that reads like a common word) account for most of the divergence. These cases are exactly the difficult, knowledge and context dependent decisions the benchmark is designed to probe.

Adjudication and final labels. For the validation set, we confirmed spans where all three annotators agreed. When there was disagreement, we discussed the cases and checked them against the

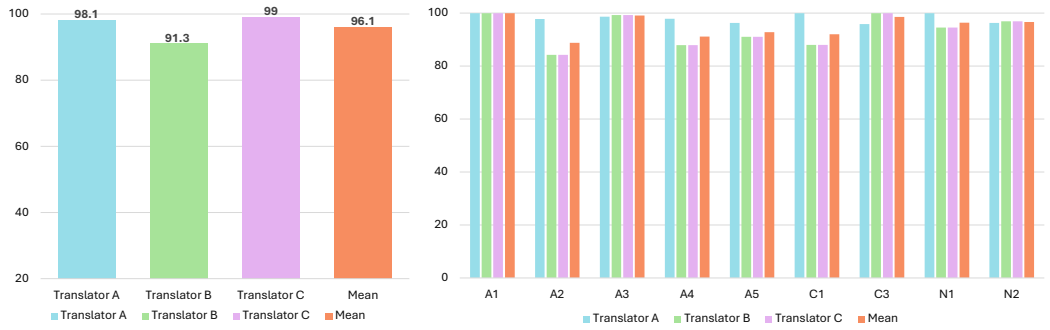


Figure 4: **Human validation of the constructional rule.** Agreement between three professional translators' manual translate/preserve labels and the constructional gold. *Left*: per-annotator agreement (Translator A/B/C) and the mean (96.1%). *Right*: agreement by rule (A1–A5 advertisements, C1/C3 captioned video, N1/N2 news). Agreement is high throughout, with the largest disagreements on borderline preservation cases (e.g., A2, A4), which are the knowledge- and context-dependent decisions the benchmark targets.

constructional rule. After validating the rule, we used it to label the remaining images consistently. This approach ensures the ground truth relies on expert judgment and scales without per-image subjectivity. This human-in-the-loop procedure uses expert labor to *validate and verify* the rule, not to decide each case by taste.

C Implementation details

Prompting. For each task, the same fixed prompt is used for all the items. In the case of the KNOW-rules probe, a blank image is shown together with one rule statement taken from Table 8 and a YES/NO answer is requested. The KNOW-spans probe shows a list of text strings without any image and asks for a TRANSLATE/PRESERVE label to be given for each string. The DECISION prompt displays the image and asks, for each detected text region, for a per-region label to be provided. The ACT prompt shows the image and includes the instruction "translate this image into {target}" and asks, for each visible region, for both its source form and its output form; the decision is determined by the output (if the surface remains unchanged then the answer is PRESERVE, if the surface has changed then the answer is TRANSLATE). We employ simple, instruction-based prompts and never include the decision rules in any of the image-conditioned tasks.

Guided inventory. In the guided ACT condition, the model is given, alongside the image, an inventory listing each span's source text and whether it should be translated or preserved, but never the gold target strings. The inventory is derived from the constructional gold, so the model still has to produce the translation itself; comparing guided and unguided ACT thus separates knowing which spans to preserve from acting on that knowledge.

Compute. Open models run locally on six NVIDIA H100 80GB GPUs. 8 to 12 billion parameter models: Two workers are used for each GPU, whereas with the 27B model there is one worker per GPU. The proprietary DECISION and image-editing models are accessed via the vendor's APIs on the 600-item subset, and with each result the exact model versions and access dates are given.

D Additional analysis

The analysis in this section supports the main conclusions namely that over-translation is a lack of restraint rather than a lack of capability, it is not overcome by increasing scale, and it is affected by the degree to which the preserve targets differ visually and orthographically.

Table 9: Scale and Know≠Act on (n=600). Know = verbal P/T without image; Pres@Dec = gold-P at Decision; Act-P = unguided preserve on gold-P. $\Delta_{\mathrm{K-A}}$ = Know − Act-P (pp).

Model	size (B)	Know	Dec.	Pres@Dec	Act-P	$\Delta_{\mathrm{K-A}}$
InternVL3.5-8B	8	74.6	91.7	91.0	43.6	31.0
Qwen3-VL-8B	8	84.9	93.0	89.0	57.1	27.8
Pixtral-12B	12	79.1	85.7	84.3	33.8	45.3
Gemma-3-12B	12	78.2	89.5	86.1	54.0	24.2
Qwen3.8-27B	27	88.2	91.2	90.0	41.0	47.2
Gemma-3-27B	27	77.6	89.5	87.3	58.0	19.6

