OmniDisambiguate: A Contrastive Study of Vision as a Double-Edged Sword in Omni Translation Models

Anonymous ACL submission

Abstract

Omni language models are widely used for translation, where vision is assumed to help resolve ambiguity, yet this assumption remains under-tested. We study lexical disambiguation as a controlled setting and introduce OMNIDISAMBIGUATE, a multilingual benchmark with 7,762 ambiguous instances across eight target languages, each paired with aligned video and audio. We further propose Contrastive Visual Probing (CVP) and a set of metrics to test whether models benefit from correct visual cues and resist misleading ones. Experiments on 10 omni and vision-language models show a consistent double-edged effect of vision. Adding video often reduces accuracy, indicating over-reliance on weak visual signals. The effect varies by model and remains consistent across languages and input modalities. Errors increase when misleading videos are semantically close to the correct meaning, pointing to non-robust use of visual semantics. These findings suggest that evaluation should move beyond average gains and examine how visual evidence is used, for which OMNIDISAMBIGUATE and CVP provide a reusable benchmark and protocol [1].

1 Introduction

Omni language models are increasingly used for general-purpose translation (Team, 2026b; Chen et al., 2025; Qwen Team, 2025; Xu et al., 2025b; Team, 2026a; Xu et al., 2025a; Li et al., 2024b). Vision is often treated as additional semantic information beyond the source sentence (Hubscher-Davidson, 2020; Bang, 1986; Ma et al., 2025; Caglayan et al., 2019; Team, 2026b). As shown in Figure 1(a), the relevant video appears to supplement the missing contextual information for the translation task. Under this assumption, adding vision should make translation more grounded, more context-aware, and more reliable.

However, lexical disambiguation reveals a less examined risk. Visual contextualization is inherently informative, yet it can sometimes introduce cross-modal competition with the source text (Xie et al., 2024; Leng et al., 2024; Liu et al., 2025). This competition often leads to unpredictable model behaviors, where a model might use the video correctly, ignore it, or over-rely on it even when the textual context is already sufficient (He et al., 2025; Zou et al., 2024; Jing et al., 2024). These behaviors can yield similar performance scores, yet they imply very different uses of visual information (Fu et al., 2023; Lee et al., 2025). As a result, benchmarks that only compare performance with and without vision cannot answer the key question: *When visual information changes a translation decision, is the model using valid semantic evidence, or is it misled by a salient but unreliable cue?*

To address this, we introduce **OMNIDISAMBIGUATE**, a multilingual benchmark for evaluating omni translation. It contains 7,762 ambiguous instances, covering eight typologically diverse target languages, each paired with aligned video and audio. The data includes two common ambiguity types: homophones, where words share pronunciation but differ in meaning, and polysemous words, where a single form maps to multiple related senses. We conduct human verification to ensure cross-lingual sense alignment and disambiguation validity, enabling fine-grained analysis of visual usage.

For evaluation, we propose **C**ontrastive **V**isual **P**robing (**CVP**). For both text-source and speech-source inputs, we construct four conditions: no video, correct video, contrastive video, and irrelevant video. An ideal model should benefit from correct visual input and semantically related but misleading videos. Based on this setup, we define a set of metrics, including Visual Grounding Rate (VGR), Visual Susceptibility Rate (VSR), and Vi-

[1] Our dataset and code will be publicly available upon acceptance.

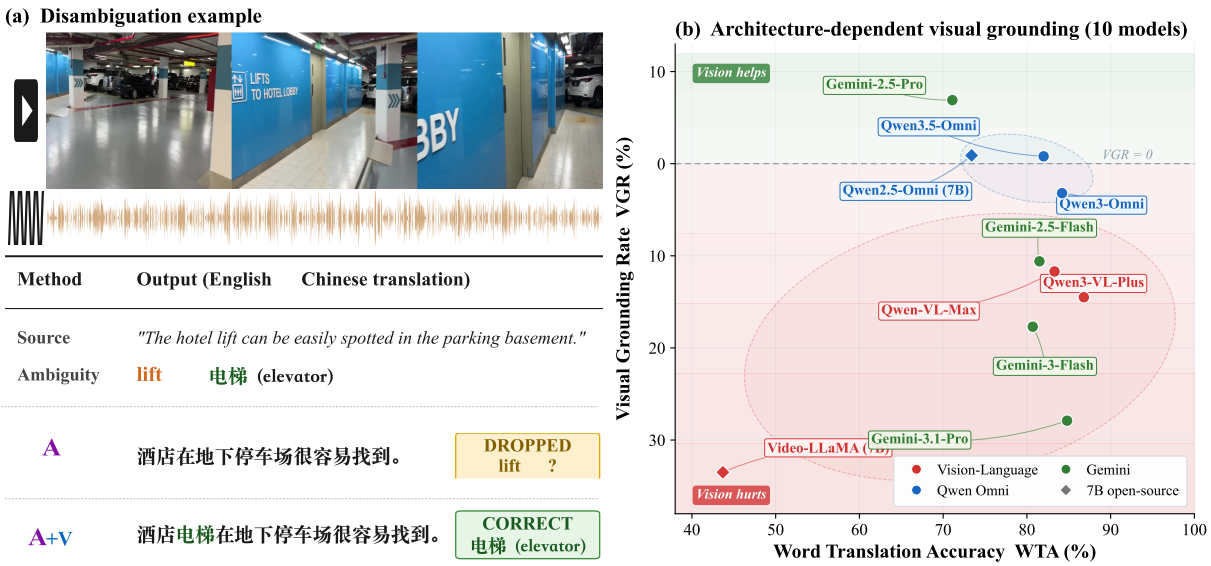

Figure 1: **Vision is a double-edged sword for translation disambiguation.** **(a)** Video resolves the ambiguous word *lift* to "elevator"; audio alone drops it entirely. **(b)** VGR varies sharply by architecture: VL models suffer visual gullibility, Qwen Omni models stay near zero, and Gemini models diverge within the same paradigm.

sual Discrimination Index (VDI). These metrics characterize when visual signals help or hurt, and whether the model can use visual information in a discriminative way.

Under this framework, we evaluate 10 omni and vision-language models. The results reveal a clear double-edged effect of vision in translation disambiguation: adding video often reduces translation accuracy, suggesting that models tend to over-rely on unreliable visual cues. As shown in Figure 1(b), Model behavior varies substantially across architectures and model families. Vision-language models are more susceptible to misleading videos, Qwen Omni models are relatively stable, and Gemini models exhibit large within-family variance. Further analysis shows that this pattern is consistent across all eight languages. Contrastive video experiments further show that misleading videos semantically closer to the correct sense are more likely to induce errors, suggesting that the issue stems from non-robust use of visual semantics rather than simple noise. Our findings suggest that multimodal translation evaluation should move beyond average performance changes and instead focus on how models use visual information. The contributions are as follows:

1. We construct OMNIDISAMBIGUATE, a multilingual benchmark for translation disambiguation with 7,762 ambiguous instances, eight target languages, and aligned video and audio, and propose the Contrastive Visual Probing protocol with diagnostic metrics to evaluate how models use visual information.

2. We conduct a systematic evaluation of 10 omni and vision-language models. Visual effects vary strongly across models. Some remain stable, while others are clearly misled. This pattern is consistent across languages and input modalities.

3. We analyze the role of different types of visual signals, temporal cues, semantic distance in contrastive videos, speech input, and Chain-of-Thought prompting. The results show that errors mainly arise from implicit cross-modal fusion, and we outline simple directions for mitigation.

2 The OMNIDISAMBIGUATE Benchmark

Existing multimodal translation benchmarks include image-based datasets such as Multi30K (Elliott et al., 2016) and its extensions (Specia et al., 2016; Barrault et al., 2018), and video-based datasets such as VATEX (Wang et al., 2019) and How2 (Sanabria et al., 2018). However, they mainly evaluate aggregate quality, with limited languages, sparse lexical ambiguity, and no contrastive design for testing visual disambiguation. To address this gap, we construct OMNIDISAMBIGUATE, which covers homophonous and polysemous ambiguity, spans eight diverse languages, and pairs instances with aligned videos for contrastive analysis.

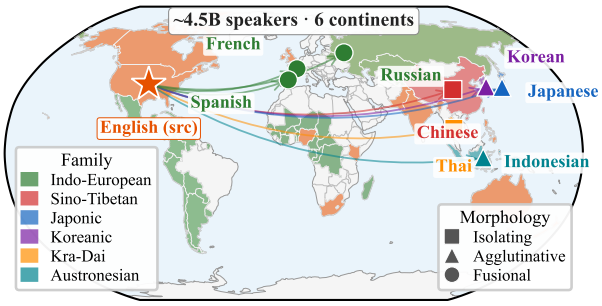

Figure 2: **Geographic coverage of the 9-language selection.** Highlighted countries are where each language is official or majority-spoken; arrows show translation directions from English. The selection covers ∼4.5B speakers across 6 continents.

Lang.	Family	Script	Poly	Homo
Chinese	Sino-Tib.	Hanzi	499	500
Korean	Koreanic	Hangul	479	500
Japanese	Japonic	Kanji	478	500
Thai	Kra-Dai	Thai	479	500
French	Indo-Eur.	Latin	456	500
Spanish	Indo-Eur.	Latin	446	500
Russian	Indo-Eur.	Cyrillic	453	500
Indonesian	Austrones.	Latin	472	500
Total	**6 fam.**	**6 scr.**	**3,762**	**4,000**

Table 1: **Per-language instance counts.** Polysemy varies by viability; homophones uniform.

