VIDA: A Dataset for Visually Dependent Ambiguity in Multimodal Machine Translation

♠Jingheng Pan and ♠♡Xintong Wang*
♡Longyue Wang and ♡Liang Ding and ♡Weihua Luo and ♠Chris Biemann*
♠Department of Informatics, Universität Hamburg ♡Alibaba Group
♠{jingheng.pan, xintong.wang, chris.biemann}@uni-hamburg.de
♡liangding.liam@gmail.com,♡{wanglongyue.wly, weihua.luowh}@alibaba-inc.com
Dataset: https://huggingface.co/datasets/p1k0/visually-dependent-ambiguity

Abstract

Ambiguity resolution is a key challenge in multimodal machine translation (MMT), where models must genuinely leverage visual input to map an ambiguous expression to its intended meaning. Although prior work has proposed disambiguation-oriented benchmarks probing the role of vision, we observe that existing benchmarks remain limited by task-format mismatch, narrow ambiguity coverage, or insufficient visual-dependency validation. Moreover, existing ambiguity evaluations are not well suited to diverse ambiguity types in open-ended translation. To address these limitations, we present **VIDA** (Visually-Dependent Ambiguity), a dataset of 2,500 carefully curated instances in which resolving an annotated source span requires visual evidence. We further propose **Disambiguation-Centric Metrics** that use an LLM-as-a-judge classifier to verify whether annotated ambiguous expressions are resolved correctly at the span level. Evaluations with stronger recent LVLMs show that visual disambiguation remains challenging. Using chain-of-thought supervised fine-tuning as a diagnostic setting, we observe stronger out-of-distribution disambiguation than with SFT, with robust gains on collective-noun ambiguities and model-dependent gains on sentence-level ambiguities.

1 Introduction

Multimodal machine translation (MMT) extends neural machine translation by incorporating visual context to improve translation quality (Lala and Specia, 2018; Yao and Wan, 2020). Recent Large Vision Language Models (LVLMs) show impressive performance on MMT benchmarks (Bai et al., 2025; Zhu et al., 2025). However, it remains unclear whether LVLMs truly leverage visual information during MMT. Earlier studies (Elliott, 2018; Wu et al., 2021) showed that replacing or perturbing images often leads to only minor degradations, raising questions about the actual contribution of the visual modality. This concern remains relevant in the LVLM era: stronger visual perception does not necessarily guarantee reliable use of visual evidence during translation.

Ambiguity resolution provides a direct probe of visual dependence in MMT, and existing work has made progress in this direction by introducing multimodal disambiguation benchmarks (Ma et al., 2024; Wang et al., 2024a; Futeral et al., 2023). Yet no existing benchmark combines open-ended MMT generation with broad ambiguity coverage and visual-dependency validation. CoMMuTE (Futeral et al., 2023) is a contrastive benchmark for word-level and gender ambiguity that scores candidate translations rather than requiring models to generate translations freely. 3AM (Ma et al., 2024) targets English–Chinese MMT and primarily focuses on word-level ambiguity, but includes many instances resolvable from text alone. MMA (Wang et al., 2024a) covers sentence-level ambiguity in a VQA-style format rather than translation generation. Additionally, prior work on MMT ambiguity evaluation spans several paradigms, including contrastive evaluation with predefined translation candidates (Futeral et al., 2023) and rule-based matching with lexical variants (Lala and Specia, 2018; Li et al., 2021). While these settings provide ambiguity-oriented evaluation, they are less suitable for open-ended MMT generation with broader ambiguity types and valid paraphrases. General MT metrics such as BLEU and COMET assess overall translation quality, but do not directly indicate whether a particular ambiguous span has been resolved correctly.

Reliable evaluation of visual disambiguation in MMT requires both an MMT dataset with visually-dependent ambiguities and an evaluation method that directly measures disambiguation accuracy. We therefore introduce **VIDA** (**Vi**sually-**D**ependent **A**mbiguity), an MMT dataset of 2,500 instances

*Corresponding authors.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 137–154
November 6-10, 2026 ©2026 Association for Computational Linguistics

capturing visually-dependent translation ambiguities at the word and sentence levels, as well as collective-noun phenomena. We further propose **Disambiguation-Centric Metrics** based on an LLM-as-a-judge classifier. These metrics complement standard translation metrics by explicitly targeting span-level disambiguation accuracy. Even recent LVLMs such as Qwen3.5-27B (Qwen Team, 2026a) resolve all annotated ambiguities in fewer than half of the test instances.