D.1 Scale does not close the knowing-doing gap

The values shown in Table 9 list each model's verbal rule knowledge (KNOW) alongside its unguided performance on the gold-preserve spans (Act-P), and give the difference $\Delta_{\mathrm{K-A}}$. Two points are clear. On the one hand, both KNOW and DECISION are consistently high (ranging from 74 to 93%) and remain almost unchanged as the model size increases: regardless of their size, all the models are able to state the rule and, when given the spans, make the correct decision. On the other hand, Act-P is much lower and does not improve with larger model sizes; in fact, the gap widens, amounting to 47.2 points for Qwen3.8-27B and 45.3 points for Pixtral-12B, while the models Gemma-3-27B and InternVL3.5-8B have lower values. The discrepancy between knowing and doing is thus not due to a capacity limitation that can be overcome by adding more parameters: simply having a better understanding of the rule does not result in the rule being applied when the model is actually processing the image. This is similar to what has been observed in embodied settings, in which increasing capacity or reasoning on its own does not lead to greater restraint, and this finding supports the idea of introducing interventions that make the decision explicit rather than just scaling the backbone.

D.2 Redundant translation of already-in-target text

A particularly sharp case of over-translation occurs when the text is *already* written in the requested target language and should therefore remain unaltered. The results shown in Table 10 gives, for each language direction, the number of times the model still selects TRANSLATE for such spans (the detection being carried out using a script heuristic for cases where the target language is the same script, and the spans being treated as edition-identical in the cross-script cases). The result is striking in that for the Latin↔Latin directions (en/es/fr) the proportion of redundant "translate again" predictions reaches 40–51% (reaching a peak of 51.4% for en→fr), whereas for directions into Korean or Chinese this figure is much lower (13–30%). Since text which is already in the target language has no need for translation, these rewrites represent the purest form of over-translation. The fact that script plays a role here is significant: when the source and target languages use the same script, the model cannot make use of a superficial clue to distinguish between "already done" and "to be translated" and therefore translates everything; but when the scripts are different, the visual mismatch itself serves as a signal that suppresses the redundant action. This provides clear evidence that the problem lies in a lack of restraint and in the failure to read the context, not in any lack of knowledge.

D.3 Preservation collapses on identity-bearing spans

The core failure, as shown in Table 11, is the frequency with which the model translates a gold-PRESERVE span, that is, it converts an identity-bearing element (such as a brand, ingredient code, or proper noun) into a translated form. The rates are high and are roughly the same in all directions (11–31%), and, in contrast to the case where the element is already in the target language, they do not disappear when the target language involves a different script: in cases such as en→ko or zh→fr translation errors for identity spans still occur at 20–30%. This points to two different mechanisms. The errors in the already-in-target case (§D.2) are reduced because of a visual script cue; the errors involving gold-preserve here are not, since deciding that a brand or ingredient should be kept requires knowledge and scene context, not just a simple script check. The fact that preservation fails specifically when the cue is semantic rather than orthographic is in line with our main claim: the

Table 10: Already-in-t→TRANSLATE (Astra, ads Decision). Script heuristic for same-script targets; edition-identical for cross-script blanks. Higher = more redundant "translate again".

Direction	Criterion	Already-t→T (%)
en→es	script	48.3
en→fr	script	**51.4**
en→ko	edition	37.5
en→zh	edition	25.9
es→en	script	46.0
es→fr	script	45.7
es→ko	edition	27.6
es→zh	edition	20.7
fr→en	script	45.1
fr→es	script	40.1
fr→ko	edition	30.6
fr→zh	edition	23.2
ko→en	script	28.3
ko→es	script	25.3
ko→fr	script	13.8
ko→zh	edition	19.8
zh→en	script	17.4
zh→es	script	19.3
zh→fr	script	27.2
zh→ko	edition	24.8

Table 11: Gold-PRESERVE→TRANSLATE (Astra, ads Decision). Same directions as left. Higher = more must-keep / identity spans flipped to translate.

Direction	Criterion	Gold-P→T (%)
en→es	gold-P	28.2
en→fr	gold-P	**28.9**
en→ko	gold-P	26.0
en→zh	gold-P	23.5
es→en	gold-P	21.3
es→fr	gold-P	24.3
es→ko	gold-P	27.9
es→zh	gold-P	17.6
fr→en	gold-P	27.4
fr→es	gold-P	18.7
fr→ko	gold-P	19.8
fr→zh	gold-P	21.7
ko→en	gold-P	**30.3**
ko→es	gold-P	23.3
ko→fr	gold-P	14.0
ko→zh	gold-P	11.4
zh→en	gold-P	21.0
zh→es	gold-P	22.0
zh→fr	gold-P	**30.9**
zh→ko	gold-P	20.5

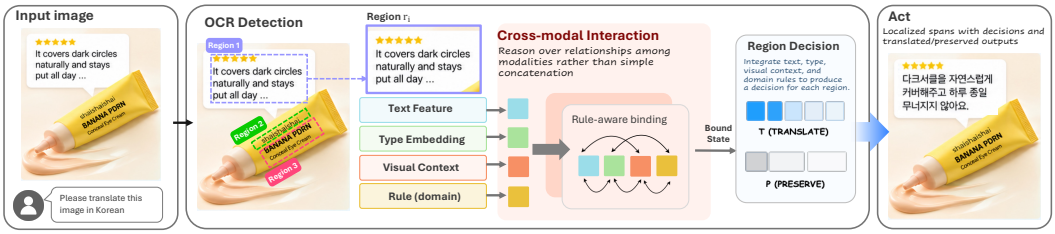


Figure 5: Bind-then-Act: each OCR span is bound to type, visual context, and a domain rule before Translate / Preserve; Act then produces the localized image (no gold T/P inventory).

models know the rule but fail to apply it to the image, and the failure is most severe precisely in those cases that involve knowledge and context and which are exactly the ones the benchmark is designed to test.