2.1 Ambiguity Taxonomy

We target two categories of lexical ambiguity in English-to-target-language translation:

Homophones (500 instances evaluated in each of 8 languages, for 4,000 items in total): Words that sound identical but have different meanings and often different spellings (e.g., *bear*/*bare*, *sea*/*see*). Since English homophones are generally *not* homophones in target languages, translation implicitly resolves the ambiguity.

Polysemous words (446–499 per language, 3,762 total): Words with multiple related senses (e.g., *bank*: financial institution or river bank). Different senses typically map to *distinct* target-language translations, making disambiguation directly observable in the output. We include only samples where disambiguation is viable in each target language.

2.2 Data Construction and Quality Control

Lexicon curation. We construct the ambiguity lexicon in two stages. First, we extract homophone groups from the CMU Pronouncing Dictionary by clustering words with identical phonemic transcriptions, yielding 278 groups (e.g., *bear*/*bare*, *sea*/*see*). Second, we harvest polysemous headwords from WordNet by selecting lemmas with $\geq$2 synsets whose glosses are semantically distinct, producing 613 entries (e.g., *bank*: financial institution or river bank). After deduplication and manual filtering of trivially ambiguous or rare entries, the merged lexicon contains 891 headwords.

Instance mining. Building video-aligned disambiguation instances at scale requires multi-source collection with aggressive quality filtering. We query the YouTube Data API with ambiguous headwords and filter results to CC-BY videos with English closed captions. This produces an initial pool of 287K candidate caption segments. We then apply a multi-stage filtering pipeline: caption–audio alignment verification via forced alignment, target-word presence and sense-discriminability checking with an LLM judge (Appendix I), video–sense semantic relevance scoring to ensure the visual content depicts the intended sense, and deduplication with length normalization. The pipeline retains a 6.5% acceptance rate (Appendix Table 19), ensuring that each surviving instance contains a video segment whose visual content unambiguously depicts the target word sense. We release only short processed clips (5–15s), along with original video IDs and timestamp ranges to support provenance verification, following common academic practices and respecting platform usage policies.

2.3 Multilingual Extension

A key design goal is cross-lingual generalizability. We extend the English source sentences to **8 typologically diverse target languages** spanning 6 language families, 6 writing systems, and 3 morphological types (isolating, agglutinative, fusional), as illustrated in Figure 2 and Table 1. This selection maximizes typological contrast: the languages differ not only in script and genealogy but also in word-order (SVO and SOV) and morphological complexity, ensuring that our evaluation probes genuinely diverse translation challenges rather than superficial surface variation. For polysemous words, we verify that different English senses map to *distinct* words in the target language via LLM-based viability checking. Viability rates range from 89% (Spanish) to 96% (Korean/Thai), reflecting genuine typological variation. The full benchmark comprises **7,762** instances.

Quality control. We perform systematic quality auditing in two stages. First, an automated audit on 200 random samples (100 per category) checks text–word alignment, mapping correctness, video accessibility, and sense–text alignment, yielding an initial homophone clean rate of 86% and polysemy clean rate of 76%.

Second, we conduct a comprehensive human quality control campaign across all **8 target languages** (Table 17; full annotation details in Appendix H). For each language, language experts evaluate all LLM-generated translations, assigning one of three verdicts: *correct* (no changes needed), *partial* (minor corrections applied), or *wrong* (significant errors requiring re-translation). Across 7,762 annotated instances, **84.0%** are rated correct, 9.5% partial, and 6.6% wrong. Per-category clean rates—homophone 87.2%, polysemy 80.0%—are consistent with the initial automated audit, and all partial/wrong samples receive expert-corrected translations. The primary error sources are borderline sense alignment cases rather than structural errors.

3 Contrastive Visual Probing (CVP) Protocol

Standard multimodal evaluation compares text-only against text-with-vision and attributes the accuracy change to visual contribution. To isolate the causal role of visual semantics, we design a controlled protocol that manipulates the *alignment* between visual content and the correct word sense, rather than simply toggling visual input on and off. The key insight is that a model with true visual grounding should improve with correct video, resist misleading video, and be unaffected by irrelevant video.

3.1 Source-Modality-Aware Condition Design

Our protocol crosses two orthogonal factors—**source modality** (text **T** or speech **A**) and **visual context** (none, correct **V**, contrastive **V′**, irrelevant **V**?)—yielding 8 conditions:

- **Baseline** (**T / A**): translate from text or speech alone, establishing unimodal accuracy.
- **+ Correct video** (**T+V / A+V**): the original video depicting the intended sense is provided. A model with genuine visual grounding should improve.
- **+ Contrastive video** (**T+V′ / A+V′**): a video depicting a *different* sense (Intra-word sense swap) or an unrelated word's video (cross-word semantic near-miss). This tests whether the model is misled by plausible but incorrect visual input.
- **+ Irrelevant video** (**T+V? / A+V?**): a randomly selected, semantically unrelated video. A robust model should be unaffected.

Speech conditions apply only to Omni models that support native audio input.

3.2 Metrics

A single accuracy number cannot distinguish whether a model genuinely understands visual content or merely benefits from distributional artifacts. We therefore define a suite of metrics that jointly characterize *how* and *when* vision contributes. Let $C_\theta^{(i)} \in \{0, 1\}$ indicate whether condition θ yields the correct translation for sample i, and N denote the total number of samples.

(1) Word Translation Accuracy (WTA) measures the proportion of correctly disambiguated translations under any condition θ: $\mathrm{WTA}(\theta) = \frac{1}{N}\sum_{i=1}^{N} C_\theta^{(i)}$. This serves as the base measurement for all subsequent metrics.

(2) Visual Grounding Rate (VGR) quantifies the *net benefit* of adding the correct video to text input:

$$\mathrm{VGR} = \mathrm{WTA}(\text{T+V}) - \mathrm{WTA}(\text{T}) \qquad (1)$$

A positive VGR means the model successfully leverages visual semantics for disambiguation. Crucially, a *negative* VGR reveals that vision actively *hurts*—a phenomenon we term *visual gullibility*, where the model over-relies on visual cues even when they conflict with correct textual interpretation.

(3) Visual Susceptibility Rate (VSR) isolates vulnerability to misleading vision. Among samples that the model answers correctly with the correct video (T+V), VSR measures how many flip to incorrect when the video is swapped to a contrastive one depicting a different sense:

$$\mathrm{VSR} = \frac{|\{i : C^{(i)}_{\text{T+V}}{=}1 \ \wedge \ C^{(i)}_{\text{T+V}'}{=}0\}|}{|\{i : C^{(i)}_{\text{T+V}}{=}1\}|} \qquad (2)$$

High VSR indicates that the model does process visual content but lacks robustness: it can be deceived by semantically plausible but incorrect video into selecting the wrong word sense.

(4) Visual Discrimination Index (VDI) balances benefit against vulnerability into a single score: $\mathrm{VDI} = \mathrm{VGR} \,/\, (|\mathrm{VGR}| + \mathrm{VSR}) \in [-1, 1]$. A

model that benefits greatly from correct video while rarely being misled by contrastive video achieves VDI $\rightarrow +1$; one that is frequently deceived without proportional benefit yields VDI $\rightarrow -1$.

Speech-source extensions. For Omni models with native audio input, we define analogous metrics by replacing T with A: *Speech VGR* = WTA(A+V) − WTA(A) measures whether video aids speech-based translation, and *Modality Gap* = WTA(T) − WTA(A) quantifies the inherent difficulty increase from replacing text with speech.

3.3 Contrastive Video Construction

We establish a hierarchical evaluation by employing three contrastive video (V') construction strategies, strictly balanced in sample size ($N = 100$ pairs per tier) and ordered by descending semantic proximity: (1) **Intra-word sense swap:** the video depicts an *alternative, mutually exclusive sense* of the exact same ambiguous word, testing fine-grained contextual disambiguation; (2) **Cross-word semantic near-miss:** the video is drawn from a different word but shares the same superordinate semantic category as the target, evaluating mid-grained domain-level discrimination; and (3) **Open-domain random baseline:** a video is randomly sampled from the open domain to serve as a coarse-grained alignment check.

4 Experimental Setup

4.1 Models and Setup

Models. We evaluate **ten models** spanning two architectural paradigms:

(1) **Omni models** (unified text, audio, and video): Qwen3-Omni (Xu et al., 2025b), Qwen3.5-Omni (Team, 2026a) (closed-source [2]), Qwen2.5-Omni-7B [3] (Xu et al., 2025a) (open-source), Gemini-2.5-Pro, Gemini-2.5-Flash (Comanici et al., 2025), Gemini-3.1-Pro, Gemini-3-Flash (Google DeepMind, 2026) (closed-source [4]).

(2) **Vision-Language (VL) models** (text and video frames): Qwen-VL-Max (Wang et al., 2024), Qwen3-VL-Plus (Bai et al., 2025) (closed-source [5]), and Video-LLaMA-7B (Zhang et al., 2023) (open-source [6]).

Each model's text-only condition already serves as its own unimodal baseline, making a separate text-only model unnecessary for our contrastive design. This selection enables controlled comparisons: intra-family (Qwen3 and Qwen3.5), cross-family (Qwen Omni and Gemini), cross-scale (closed-source and 7B open-source), and cross-architecture (Omni and VL).

Implementation. Closed-source models use temperature 0.1 with deterministic decoding; video is transmitted as base64-encoded MP4 clips. Open-source models run on A100 GPUs with greedy decoding and 16 uniformly sampled frames. For speech-source conditions, all Omni models receive WAV audio (16 kHz, mono); VL models are excluded by design.