Beyond dataset and evaluation, we use chain-of-thought supervised fine-tuning (CoT-SFT) (Muennighoff et al., 2025) with a manually designed template as a diagnostic setting to test whether explicit visual reasoning supervision improves disambiguation beyond standard SFT. Experiments show robust gains on VIDA-CollN across models, while the effect on VIDA-Sent is model-dependent. The main contributions of this paper are:

- We introduce **VIDA**, an MMT dataset featuring visually-dependent ambiguities at both word and sentence levels.
- We propose **Disambiguation-Centric Metrics** to directly measure disambiguation accuracy, complementing standard translation metrics for MMT disambiguation.
- We use CoT-SFT as a diagnostic setting to examine whether explicit visual-reasoning supervision improves disambiguation beyond standard SFT.

2 Related Work

Multimodal Disambiguation Datasets. Prior work has studied visual disambiguation through translation-based benchmarks (Lala and Specia, 2018; Li et al., 2021; Ma et al., 2024), contrastive MMT benchmarks (Futeral et al., 2023), and VQA-style ambiguity benchmarks (Wang et al., 2024a). Building on such benchmarks, DeDiT (Liu et al., 2025) examines explicit visual reasoning in MMT and evaluates cross-dataset transfer from Multi30K to CoMMuTE. In comparison, VIDA targets open-ended MMT generation with broader ambiguity coverage, retaining only cases whose correct interpretation requires visual evidence. Sharing the motivation of DeDiT, we use CoT-SFT as a diagnostic setting, but since VIDA separates ambiguity types into subsets, we test whether visual reasoning transfers from word-level ambiguities to held-out sentence-level and collective-noun ambiguities rather than across datasets.

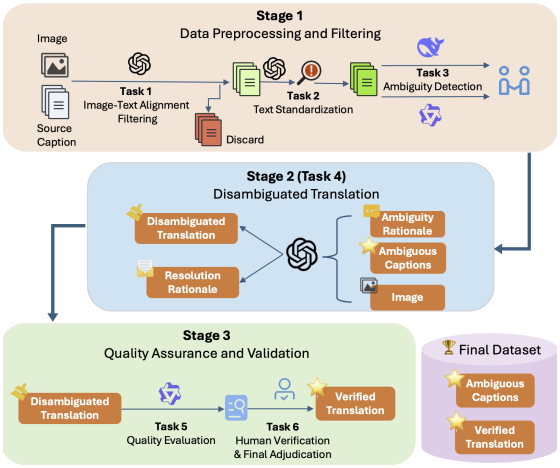


Figure 1: Three-stage VIDA curation pipeline

Disambiguation Metrics for MMT. Prior work has evaluated multimodal ambiguity with contrastive candidate ranking (Futeral et al., 2023), rule-based matching over predefined lexical variants (Lala and Specia, 2018; Li et al., 2021), and general MT metrics (Papineni et al., 2002; Rei et al., 2020). These approaches are less suitable for open-ended MMT generation with broader ambiguity types, where correct disambiguation may be expressed through paraphrases or lexical variation. We therefore propose Disambiguation-Centric Metrics using an LLM-as-a-judge classifier to directly assess span-level disambiguation accuracy. A detailed comparison is provided in Appendix A.

3 Dataset Curation

To construct an MMT dataset with visually-dependent translation ambiguities, we adopt a three-stage semi-automatic pipeline, as shown in Figure 1. Our goal is to collect instances where the disambiguating evidence is grounded in the image, and to provide human-verified reference translations. VIDA is built on two existing resources: VIDA-Base and VIDA-CollN are curated from 3AM (Ma et al., 2024), and VIDA-Sent is adapted from MMA (Wang et al., 2024a).

3.1 Stage 1: Data Preprocessing and Filtering

The goal of this stage is to curate image–text aligned and textually clean source captions with visually-dependent ambiguities. We use GPT-4o (Hurst et al., 2024) to filter mismatched image–text pairs and normalize text. We then apply a dual-

model consensus strategy with two state-of-the-art commercial LLMs, Qwen-Max (Qwen Team, 2024) and DeepSeek-v3 (Liu et al., 2024): each model independently assesses whether the caption remains ambiguous without the image. We retain only captions that both models judge as ambiguous, and record an *Ambiguous Caption* and an *Ambiguity Rationale* that points to the ambiguous span and the disambiguation cue.