E BIND-THEN-ACT FOR FUTURE WORK

The analysis shows that the problem does not lie in a lack of knowledge but in the fact that models hold the translate/preserve rule and can read the image, yet fail to link the two when carrying out the translation. This suggests a straightforward, training-free approach. Instead of mapping an image directly onto the translated output, we make the model *bind* each visible span to its role and context before deciding, so that the decision results from that binding rather than from a reflex to translate everything. As shown in Figure 5, the pipeline binds each OCR span to a type, a visual context, and a domain rule before the TRANSLATE/PRESERVE decision, after which the model acts to generate the localized image, with no gold translate/preserve inventory.

Method. The input is still a single image, with no gold translate/preserve inventory and no target strings; the only change is the required output structure. Alongside the image we supply the constructional rules themselves (e.g., brands and ingredient codes are preserved, marketing copy and headlines are translated; the full rule set is in Appendix A), but never which specific spans they apply to. For each detected span, the model must produce, in a fixed order, its `type` (semantic role), its `visual_context` (where it sits in the image), the `rule` that applies, and only then its

`decision` and `output`; the prompt explicitly instructs the model not to jump to the final string (Figure 7). Because the model emits the decision only after the type, context, and rule, it cannot default to translating everything, and the decision becomes a deterministic function of type, visual context, and rule. At the generation step, the model additionally receives each span's inferred type and visual context, so that the final image is rendered conditioned on this structure rather than on the raw text alone.

Three settings. We situate Bind-then-Act between the deployable unguided baseline and the non-deployable oracle ceiling. All three take a single image and never receive the gold target strings; they differ only in what structure or information is supplied. Crucially, Bind-then-Act is not oracle guidance: we never reveal which spans to translate or preserve, so any gain over the unguided baseline reflects decision *structure* alone, not privileged information.

Effect and outlook. By first identifying what each span is and where it occurs before carrying out the translation, the model avoids automatically falling back on a global translation, thereby reducing over-translation and increasing preservation compared to the unguided baseline (Table 7), so that the gap between the current performance and the guided upper limit is narrowed without the need for any training. We regard this as a potential research direction rather than as a complete method: it is still an open question how to close the knowing-doing gap more robustly, whether by applying explicit binding at inference time or through training that rewards restraint, and whether the same approach can be generalised beyond the case of translation.



Figure 6: **Mistranslation example.** Top: input advertisement (French). Bottom: outputs localizing into Korean, over-translated spans in red. Several models translate the brand *Mardi Mercredi* instead of preserving it, and even rewrite the on-garment logo, erasing brand identity.

Localize this {ads|news} image into {tgt_lang}.
You MUST follow **Bind-then-Act**. Do not jump directly to the final string.

RULES

Ads (A1–A5)

[A1] brand → PRESERVE
[A2] product name on pack → PRESERVE
[A3] marketing copy (headline/slogan/CTA/claims) → TRANSLATE
[A4] pack print → PRESERVE
[A5] INCI / SPF / net weight / model code → PRESERVE

News (N1–N2, NP/NO)

[N1] headline / standfirst / body → TRANSLATE
[N2] text inside the photograph (diegetic signs) → PRESERVE
[NP/NO] person / org names → official target-language form

BIND VOCABULARY

TYPE **Ads:** brand, product, copy, pack_print, ingredient, other
News: headline/body/standfirst, masthead, photo_sign, person/org, other

VISUAL CONTEXT **Ads:** on_pack_surface, ad_overlay, logo_mark, uncertain
News: cms_text_layer, masthead_chrome, on_photograph, caption_overlay, uncertain

BIND → ACT

For each visible text span, output one JSON object with fields:

`text → type → visual_context → rule → decision → output`

The decision must follow the bound type, visual context, and rule:

Ads brand/product/pack_print/ingredient + on_pack_surface → PRESERVE
copy + ad_overlay → TRANSLATE

News photo_sign + on_photograph → PRESERVE
headline/body/masthead + cms_text_layer → TRANSLATE

OUTPUT

Return a JSON list only. No markdown fences.

Figure 7: Bind-then-Act prompt (Method 1), unified across ads and news. Domain-specific type and visual-context vocabularies, together with constructional rules, are selected at runtime.