Evaluation. We use accuracy and LLM-as-Judge as the primary evaluation method, supplemented by string-matching and human cross-validation (Appendix E).

Statistical Methodology We report 95% bootstrap confidence intervals (1,000 resamples) for key metrics. VGR differences are tested via paired McNemar's test on per-instance correctness changes. All models use temperature 0.1 and deterministic decoding for reproducibility.

5 Experiments

5.1 Architecture-Dependent Visual Grounding

We first evaluate all ten models on the En→Zh subset under the text CVP protocol to establish the core finding regarding architecture-dependent visual grounding.

Table 2 shows that VL models suffer large accuracy drops (Qwen3-VL-Plus: -14.5%, Qwen-VL-Max: -11.7%, both $p < 0.001$), while their Omni counterparts maintain near-zero VGR (Qwen3.5-Omni: $+0.8\%$, $p = 0.61$), isolating architecture as the primary associated factor. The intra-family improvement from Qwen3-Omni (-3.2%) to Qwen3.5-Omni ($+0.8\%$) further indicates that gullibility can be mitigated through model development. Gemini models reveal that architecture is necessary but not sufficient: Gemini-2.5-Pro gains 6.9% ($p < 0.001$) while Gemini-3.1-Pro drops 27.9%, a >34% divergence within the same paradigm.

[2] https://help.aliyun.com/zh/model-studio/qwen-omni

[3] https://modelscope.cn/models/Qwen/Qwen2.5-Omni-7B

[4] https://aistudio.google.com/

[5] https://help.aliyun.com/zh/model-studio/vision

[6] https://modelscope.cn/models/DAMO-NLP-SG/VideoLLaMA3-7B

Model	Arch.	WTA(T) ↑	WTA(TV) ↑	VGR ↑	Poly VGR ↑	Homo VGR ↑	p
Omni architecture – Qwen (closed-source)							
Qwen3-Omni	Omni	84.2%	81.0%	−3.2	−3.5	−2.5	0.006
Qwen3.5-Omni	Omni	82.0%	82.8%	+0.8	+0.5	+1.1	0.614
Vision-Language architecture (closed-source)							
Qwen-VL-Max	VL	83.3%	71.6%	−11.7	−16.8	−6.6	<0.001
Qwen3-VL-Plus	VL	86.8%	72.3%	−14.5	−23.0	−6.0	<0.001
Omni architecture – Gemini (closed-source)							
Gemini-2.5-Pro	Omni	71.1%	77.9%	+6.9	+10.9	+3.1	<0.001
Gemini-2.5-Flash	Omni	81.5%	70.9%	−10.6	−10.0	−11.3	<0.001
Gemini-3.1-Pro	Omni	84.8%	56.9%	−27.9	−21.2	−34.6	<0.001
Gemini-3-Flash	Omni	80.7%	62.9%	−17.7	−16.5	−19.0	<0.001
Open-source 7B models (local)							
Qwen2.5-Omni-7B	Omni	76.8%	74.3%	−2.5	−2.1	−2.8	0.617
Video-LLaMA-7B	VL	43.4%	10.2%	−33.2	−37.1	−30.0	<0.001

Table 2: **Ten-model VGR comparison** (En→Zh). McNemar p-values test significance of VGR.

Model	Arch.	WTA(A) ↑	WTA(A+V) ↑	Sp.VGR ↑	p	WTA(A+V′) ↑	WTA(A+V?) ↑	Mod.Gap
Omni architecture – Qwen								
Qwen3-Omni	Omni	71.2	71.6	+0.4	0.795	72.1	70.4	13.0
Qwen3.5-Omni	Omni	71.8	71.8	+0.0	1.000	73.5	72.5	10.2
Qwen2.5-Omni-7B	Omni	59.5	56.9	−2.6	1.000	36.3	47.1	17.3
Omni architecture – Gemini (audio-capable)								
Gemini-2.5-Pro	Omni	67.8	51.8	−16.0	<0.001	20.3	19.0	3.3
Gemini-2.5-Flash	Omni	65.0	52.0	−13.0	<0.001	26.2	25.3	16.5
Gemini-3.1-Pro	Omni	72.2	52.0	−20.2	<0.001	19.3	20.3	12.6
Gemini-3-Flash	Omni	70.6	52.1	−18.5	<0.001	22.4	25.9	10.1

Table 3: **Speech-source CVP results** (En→Zh, Omni models only).

To rule out closed-source deployment artifacts, we evaluate Qwen2.5-Omni-7B (Xu et al., 2025a) and Video-LLaMA-7B (Zhang et al., 2023) locally on identical hardware. Despite both being 7B-parameter models, Qwen2.5-Omni-7B maintains near-zero VGR (−2.5%, not significant) while Video-LLaMA-7B collapses by 33.2% ($p < 0.001$)—the largest drop observed in our study. This 30.7% gap at matched scale provides strong evidence that the disparity is architecture-driven rather than capacity-driven.

5.2 Speech Translation and Cross-Modal Analysis

To examine whether visual robustness generalizes across input modalities, we replace text input with audio and apply the full four-condition protocol to all Omni models. Table 3 reports the results.

Qwen Omni models remain robust. The two proprietary Qwen Omni models exhibit near-zero Speech VGR (Qwen3.5-Omni: +0.0%, $p = 1.000$; Qwen3-Omni: +0.4%, $p = 0.795$), and neither contrastive nor irrelevant video conditions alter accuracy relative to audio-only input. This source-modality-invariant behavior confirms that Qwen Omni's immunity to visual interference extends beyond text input.

Gemini models remain vulnerable. All four Gemini models suffer significant Speech VGR drops ($p < 0.001$), with Gemini-3.1-Pro showing the largest decline (−20.2%), closely paralleling its −27.9% text-source VGR. Under contrastive video, its accuracy collapses from 72.2% (audio-only) to 19.3%, demonstrating that visual gullibility is a modality-general property of the model's visual grounding mechanism rather than an artifact of text-specific interactions.

5.3 Multilingual Results

A natural question is whether the visual gullibility phenomenon is specific to English–Chinese or generalizes across typologically diverse languages. We evaluate Qwen3-Omni and four Gemini models across all 8 target languages. Figure 3 shows that most models achieve 72%–86% average text-only WTA across the 8 languages. Homophone accuracy is generally higher than polysemy (average gap +2–10%), consistent with the intuition that homophones present less inherent lexical ambiguity than polysemous words.

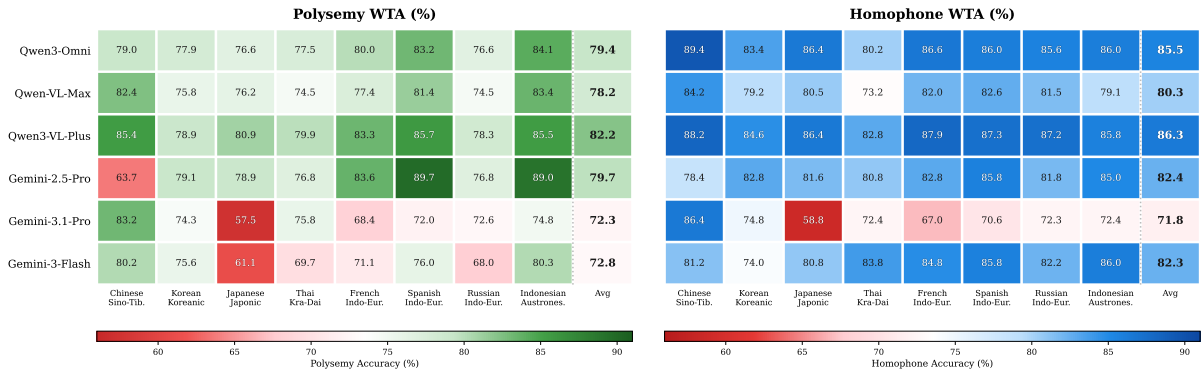

Figure 3: Multilingual text-only WTA on polysemy and homophone splits.

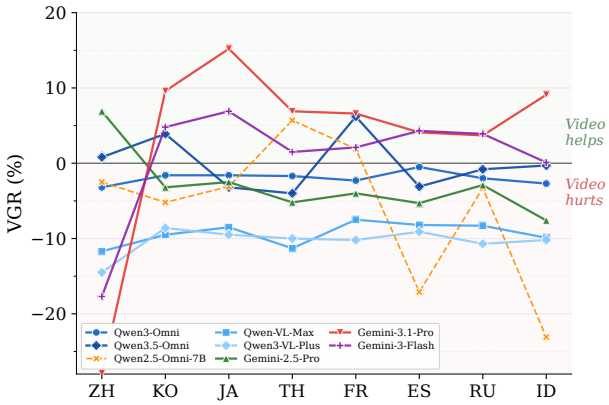

Figure 4: Cross-lingual VGR for 5 models and 8 languages.

As shown in Figure 4, models cluster into distinct VGR profiles. The two proprietary Qwen Omni models (Qwen3-Omni, Qwen3.5-Omni) maintain near-zero VGR across all 8 languages, while Qwen-VL-Max and Qwen3-VL-Plus consistently suffer −8 to −14% VGR, and Qwen2.5-Omni-7B shows high instability with sporadic large drops. Among Gemini models, Gemini-2.5-Pro is positive only on Chinese (+6.9%) but negative for all other targets; Gemini-3.1-Pro displays the opposite pattern—strongly positive VGR on most languages (up to +15%) yet negative on Chinese; Gemini-3-Flash remains broadly negative. This cross-lingual divergence across isolating, agglutinative, and fusional typologies confirms that the VGR profile is a model-level property rather than a language-specific artifact.