3.2 Stage 2: Disambiguated Translation

In this stage, we produce high-quality disambiguated translations for each retained caption. For each caption from Stage 1, we generate a disambiguated translation using GPT-4o with a structured input that includes the *Ambiguous Caption*, the paired *Image*, and the *Ambiguity Rationale* from the previous stage. The model outputs both a *Disambiguated Translation* that resolves the ambiguity and a *Resolution Rationale* explaining how visual information is used to support the decision.

3.3 Stage 3: Quality Assurance and Validation

This stage finalizes the dataset with an LLM-as-a-judge quality check followed by human verification, ensuring both translation quality and correct visual disambiguation. We employ Qwen-Max to evaluate each disambiguated translation along two dimensions: semantic preservation and fluency. For each dimension, Qwen-Max outputs a score from 1 to 5 together with a brief justification. We flag cases with scores below 4 as potentially problematic and prioritize them for closer inspection.

To ensure translation correctness, we further conduct human verification with two native Chinese speakers (one Ph.D. and one M.S. in Computer Science). For each instance, annotators are shown the image, the source caption, the ambiguity rationale, the candidate translation, and Qwen-Max's judgment. Each annotator independently reviews the translation using three criteria: (i) whether the annotated ambiguity is correctly resolved, (ii) whether the translation is fluent, and (iii) whether the meaning of the translation is preserved. If the candidate fails any criterion, the annotator provides a corrected translation based on the image and the ambiguity rationale. During verification, we identified collective noun cases where underspecified terms must be concretized based on the image content and group these cases into the **Collective Noun Subset**. Finally, **VIDA** comprises 2,500 instances across three subsets (**VIDA-Base**, **VIDA-Sent**, and **VIDA-CollN**). See Appendix C for details.

4 Evaluation Metrics for Disambiguation

In MMT disambiguation, the objective of evaluation is to measure disambiguation accuracy rather than overall translation quality. We use an LLM-as-a-judge to classify span-level correctness and compute accuracy from binary outputs. Specifically, we fine-tune Qwen3-8B (Yang et al., 2025) on **VIDA** to determine whether annotated ambiguous expressions are correctly resolved in translation. We train the classifier in a contrastive setting. Gold disambiguated translations are used as positive samples. Negative samples are candidate translations from our curation pipeline that fail to resolve the annotated ambiguity. Building on this, we introduce two complementary **Disambiguation-Centric Metrics** for a comprehensive evaluation of disambiguation performance: **Disambi-Term** measures term-level disambiguation accuracy by evaluating each annotated ambiguous term in the entire dataset. **Disambi-Inst.** reports instance-level accuracy, counting a sentence as correct only if all ambiguous expressions within it are correctly resolved. Judge reliability and training details are provided in Appendix D.

5 Experiments

5.1 Experimental Settings

Dataset We conduct experiments on the VIDA dataset. VIDA-Base is split 7:3 into 1,352 training and 580 test samples. VIDA-Sent and VIDA-CollN are held out from training as out-of-distribution (OOD) test sets due to limited size, which also allows evaluation on unseen ambiguity types.

Metrics We report standard MT metrics (*BLEU*, *chrF*/*chrF++* (Popović, 2015, 2017), *METEOR* (Banerjee and Lavie, 2005), *TER* (Snover et al., 2006), *BERT-F1* (Devlin et al., 2019), and *COMET*) to assess overall translation quality, together with Disambi-Term and Disambi-Inst. for disambiguation accuracy. To assess the reliability of gains on the small OOD subsets, we estimate 95% CIs using paired bootstrap for Disambi-Inst. and instance-level cluster bootstrap for Disambi-Term, and test instance-level significance with exact McNemar tests (Appendix E).

Models and Baselines We evaluate **Qwen2.5-VL-7B** (Bai et al., 2025) and **InternVL3-8B** (Zhu