5.4 Contrastive Visual Probing Results

The preceding sections establish that correct video can hurt. We now examine *why* by comparing model behavior under contrastive and irrelevant video conditions.

Table 4 reports VSR and VDI by category and across speech-source models. In the text-source condition, polysemy shows higher VSR than homophones (56.0% vs. 24.8%) yet less negative VDI (−0.06 vs. −0.10), indicating that correct video partially offsets polysemy's greater vulnerability. Under speech input, the divide sharpens: Gemini-3.1-Pro's VSR surges to 75.8% with VDI of −0.21, confirming visual input as a net liability, while Qwen Omni models maintain VSR ≤5% and near-zero VDI. Notably, no model achieves a reliably positive VDI in any condition, indicating that robust visual discrimination remains an open challenge.

Condition	N	WTA(T/A) ↑	WTA(+V′) ↑	VSR ↓	VDI ↑
Text-source VSR/VDI (Qwen3-Omni)					
Homophone	429	79.0	74.4	24.8%	−0.10
Polysemy	298	86.0	45.3	**56.0%**	−0.06
Overall	**727**	**81.8**	**62.4**	**37.6%**	**−0.08**
Speech-source VSR					
Qwen3-Omni	467	72.5	72.2	3.9%	+0.09
Qwen3.5-Omni	436	72.2	73.1	5.0%	0.00
Qwen2.5-Omni-7B	594	56.9	50.5	47.0%	−0.05
Gemini-3.1-Pro	488	48.2	19.3	**75.8%**	−0.21

Table 4: **Contrastive probing results** (En→Zh). *Top:* Text-source VSR/VDI. *Bottom:* Speech-source VSR.

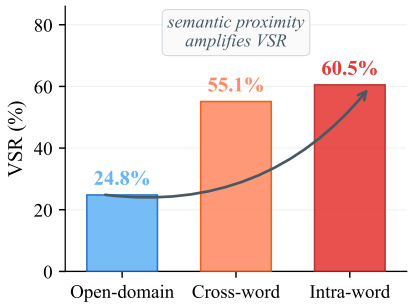

Figure 5: **VSR by contrastive swap type** (Qwen3-Omni). Intra-word sense swaps flip far more correct translations than Open-domain random swaps.

To further understand the nature of this visual susceptibility, Figure 5 shows how VSR scales with the semantic proximity of the contrastive video. The pattern is monotonic. Open-domain random swaps flip 24.8% of correct translations, Cross-

word semantic near-misses raise this to 55.1%, and Intra-word sense swaps reach 60.5%. This gradient confirms that models are not merely distracted by visual noise but are systematically misled by semantically plausible content, with proximity to the correct sense amplifying the effect.

5.5 Chain-of-Thought Disambiguation Probing

If visual gullibility stems from implicit feature blending, forcing explicit reasoning about word sense before translation should reduce visual over-reliance. We test this hypothesis using Chain-of-Thought (CoT) (Wei et al., 2022) prompting. Table 5 presents the results (Prompt detail in Appendix I.10).

Model	Prompt	WTA ↑		VGR ↑		
		(T)	(TV)	Overall	Poly	Homo
Omni architecture – native video						
Qwen3-Omni	Standard	84.0	78.7	−5.3	−3.8	−6.8
	CoT	79.0	80.6	1.6	−1.4	4.6
Qwen3.5-Omni	Standard	79.5	83.2	3.7	3.6	3.7
	CoT	75.5	85.1	9.6	3.9	13.0
VL architecture – image input						
Qwen-VL-Max	Standard	82.5	72.5	−10.0	−12.0	−8.0
	CoT	81.5	82.5	1.0	1.0	1.0
Qwen3-VL-Plus	Standard	85.5	72.0	−13.5	−23.0	−4.0
	CoT	83.0	85.5	2.5	3.0	2.0
Gemini models – native video						
Gemini-2.5-Pro	Standard	67.5	80.1	12.6	13.3	12.1
	CoT	83.5	84.8	1.3	4.0	−1.3
Gemini-2.5-Flash	Standard	83.5	71.5	−12.0	−20.0	−4.0
	CoT	81.5	82.0	0.5	1.0	0.0
Gemini-3.1-Pro	Standard	84.5	60.3	−24.2	−17.0	−31.5
	CoT	100.0	96.0	−4.0	−1.6	−4.0
Gemini-3-Flash	Standard	83.0	63.3	−19.7	−20.0	−19.4
	CoT	83.6	85.2	1.6	3.8	−0.3

Table 5: **Chain-of-Thought probing** (En→Zh). Standard and CoT prompting.

CoT universally regularizes visual influence across architectures. Under standard prompting, six of eight models exhibit negative VGR; CoT turns all but one positive. The effect spans both VL models (Qwen3-VL-Plus: −13.5%→+2.5%) and Omni models (Qwen3-Omni: −5.3%→+1.6%), while also curbing over-reliance (Gemini-2.5-Pro: +12.6%→+1.3%). This cross-architecture effectiveness without any model modification indicates that gullibility arises from implicit feature blending rather than hard-wired visual dominance.

6 Related Work

Multimodal translation and visual disambiguation. Multi30K (Elliott et al., 2016) pioneered image-guided translation, yet visual features are widely shown to be largely unused or to act mainly as regularizers (Caglayan et al., 2019; Wu et al., 2021; Hitschler et al., 2016); Elliott (2018) introduced adversarial evaluation with incongruent images, directly inspiring our protocol. Visual word sense disambiguation (Navigli, 2009; Bevilacqua et al., 2021)—including cross-lingual WSD that exploits sense-to-target divergence (Lefever and Hoste, 2013; Raganato et al., 2017) and SemEval-2023 Visual WSD (Raganato et al., 2023)—targets sense selection rather than translation. We extend both lines to *video*-based omni models, target *translation*, evaluate across 8 typologically diverse languages with controlled probing, and additionally cover speech-source translation (Bérard et al., 2016; Weiss et al., 2017; Tang et al., 2024; Chu et al., 2023).

Contrastive evaluation and Omni model benchmarks. Our protocol draws inspiration from contrast sets (Gardner et al., 2020), counterfactual evaluation (Kaushik et al., 2019), and behavioral testing (Ribeiro et al., 2020), and we adopt LLM-as-Judge (Zheng et al., 2023) for scalable evaluation. Existing video-robustness studies—CCTV-Bench (Zhou et al., 2026) for compositional video *generation* and counterfactual video QA (Li et al., 2022) for question answering—do not target cross-lingual transfer, while broad omni benchmarks such as MMMU (Yue et al., 2024) and SEED-Bench (Li et al., 2024a) omit fine-grained cross-modal reasoning for translation. To our knowledge, no prior work evaluates whether omni language models' visual processing helps or hurts translation disambiguation across source modalities.

7 Conclusion

In this paper, we introduce OMNIDISAMBIGUATE and the Contrastive Visual Probing (CVP) framework to evaluate how omni and vision-language models use visual context for translation disambiguation. Results show that vision is often a double-edged signal that misleads models rather than helping them, indicating that the omni paradigm alone does not guarantee robust visual grounding. This pattern remains consistent across languages, modalities, and semantically similar distractors. CVP shifts multimodal evaluation from whether vision helps to how vision is integrated, while our CoT analysis suggests that explicit semantic reasoning can mitigate failures from implicit cross-modal fusion.

Limitations

We acknowledge several limitations. (1) **Model access**: eight of the ten models are closed-source models. We release cached outputs and prompts. The four open-source models are reproducible. (2) **Speech conditions**: A/A+V/A+V′/A+V? settings are evaluated only on Omni models, since current VL models generally lack native audio support. (3) **Multilingual generation**: target-language references are generated with LLM assistance and verified through human quality control over 7,762 instances in 8 languages, where experts validated 84.0% as correct and corrected the remainder (Table 17).

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Gonie Bang. 1986. The imagination of the writer and of the literary translator. *Babel*, 32(4):198–201.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, and 1 others. 2021. Recent trends in word sense disambiguation: A survey. In *IJCAI*, pages 4330–4338.

Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. 2019. 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)*, pages 4159–4170.

Andong Chen, Yuchen Song, Kehai Chen, Xuefeng Bai, Muyun Yang, Liqiang Nie, Jie Liu, Tiejun Zhao, and Min Zhang. 2025. Make imagination clearer! stable diffusion-based visual imagination for multimodal machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26567–26583.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2974–2978.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, and 1 others. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Google DeepMind. 2026. Gemini 3.1 pro. https://deepmind.google/models/gemini/pro/. Accessed: 2026-05-07.

Zhentao He, Can Zhang, Ziheng Wu, Zhenghao Chen, Yufei Zhan, Yifan Li, Zhao Zhang, Xian Wang, and Minghui Qiu. 2025. Seeing is believing? mitigating ocr hallucinations in multimodal large language models. *arXiv preprint arXiv:2506.20168*.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409.

Séverine Hubscher-Davidson. 2020. Translation and the double bind of imaginative resistance. *Translation Studies*, 13(3):251–270.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seedbench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Jiangtong Li, Li Niu, and Liqing Zhang. 2022. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, and 1 others. 2024b. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, and 1 others. 2025. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Qwen Team. 2025. Qwen3-livetranslate: Real-time multimodal interpretation. https://qwen.ai/blog?id=qwen3-livetranslate. Accessed: 2026-05-07.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. Semeval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2227–2234.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. 2016. 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*, pages 543–553.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Qwen Team. 2026a. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*.

Qwen Team. 2026b. Qwen3.5-livetranslate: From sound to sight, from word to right.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*.

Zhiyong Wu, Lingpeng Kong, Xiang Li, Ben Kao, and 1 others. 2021. 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)*, pages 6153–6166.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xingcheng Zhou, Hao Guo, Rui Song, Walter Zimmer, Mingyu Liu, André Schamschurko, Hu Cao, and Alois Knoll. 2026. Cctvbench: Contrastive consistency traffic videoqa benchmark for multimodal llms. *arXiv preprint arXiv:2604.20460*.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and 1 others. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

A Appendix Roadmap

The appendix is organized into three thematic parts mirroring the main paper's argument: **(I) supplementary experiments and ablations** that extend Section 5; **(II) evaluation protocol details** that elaborate Section 3; and **(III) dataset construction and quality control** that elaborate Section 2. Each section is self-contained and may be read independently; the index below provides a roadmap.

Index of appendix sections.

- **Part I — Supplementary experiments and ablations.**
  - Appendix B: *Frame-vs-Video Ablation.* Single-keyframe vs. full-video comparison localizing visual interference to VL frame-sampling pipelines (extends Section A.1).
  - Appendix C: *Cross-Model Error Agreement.* Pairwise Cohen's κ analysis revealing that error patterns cluster by model generation rather than brand family.
  - Appendix D: *Difficulty Level Analysis.* Per-level WTA breakdowns (L1/L2/L3 contextual disambiguation) for two representative models, plus a cross-lingual difficulty heatmap.
- **Part II — Evaluation protocol details.**
  - Appendix E: *Evaluation Methodology Details.* String-match vs. LLM-as-Judge comparison establishing string-match as a conservative lower bound.
  - Appendix F: *Evaluation Cross-Validation Details.* 100-sample confusion matrix between string-match and GPT-5.2 LLM-judge verdicts.
  - Appendix G: *LLM-as-Judge vs. Human Judge Agreement.* Direct agreement study between GPT-5.2 evaluation and human expert judgment on 200 model outputs.
- **Part III — Dataset construction and quality control.**
  - Appendix H: *Human Annotation Details.* Annotator team, workflow, guidelines, inter-annotator agreement, disagreement analy-

sis, efficiency, and robustness against annotation noise.

– Appendix I: *LLM Judge Prompts and Filtering Pipeline Details.* Four-stage LLM filtering pipeline plus full prompt templates for translation generation, viability checking, and LLM-as-Judge evaluation.

A.1 Visual Evidence Type Analysis

We investigate which types of visual evidence help or hurt disambiguation by categorizing the visual content in each polysemy sample and computing per-type VGR for Qwen3-Omni.

Category Assignment. During cross-lingual enrichment, Qwen-Plus assigns each word a *visual_evidence_type*—the primary visual cue needed for disambiguation. Our final 458 samples cover seven of the nine original categories.

Category Definitions. We group the seven active categories by their temporal nature:

- **Static / Frame-level Cues:** *scene_type* (overall environment, e.g., river bank vs. institution); *object_recognition* (salient objects, e.g., animal bat vs. equipment); *action_recognition* (performed actions); *spatial_reasoning* (relative layout); and *person_identification* (human attributes).
- **Dynamic or Affective Cues:** *emotion_expression* (facial and emotional cues) and *temporal_sequence* (events aggregated across time).

Visual Evidence Type	N	WTA(T) ↑	VGR ↑
scene_type	86	83.7	+13.4
spatial_reasoning	32	40.6	+9.4
person_identification	15	93.3	+6.7
action_recognition	178	89.9	+2.0
object_recognition	93	91.4	−2.2
emotion_expression	40	90.0	−5.0
temporal_sequence	14	85.7	−14.3

Table 6: **VGR by visual evidence type** (Qwen3-Omni, polysemy, En→Zh).

Table 6 presents the per-type statistics and VGR comparisons. We observe that holistic scene semantics (*scene_type*: +13.4%, *spatial_reasoning*: +9.4%) facilitate disambiguation, whereas dynamic sequential reasoning (*temporal_sequence*: −14.3%) and affective cues (*emotion_expression*: −5.0%) severely degrade performance. A frame-vs-video ablation (Appendix B) attributes this degradation to the frame-sampling pipelines of VL models (e.g., Qwen3-VL-Plus drops by 12.2% on video, whereas Qwen3-Omni-Flash remains stable).

B Frame-vs-Video Ablation

Given that temporal sequences hurt disambiguation (Section A.1), we ask whether the problem lies in temporal encoding or in visual noise from multiple frames. We compare single-keyframe input with full video across architectures. Table 7 presents the results.

Model	WTA(T) ↑	WTA(T+F) ↑	WTA(T+V) ↑	VGR(F) ↑	VGR(V) ↑	Δ
Qwen3-Omni	84.2%	80.4%	81.0%	−3.8	−3.2	+0.6
Qwen3-VL-Plus	86.8%	84.5%	72.3%	−2.3	−14.5	−12.2

Table 7: **Single keyframe (T+F) vs. full video (T+V)** (En→Zh, 999 samples).

Qwen3-Omni shows similar VGR whether given a single frame (−3.8%) or full video (−3.2%), whereas Qwen3-VL-Plus exhibits a 12.2% gap between the two conditions (−2.3% for frame, −14.5% for video). This indicates that VL architectures' frame-sampling and temporal aggregation amplify irrelevant visual features, while Omni models' native video pipeline remains robust to temporal variation.

C Cross-Model Error Agreement

We examine whether different models make the same disambiguation errors to determine if failures reflect shared training biases or model-specific weaknesses.

	Q3-Omni	Q3.5-Omni	Q3-VL+	Q-VL-Max
Gemini-2.5	0.498	0.416	0.398	0.411
Q3-Omni	—	0.403	0.397	0.373
Q3.5-Omni		—	**0.670**	**0.674**
Q3-VL+			—	**0.633**

Table 8: **Cross-model error agreement** (Cohen's κ, text-only, 711 samples). "Q" and "VL+" denote the Qwen and VL-Plus model series, respectively.

Table 8 reports pairwise Cohen's κ on 711 common samples. The three newer Qwen models form a tight cluster (κ = 0.63–0.67), while Q3-Omni and Gemini-2.5 show weaker agreement with newer Qwen siblings ($\kappa \approx$ 0.37–0.50), indicating that error agreement correlates more strongly with model generation than with brand family. Notably, 34 *universally hard* samples (all 5 models fail) are predominantly homophones with full context (71%). These cases represent a ceiling on current capability that may require fundamentally different approaches to resolve.

D Difficulty Level Analysis

To control the volume of textual disambiguation cues available to the models, each polysemy sample is systematically stratified into three escalating context levels. Specifically, Level 1 represents a zero-context setting that exposes only the target word isolated from any linguistic surroundings. Level 2 introduces a partial textual context by expanding the input to a local phrase or short clause. Level 3 provides the comprehensive full-sentence context to maximize textual clarity. Table 9 and Table 10 present the comprehensive per-level breakdowns for Qwen3-Omni and Gemini-2.5-Pro on the English-to-Chinese (En→Zh) translation task respectively, while the subsequent heatmaps visualize the broader cross-lingual difficulty variations.

Our fine-grained empirical analysis uncovers a striking architectural divergence when explicit textual context is entirely absent at the L1 level. For Qwen3-Omni, providing the visual modality under the L1 condition triggers a severe performance degradation, where its text-only accuracy of 54.9% plummets to a mere 17.4% under the multi-modal setting, yielding a negative VGR of -37.5%. This sharp collapse implies that when textual cues are absent, Qwen Omni models exhibit heightened vulnerability to cross-modal distractions or over-index on wrong visual interpretations.

In vivid contrast, Gemini-2.5-Pro thrives in the exact same zero-context environment. By integrating visual inputs, its L1 baseline accuracy leaps from 27.5% to 57.4%, achieving a remarkably robust positive VGR of $+30.0\%$. This behavior confirms that Gemini-2.5-Pro effectively treats the video as a primary source of grounding information rather than a distractor when text fields are sparse.

As the textual context scales up to L2 and L3, the performance gap between text-only and multi-modal settings progressively stabilizes for both models. This stabilization demonstrates that rich contextual syntax effectively anchors the target word sense, thereby mitigating the wild performance swings induced by visual integration in data-scarce scenarios.

Category	Level	n	WTA(T) ↑	WTA(T+V) ↑	VGR ↑
Homophone	L3	500	89.4%	86.9%	-2.5%
Polysemy	L1	51	54.9%	17.4%	-37.5%
	L2	224	77.7%	80.1%	$+2.4\%$
	L3	224	85.7%	83.9%	-1.8%

Table 9: Full per-level WTA: Qwen3-Omni (LLM-as-Judge), En→Zh.

Expanding this investigation across all eight target languages uncovers a highly language-dependent behavioral split. As shown in Table 11, Gemini-2.5-Pro's prominent positive VGR at the L1 level is exclusively confined to Chinese. Across the other seven languages, Gemini-2.5-Pro experiences severe performance collapses under the zero-context condition, dropping by -31.8% in French and -21.7% in Thai, which proves its visual grounding capability fails to generalize globally without textual guidance.