Model	Dataset	Model Setting	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
InternVL3-8B	All-Test	Vanilla	48.04	41.95	32.98	**40.29**	86.63	58.47	84.49	50.86	39.81
		SFT	**49.20**	**42.73**	32.82	43.44	87.07	**58.95**	85.55	54.36	43.77
		CoT-SFT	47.64	41.16	**32.99**	41.61	**87.18**	58.78	**85.88**	**58.45**	**48.78**
	VIDA-Base-Test	Vanilla	53.51	46.76	36.87	35.66	88.84	65.24	86.08	60.18	46.55
		SFT	**55.31**	**48.31**	**37.13**	**34.28**	**89.61**	**66.55**	**87.30**	62.67	50.17
		CoT-SFT	51.10	44.41	35.75	38.06	88.56	63.25	86.44	**64.89**	**51.38**
	VIDA-Sent	Vanilla	42.51	36.85	33.01	**44.76**	84.31	52.54	84.21	50.00	50.00
		SFT	36.99	35.69	31.96	67.93	83.93	51.67	84.70	55.45	55.45
		CoT-SFT	**44.22**	**38.19**	**34.28**	45.27	**85.70**	**55.32**	**86.39**	**58.97**	**58.97**
	VIDA-CollN	Vanilla	36.56	31.66	27.14	49.24	84.63	48.79	81.39	18.36	12.16
		SFT	37.97	32.92	**28.22**	49.11	85.11	50.56	82.60	22.62	14.90
		CoT-SFT	**39.26**	**33.89**	25.70	**48.52**	**85.70**	**51.05**	**83.96**	**38.36**	**32.55**
Qwen2.5-VL-7B	All-Test	Vanilla	47.85	41.67	34.12	42.74	86.56	58.88	84.83	50.08	39.81
		SFT	**49.13**	**42.85**	**34.58**	**40.34**	**87.38**	**59.51**	85.82	52.81	42.46
		CoT-SFT	47.59	41.39	33.26	42.49	87.06	58.60	**85.84**	**55.51**	**46.08**
	VIDA-Base-Test	Vanilla	52.38	45.66	37.66	37.64	88.53	64.45	86.30	58.49	46.38
		SFT	**53.88**	**47.09**	**38.35**	**36.17**	**89.11**	**65.48**	**87.07**	**61.42**	**49.31**
		CoT-SFT	50.41	44.57	36.00	39.44	88.32	62.75	86.35	60.71	46.90
	VIDA-Sent	Vanilla	44.46	38.92	34.97	50.95	84.21	54.78	84.41	51.28	51.28
		SFT	45.12	39.52	**35.59**	**42.66**	**85.86**	**55.41**	86.06	52.56	52.56
		CoT-SFT	**45.54**	**39.79**	35.43	45.17	85.66	55.02	**86.41**	**60.26**	**60.26**
	VIDA-CollN	Vanilla	38.06	32.83	24.63	50.54	84.87	51.16	82.06	19.02	12.16
		SFT	**39.02**	**33.71**	**25.28**	**49.24**	85.30	50.96	82.69	21.31	14.51
		CoT-SFT	38.21	32.49	24.51	50.71	**85.32**	**51.34**	**83.39**	**33.77**	**27.45**

Table 1: Performance comparison of **InternVL3-8B** and **Qwen2.5-VL-7B** under Vanilla, SFT, and CoT-SFT settings. **Box** highlights best standard metrics, **Box** highlights best disambiguation metrics.

et al., 2025) under three settings: vanilla inference, supervised fine-tuning (SFT), and chain-of-thought supervised fine-tuning (CoT-SFT) with manually synthesized ambiguity-resolution traces (Appendix G). We additionally evaluate Qwen3.5-27B and Gemma-4-26B-A4B (Google DeepMind, 2026) under vanilla inference and extend the SFT/CoT-SFT comparison to InternVL3-14B to examine CoT-SFT's effectiveness at a larger model scale.

5.2 Experimental Results

Analysis on In-Domain Dataset. Evaluation on the in-domain test set (**VIDA-Base-Test**) in Table 1 examines how well the models fit the training distribution (VIDA-Base-Train). The SFT setting achieves the strongest overall translation quality for both models across most standard translation metrics. CoT-SFT remains competitive on semantic metrics (e.g., COMET) for both models, but can slightly underperform SFT on surface-overlap metrics such as BLEU, likely because ambiguity resolution often requires paraphrasing or structural reformulation that deviates from reference wording. To further assess whether these surface-metric gaps indicate an overall quality loss, we use an LVLM judge (Qwen3.6-Plus; Qwen Team, 2026b) and find CoT-SFT comparable to SFT (Appendix B). Under Disambiguation-Centric evaluation, CoT-SFT yields the best performance for InternVL3-8B (Disambi-Term/Inst. **64.89/51.38**), surpassing both SFT and Vanilla. In contrast, for Qwen2.5-VL-7B, CoT-SFT improves over Vanilla but remains slightly below SFT (60.71/46.90 vs. 61.42/49.31), suggesting that in-domain gains from reasoning supervision are model-dependent. We attribute part of the CoT-SFT disambiguation gap to overthinking behaviors and provide quantitative and qualitative analyses in Appendix H.