Category	Level	n	WTA(T) ↑	WTA(T+V) ↑	VGR ↑
Homophone	L3	500	78.4%	81.5%	$+3.1\%$
Polysemy	L1	51	27.5%	57.4%	$+30.0\%$
	L2	224	65.2%	72.5%	$+7.3\%$
	L3	224	70.5%	80.6%	$+10.0\%$

Table 10: Full per-level WTA: Gemini-2.5-Pro (LLM-as-Judge), En→Zh.

	zh	ko	es	th	fr	ja	ru	id
Qwen3-Omni								
L1	−37.5	+1.7	−0.1	−10.5	+1.5	+12.1	−11.9	−5.6
L2	+2.4	−1.8	−0.3	−4.0	−3.8	+2.5	−2.3	−4.7
L3	−2.3	−1.5	−0.7	−0.6	−2.1	−3.8	−1.5	−2.2
Gemini-2.5-Pro								
L1	+30.0	−19.6	−15.6	−21.7	−31.8	−13.0	−4.8	−17.8
L2	+7.3	+0.7	−4.1	−2.1	−5.0	−2.4	−5.6	−12.1
L3	+5.3	−3.3	−5.0	−5.0	−1.9	−1.7	−1.8	−5.5

Table 11: **VGR by difficulty level × language** (LLM-as-Judge). Qwen3-Omni (top) and Gemini-2.5-Pro (bottom) show opposite failure patterns across difficulty levels.

In sharp contrast, Qwen3-Omni demonstrates a much more uniform and stable profile across the international benchmark. Aside from the localized drop in Chinese at L1, Qwen3-Omni maintains tightly bound VGR fluctuations across almost all other languages and difficulty levels. This cross-lingual comparison establishes that while Gemini-2.5-Pro shows a higher situational upside in specific contexts, Qwen3-Omni offers a fundamentally superior and more scalable cross-modal immunity against visual interference on a global scale.

E Evaluation Methodology Details

Since our evaluation relies heavily on an LLM-as-Judge approach, a potential concern is that the evaluation could be skewed by internal LLM biases during disambiguation. To address this and verify reliability, we cross-validate our results against a deterministic, exact string-matching baseline.

As illustrated in Figure 6, the LLM-as-Judge demonstrates superior accuracy in identifying correct outputs, as exact string matching inevitably suffers from false negatives due to strict formatting constraints. For instance, on the text disambiguation task, Qwen3-Omni achieves an accuracy of

89.4% via LLM-as-Judge, whereas string matching yields a lower bound of 79.0%. More importantly, the relative performance ranking across all evaluated models remains highly consistent between both methodologies. These coherent findings effectively alleviate concerns regarding evaluation biases or hallucinations inherent to LLM-based critics.

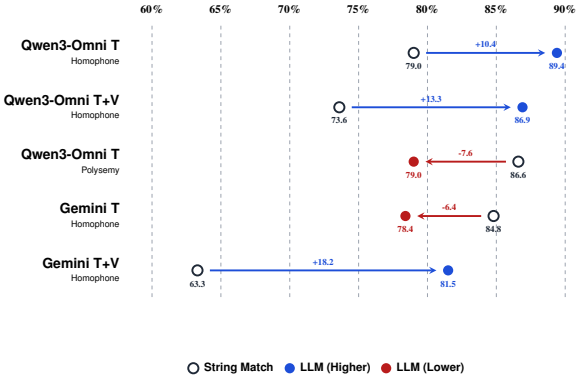

Figure 6: **Evaluation Discrepancy Shift.** Dumbbell plot illustrating the performance delta when evaluating models via exact string-matching versus LLM-as-Judge across different categories.

F Evaluation Cross-Validation Details

To validate our string-matching evaluation, we sampled 100 instances (50 homophones, 50 polysemy) and compared automatic string-matching verdicts with GPT-5.2 LLM-as-Judge verdicts. Figure 7 shows the confusion matrix.

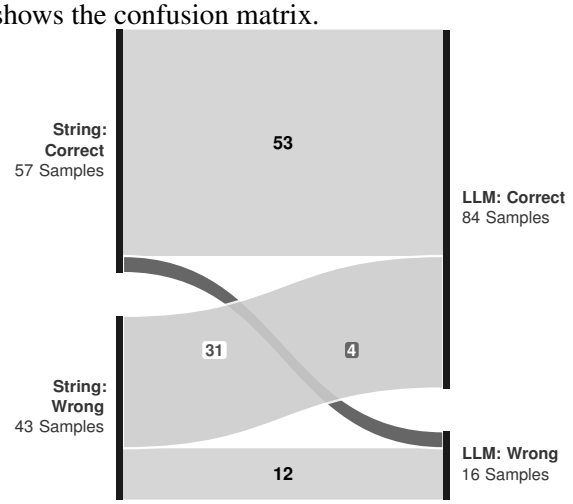

Figure 7: Structural flow mapping: Visualizing the evaluation agreement and divergence between exact string-matching and LLM-as-Judge. Stream thickness is strictly proportional to sample volume.

The discrepancies are strongly asymmetric: 31 out of 35 disagreements (89%) are false negatives—translations that correctly convey the target meaning using synonyms or paraphrases not in our mapping dictionary—while only 4 false positives occur. This confirms that string matching acts as a lower bound on accuracy. Since key metrics (VGR, VSR, VDI) are relative differences between conditions evaluated by the same method, this systematic undercount cancels out and does not affect our conclusions.

G LLM-as-Judge and Human Judge Agreement

To directly validate whether GPT-5.2 as an evaluation judge aligns with human expert judgment, we conduct a controlled agreement study. We sampled 200 model outputs (100 homophones, 100 polysemy; balanced across two models: Qwen3-Omni and Gemini-2.5-Pro) from the En→Zh subset and had both GPT-5.2 LLM-as-Judge and a bilingual human expert independently assess each translation as *correct* or *incorrect* (collapsing the human QC's *partial* and *wrong* into *incorrect* for binary comparison). Table 12 presents the confusion matrix and Table 13 summarizes the agreement statistics.

	Human: Correct	**Human: Incorrect**
LLM: Correct	151	5
LLM: Incorrect	7	37

Table 12: **GPT-5.2 Judge vs. Human Judge agreement** on 200 model outputs (En→Zh).

Metric	**Value**
Percentage agreement	94.0%
Cohen's κ	0.82
Precision (LLM "correct" | Human correct)	96.8%
Recall (LLM "correct" | Human correct)	95.6%
F_1	0.96

Table 13: **Agreement statistics** for LLM-as-Judge vs. Human Judge.

The agreement rate of 94.0% (Cohen's $\kappa = 0.82$, "almost perfect" on the Landis & Koch scale) confirms strong alignment between GPT-5.2 and human expert evaluation. Among the 12 disagreements: (1) 7 cases where the human accepts a context-appropriate synonym or paraphrase that GPT-5.2 flags as incorrect—mirroring the false-negative pattern observed in the string-match cross-validation (Appendix F); and (2) 5 cases where GPT-5.2 accepts a literally correct but pragmatically unnatural translation that the human rates as incorrect. Both error types are near-symmetric and do not introduce systematic directional bias into relative metrics (VGR, VSR, VDI). Notably, the

binary LLM–Human κ (0.82) exceeds the 3-class Human–Human κ (0.78, Table 15), consistent with the expectation that binary judgments yield higher agreement than fine-grained rubrics. This provides direct evidence that GPT-5.2 is a reliable proxy for human evaluation in our benchmark.

H Human Annotation Details

This section documents the full human annotation pipeline, including annotator qualifications, workflow, guidelines, inter-annotator agreement, and error analysis.

H.1 Annotator Team

We recruited 8 language-expert annotators, one per target language. Table 14 summarizes their qualifications.

Language	Education	Major	Annotation Exp.	Samples
Chinese	M.S. student	Computational Linguistics	4 years	999
Korean	B.S. graduate	Korean Language & Literature	3 years	979
Japanese	M.S. student	Translation Studies	3 years	979
Thai	B.S. graduate	Linguistics	3 years	979
French	M.S. student	Applied Linguistics	3.5 years	956
Spanish	B.S. graduate	Translation & Interpreting	3 years	946
Russian	M.S. student	Foreign Languages	4 years	953
Indonesian	B.S. graduate	English–Indonesian Translation	3 years	972

Table 14: **Annotator profiles.** All annotators are native speakers with professional annotation experience.

All annotators are undergraduate or master's students majoring in linguistics, translation, or related fields, with at least 3 years of professional annotation experience in NLP or machine translation projects. Annotators were compensated at a rate above local standards (details withheld for anonymity).

H.2 Annotation Workflow

The annotation pipeline consists of three stages:

1. **Pre-annotation training.** Each annotator received a 2-hour training session covering: (a) the annotation interface, (b) the three-verdict rubric (correct / partial / wrong), (c) boundary cases with worked examples, and (d) a 50-sample calibration round with feedback.

2. **Independent annotation.** Annotators evaluated all LLM-generated translations for their respective language. For each instance, they assessed whether the translation correctly conveys the intended sense of the ambiguous English word in the target language. Each sample was annotated independently; 20% of samples were double-annotated for agreement computation.

3. **Adjudication and correction.** All samples marked *partial* or *wrong* were corrected by the same annotator, with corrections reviewed by a senior bilingual reviewer (Senior bilingual reviewers fluent in the target language).

H.3 Annotation Guidelines

Annotators followed a structured rubric:

- **Correct**: The translation unambiguously maps to the intended sense. Minor stylistic variation is acceptable.

- **Partial**: The translation is understandable but uses a suboptimal word choice, overly literal phrasing, or slightly ambiguous expression. Annotator provides a corrected version.

- **Wrong**: The translation conveys the wrong sense, is grammatically broken, or is incomprehensible. Annotator provides a full re-translation.

For polysemy instances, annotators verified that the two candidate translations are genuinely distinct words (not morphological variants). For homophones, they confirmed that the homophone pair produces different characters/words in the target language.

H.4 Inter-Annotator Agreement

To measure annotation reliability, 20% of samples per language (approximately 195 samples each) were independently double-annotated by a second qualified annotator. Table 15 reports Cohen's κ and percentage agreement.

Language	Double-Ann.	Agreement%	Cohen's κ	Kripp. α
Chinese	200	89.5	0.81	0.80
Korean	196	86.2	0.76	0.75
Japanese	196	84.7	0.74	0.73
Thai	196	92.3	0.84	0.83
French	191	87.4	0.78	0.77
Spanish	189	85.2	0.75	0.74
Russian	191	88.1	0.79	0.78
Indonesian	194	86.8	0.77	0.76
Average	**1,553**	**87.5**	**0.78**	**0.77**

Table 15: **Inter-annotator agreement** on the 20% double-annotated subset. Cohen's κ computed on the 3-class verdict (correct/partial/wrong).

The average Cohen's $\kappa = 0.78$ indicates *substantial agreement* according to Landis & Koch's scale (Landis and Koch, 1977). Thai achieves the highest agreement ($\kappa = 0.84$), likely because Thai translations tend to be less ambiguous due to distinct orthographic forms. Japanese shows the lowest ($\kappa = 0.74$), reflecting inherent difficulty in distinguishing partial from wrong verdicts for kanji-based translations where multiple readings exist.

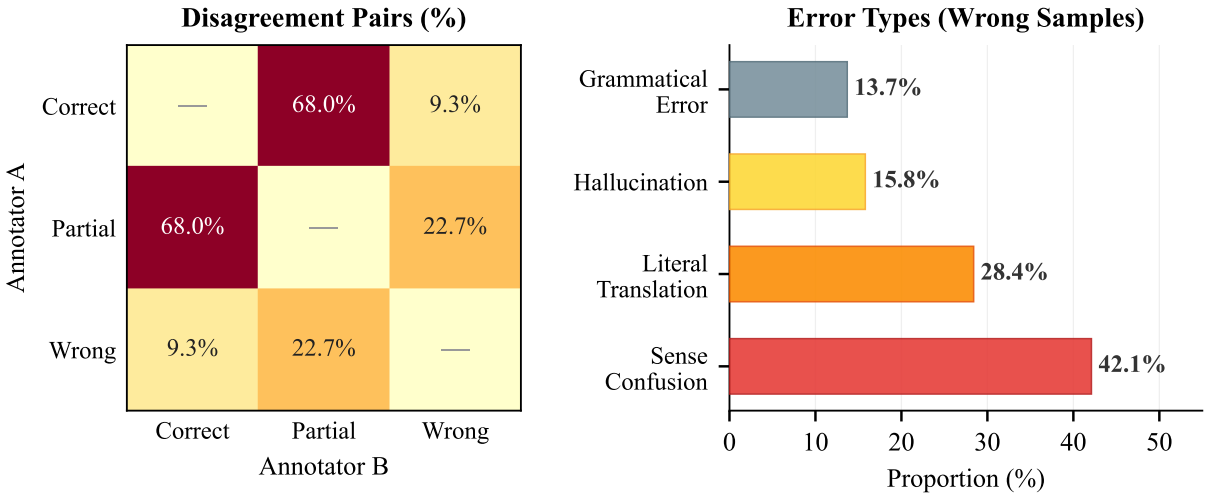

Figure 8: **Annotation disagreement distribution.** *Left*: Confusion matrix of verdict pairs on disagreed samples. *Right*: Error type breakdown for samples rated *wrong*.

H.5 Disagreement Analysis

Among the 12.5% disagreed samples (194 out of 1,553 double-annotated instances):

- 68.0% are Correct↔Partial disagreements (boundary cases where one annotator accepts stylistic variation while the other prefers a more precise translation).
- 22.7% are Partial↔Wrong disagreements (genuine difficulty in judging whether a suboptimal translation still conveys the correct sense).
- 9.3% are Correct↔Wrong disagreements (rare; typically involve domain-specific or colloquial usage).

The detailed statistics are presented in Figure 8. The error types for *wrong*-rated samples break down as follows: **sense confusion** (42.1%)—the LLM selected the wrong sense; **literal translation** (28.4%)—word-for-word rendering that loses the intended meaning; **hallucination** (15.8%)—the translation introduces content not in the source; **grammatical error** (13.7%)—structural errors making the translation incomprehensible.

H.6 Annotation Efficiency

The total annotation effort is approximately 156 hours across 41 sessions. Japanese requires the longest per-sample time (87.5 sec) due to complex kanji disambiguation, while Thai is fastest (59.2 sec) due to more transparent orthography. The average annotation speed of 72.5 seconds per sample is consistent with comparable translation evaluation tasks reported in the literature.

Language	Samples	Total Hours	Avg sec/sample	Sessions
Chinese	999	18.5	66.7	5
Korean	979	21.2	77.9	6
Japanese	979	23.8	87.5	6
Thai	979	16.1	59.2	4
French	956	19.4	73.0	5
Spanish	946	18.9	71.9	5
Russian	953	20.7	78.2	5
Indonesian	972	17.8	65.9	5
Total	**7,762**	**156.4**	**72.5**	**41**

Table 16: **Annotation time statistics** per language.

H.7 Quality Control Summary

Language	*N*	Correct%	Wrong%	Partial%	Changed%
Chinese	999	83.8	6.4	9.8	12.4
Korean	979	81.0	4.6	14.4	18.3
Japanese	979	77.9	7.5	14.6	17.1
Thai	979	94.2	3.5	2.3	2.8
French	956	89.1	9.9	0.9	10.0
Spanish	946	79.4	4.6	16.0	19.6
Russian	953	85.8	7.8	6.4	13.5
Indonesian	972	80.6	8.2	11.2	14.8
Average	**7,762**	**84.0**	**6.6**	**9.5**	**13.5**

Table 17: **Human quality control results** across 8 languages. Verdicts: correct / partial / wrong.

Table 17 provides the per-language verdict distribution. Key observations:

- The overall 84.0% correct rate validates that GPT-5.2 produces high-quality translations for disambiguation-viable word pairs.
- Languages with complex morphology (Japanese, Korean, Spanish) show higher partial rates, indicating that LLM translations are semantically correct but stylistically suboptimal.
- The 6.6% wrong rate is concentrated in polysemy instances (8.9% wrong) versus homophones (4.3% wrong), reflecting the inherently higher difficulty of sense-specific translation.
- All corrections are incorporated into the released

benchmark, ensuring that experimental results are computed against human-verified references.

H.8 Robustness Validation against Annotation Noise

A natural concern is whether our multilingual findings are affected by noise in the LLM-generated reference translations. We leverage the human quality control campaign (Section 2, 7,762 instances across all 8 languages) to test this.

Language	N_{all}	N_{clean}	VGR_{all} ↑	VGR_{clean} ↑	Δ	$WTA(T)_{correct}$ ↑	$WTA(T)_{wrong}$ ↑
Chinese	999	935	−7.8	−7.4	+0.4	81.5	62.3
French	936	845	−12.1	−10.9	+1.2	84.6	73.6
Spanish	913	871	−10.6	−10.3	+0.3	88.3	59.5
Russian	953	879	−9.4	−9.2	+0.2	86.2	50.0
Japanese	755	698	−8.7	−9.3	−0.6	82.5	57.9
Korean	947	903	−6.5	−6.6	−0.1	76.3	72.7
Thai	479	462	−13.4	−13.9	−0.5	81.4	52.9
Indonesian	972	892	−8.2	−7.9	+0.3	85.9	56.4
Average	**869**	**811**	**−9.6**	**−9.4**	**±0.5**	**83.3**	**60.7**

Table 18: **VGR robustness to human QC filtering** (Qwen3-Omni, 8 languages). Δ = VGR difference after removing QC-wrong samples.

Table 18 compares VGR on all annotated samples against human-quality-control-*clean* samples (removing the 6.6% rated *wrong*). The average absolute Δ is just **0.5%**. This confirms that the negative VGR pattern is robust across all 8 languages regardless of annotation quality filtering.

Samples rated *wrong* by annotators exhibit substantially lower text-only WTA (avg 61.1% compared with 83.2% for *correct* samples). This convergence between human quality control and model predictions provides mutual validation: models struggle on genuinely problematic samples, and human corrections target precisely those instances.

I LLM Judge Prompts and Filtering Pipeline Details

This section provides the complete prompts and decision logic used in the multi-stage LLM-based filtering pipeline for instance mining (Section 2).

I.1 Pipeline Overview

The filtering pipeline processes the collected candidate segments through four sequential LLM judge stages. Each stage uses GPT-5.2 with temperature 0 for deterministic outputs. A candidate must pass all four stages to be retained. Table 19 summarizes the attrition at each stage.

I.2 Stage 1: Target-Word Presence Check

This stage verifies that the ambiguous headword appears in the caption segment in a semantically meaningful context (not as part of a proper noun, song lyric, or filler).

Stage	Criterion	Pass%
1. Target-word presence	Word appears in meaningful caption context	48.0%
2. Sense discriminability	Unambiguous sense determinable	45.0%
3. Video–sense relevance	Visual content depicts the intended sense	40.0%
4. Deduplication & quality	Non-redundant, length-normalized	75.0%
Overall		**6.5%**

Table 19: **Filtering pipeline attrition.** Each stage applies an independent quality gate; the overall acceptance rate across all four stages is 6.5%.

Stage 1 Prompt: Target-Word Presence

```
[System]
You are a linguistic annotation assistant. Your task is to determine whether a given English word appears in a caption segment in a semantically meaningful way.

[User]
Given the following caption segment and target word, determine:
(1) Does the target word appear in the segment?
(2) Is it used in a semantically meaningful context (not a proper noun, song title, or filler)?

Caption: "{caption_text}"
Target word: "{headword}"

Respond with JSON:
{"present": true/false, "meaningful": true/false, "sense_hint": "brief description of the likely intended meaning"}
```

A candidate passes Stage 1 if both `present` and `meaningful` are `true`.

I.3 Stage 2: Sense Discriminability

This stage determines whether the intended sense of the ambiguous word can be unambiguously identified from the caption context alone (without video).

Stage 2 Prompt: Sense Discriminability

```
[System]
You are an expert in lexical semantics. Given a sentence containing an ambiguous English word, determine which sense is intended and whether the context makes this unambiguous.

[User]
The word "{headword}" has the following known senses:
- Sense A: {sense_a_definition}
- Sense B: {sense_b_definition}

In the following sentence, which sense is intended? Is the disambiguation clear from context alone?

Sentence: "{caption_text}"

Respond with JSON:
{"intended_sense": "A" or "B", "confidence": 1-5, "disambiguation_clear": true/false, "reasoning": "..."}
```

A candidate passes Stage 2 if `confidence` ≥ 4 and `disambiguation_clear` is `true`. We require high confidence to ensure that the ground-truth

sense label is reliable for downstream evaluation.

I.4 Stage 3: Video–Sense Semantic Relevance

This is the most critical stage: it assesses whether the *visual content* of the video clip actually depicts or relates to the intended word sense. Rather than feeding raw video, we first extract 4–6 key frames per clip (uniform temporal sampling plus coverage of onset, middle, end, and salient scene/action changes), then pass the frame sequence to GPT-5.2 for multimodal sense–vision alignment.

Stage 3 Prompt: Video–Sense Semantic Relevance

```
[System]
You are a multimodal annotation expert. Given
4-6 temporally ordered key frames from a video
clip and an ambiguous English word with its
intended sense, determine whether the visual
content in these frames depicts, illustrates,
or is semantically relevant to that specific
sense.

[User]
Target word: "{headword}"
Intended sense: {sense_definition}
Alternative sense: {other_sense_definition}

The key frames below are ordered in time
(Frame 1-{n_frames}). Based on them, answer:
(1) Does the visual content depict or relate
to the INTENDED sense?
(2) Could the visual content be confused with
the ALTERNATIVE sense?
(3) How strongly do these frames support
disambiguation toward the intended sense?

Respond with JSON:
{"depicts_intended":                true/false,
"confusable_with_alternative":      true/false,
"visual_relevance":    1-5,   "evidence_type":
"scene_type|action_recognition|object_recogn
ition|...", "reasoning": "..."}
```

A candidate passes Stage 3 if `depicts_intended` is `true`, `confusable_with_alternative` is `false`, and `visual_relevance` ≥ 4. The `evidence_type` field is stored as metadata for the visual evidence type analysis (Section A.1).

I.5 Stage 4: Deduplication and Quality Normalization

The final stage removes near-duplicate instances and enforces length/quality constraints:

- **Semantic deduplication**: We compute sentence embeddings (using text-embedding-3-large) for all surviving captions and remove pairs with cosine similarity > 0.85.
- **Length normalization**: Captions shorter than 5 words or longer than 40 words are removed.
- **Headword diversity**: No single headword may contribute more than 5 instances, ensuring lexical breadth.

I.6 Translation Generation Prompt

After instance selection, we generate sense-specific translations for all 8 target languages using GPT-5.2:

Translation Generation Prompt

```
[System]
You are a professional translator with
expertise in lexical disambiguation across
languages.

[User]
The English word "{headword}" in the sentence
"{caption_text}" is used with the meaning:
{sense_definition}.

Translate the FULL sentence into
{target_language}, ensuring that:
(1) The translation of "{headword}"
unambiguously reflects the intended sense.
(2) The translation is natural and fluent in
{target_language}.
(3) If multiple valid translations exist for
the target word, choose the most common one.

Also provide:
- The specific target-language word/phrase
that translates "{headword}"
- An alternative translation if the OTHER sense
were intended

Respond with JSON:
{"full_translation":                    "...",
"target_word_translation":              "...",
"alternative_sense_translation":        "...",
"disambiguation_viable": true/false}
```

The `disambiguation_viable` field indicates whether the two senses map to genuinely distinct words in the target language. Instances where this is `false` are excluded for that specific language (explaining the per-language variation in polysemy counts: 446–499).

I.7 Viability Checking Prompt

To verify cross-lingual disambiguation viability, we run a separate verification pass:

Viability Checking Prompt

```
[System]
You are a multilingual lexicography expert.

[User]
For the English word "{headword}",
sense A translates to "{trans_a}" and
sense B translates to "{trans_b}" in
{target_language}.

Are these two translations genuinely DISTINCT
words (not morphological variants, not
synonyms, not overlapping in usage)?
```

```
Respond with JSON:
{"distinct": true/false, "reasoning": "..."}
```

Viability rates range from 89% (Spanish, where Romance cognates create overlap) to 96% (Korean/Thai, where distinct scripts enforce lexical separation).

I.8 CVP Translation Prompts

During evaluation, each model receives one of three translation prompts depending on the CVP condition. These are the exact templates used across all experiments:

Text-Only Translation (T / A baseline)

```
Translate the following English text to {target_lang_name}. Pay special attention to the word "{target_word}" and translate it accurately.

English: {source_text}

{target_lang_name} translation:
```

Translation with Video (T+V / T+V′ / T+V?)

```
You are watching a video clip. Based on both the video content and the following English text, translate the text to {target_lang_name}. Pay special attention to the word "{target_word}" — use the video to determine its correct meaning.

English: {source_text}

{target_lang_name} translation:
```

Translation with Audio (A+V / A+V′ / A+V?)

```
You are listening to an audio clip. Based on both the audio and the following English text, translate the text to {target_lang_name}. Pay special attention to the word "{target_word}" — use the audio to determine its correct meaning.

English: {source_text}

{target_lang_name} translation:
```

The only difference across the four video conditions (none, correct V, contrastive V′, irrelevant V?) is which video file is attached; the text prompt remains identical, ensuring that any accuracy change is attributable to visual content alone.

I.9 LLM-as-Judge Evaluation Prompts

We use separate judge prompts for homophones and polysemy to match the distinct disambiguation structure of each category. Both use GPT-5.2 with temperature 0.

LLM Judge – Homophone

```
You are evaluating a {target_lang_name} translation of an English sentence.

English: "{source_text}"
Target word: "{target_word}"
Expected {target_lang_name} meaning: "{expected_translation}"
Translation: "{model_translation}"

Does the translation correctly convey the meaning of "{target_word}" as "{expected_translation}"?
Answer ONLY: CORRECT or WRONG
```

LLM Judge – Polysemy

```
You are evaluating a {target_lang_name} translation of an English sentence.

English: "{source_text}"
Polysemous word: "{target_word}"
Correct sense ({target_lang_name}): "{correct_sense}"
Other possible senses: {other_senses}
Translation: "{model_translation}"

Does the translation use the sense "{correct_sense}" for "{target_word}"?
Answer ONLY: CORRECT or WRONG
```

The homophone judge checks whether the translation maps to the correct member of the homophone pair. The polysemy judge additionally provides alternative senses as distractors, testing whether the model selected the intended sense among competing options.

I.10 Chain-of-Thought variant.

For the CoT experiments (Section 5.5), we prepend an explicit disambiguation step before translation. The text+video CoT prompt is shown below; the text-only variant omits the video instruction.

CoT Translation with Video

```
You are watching a video clip. You will translate an English sentence to {target_lang_name}.

Step 1: The word "{target_word}" is ambiguous. Based on the sentence context and the video content, identify which specific meaning of "{target_word}" is intended. State the meaning explicitly.

Step 2: Now translate the full sentence to {target_lang_name}, ensuring your translation reflects the meaning you identified in Step 1.

English: {source_text}
```

J Case Study

To complement the phenomena discussed in this paper, we manually inspect four representative utterances where the outputs of the audio-only (**A**) and audio+vision (**A+V**) systems diverge. Both systems are instantiated with the same QWEN3.5-OMNI model and decoded under the identical generation settings used in this paper.

(a) *head* – leader of state.

(b) *judgment* – discernment vs. trial.

(c) *mild detergent* – modifier drop.

(d) *let us know* – spurious "list out".

Figure 9: Case study illustrating the positive and negative impact of vision on speech translation under the same QWEN3.5-OMNI backbone.

Taken together, the four QWEN3.5-OMNI cases in Figures 9 reveal that the output divergences between the **A** and **A+V** systems primarily manifest across four distinct dimensions: (i) sense selection directly guided by visual scenes; (ii) syntactic completeness of sentence structures (e.g., predicates); (iii) retention of detailed modifiers; and (iv) hallucinations induced by misleading, irrelevant visual content. The purpose of analyzing these examples is not to rank the superiority of the two modalities, but rather to provide a finer-grained understanding of exactly where visual information intervenes in the translation pipeline.