Analysis on Out-of-Distribution Dataset. Evaluation on the out-of-distribution (OOD) datasets (**VIDA-Sent** and **VIDA-CollN**) examines whether models can generalize to unseen ambiguity types. Compared to in-domain results, CoT-SFT shows clearer advantages under distribution shift. On standard translation metrics, CoT-SFT achieves the highest COMET on both OOD subsets for both models, indicating stronger semantic adequacy. Disambiguation-centric metrics make the advantage more explicit. On VIDA-Sent, CoT-SFT achieves consistent gains over SFT (approximately **+3.5** points for InternVL3-8B and **+7.7** points for Qwen2.5-VL-7B on both Disambi-Term and Disambi-Inst.). The gains are substantially larger on VIDA-CollN (over **+15** points for InternVL3-8B and **+12** points for Qwen2.5-VL-7B), suggesting that CoT-SFT demonstrates stronger generalization beyond the training distribution.

Subset	Δ Inst. [95% CI]	McNemar p	Δ Term [95% CI]
Qwen2.5-VL-7B			
VIDA-Sent	**+7.69 [3.21, 12.18]**	0.0012	**+7.69 [3.21, 12.18]**
VIDA-CollN	**+12.94 [7.84, 18.04]**	<0.00001	**+12.46 [7.33, 17.78]**
InternVL3-8B			
VIDA-Sent	+3.53 [−0.64, 8.01]	0.1352	+3.53 [−0.64, 8.01]
VIDA-CollN	**+17.65 [12.16, 23.14]**	<0.00001	**+15.74 [10.36, 21.38]**

Table 2: CoT-SFT vs. SFT on the OOD subsets. Δ: CoT-SFT minus SFT; p: exact McNemar test. Bold: 95% CI excludes zero.

Model	Setting	BLEU	COMET	Disambi-Term	Disambi-Inst.
Qwen3.5-27B	Vanilla	48.14	86.38	57.25	46.47
Gemma-4-26B-A4B	Vanilla	45.82	85.53	54.86	44.46
InternVL3-14B	Vanilla	49.93	84.90	50.28	39.67
	SFT	**51.23**	85.75	52.97	42.63
	CoT-SFT	49.55	**86.20**	**58.47**	**50.04**

Table 3: All-Test results of stronger LVLMs (top) and InternVL3-14B (bottom). Highlighting follows Table 1; full results are in Appendix F.

Statistical Significance of OOD Gains. Because the OOD subsets are small, we further test whether these gains are statistically reliable (Table 2). Three of the four instance-level gains are significant at $\alpha = 0.05$ and remain significant after Bonferroni correction ($\alpha = 0.0125$): Qwen2.5-VL-7B on VIDA-Sent and both models on VIDA-CollN, and the term-level CIs lead to the same conclusion. The gain for InternVL3-8B on VIDA-Sent is positive but not significant ($p = 0.1352$). These results indicate that the benefit of CoT-SFT on VIDA-CollN is significant for both models and consistent at the term and instance levels. In contrast, the benefit on VIDA-Sent is model-dependent, possibly because sentence-level ambiguities such as idioms and pragmatic interpretations differ more from the word-level training instances, so the learned reasoning patterns transfer less uniformly across models.

Analysis on All-Test Dataset. All-Test pools all subsets into a single evaluation rather than averaging per-subset scores, yielding an overall score weighted by subset size. On standard translation metrics, SFT generally improves over Vanilla for both models, indicating better overall translation quality on the mixed test distribution. In contrast, CoT-SFT consistently yields higher semantic adequacy and disambiguation accuracy than SFT: it achieves the highest COMET and improves Disambi-Term/Inst. by about **+4.1/+5.0** for InternVL3-8B and **+2.7/+3.6** for Qwen2.5-VL-7B. Overall, CoT-SFT trades a small loss in surface overlap for higher semantic adequacy and disambiguation accuracy, with most of the disambiguation gain coming from the OOD subsets.

5.3 Stronger LVLMs and Larger-Scale CoT-SFT

Evaluation with Stronger LVLMs. As shown in Table 3, Qwen3.5-27B and Gemma-4-26B-A4B achieve higher disambiguation accuracy than the vanilla 7B/8B models in Table 1 (up to 57.25/46.47 vs. 50.86/39.81 Disambi-Term/Inst.). Nevertheless, fewer than half of the All-Test instances have all annotated ambiguities correctly resolved, indicating that visual disambiguation on VIDA is not solved by stronger recent LVLMs. Notably, InternVL3-8B with CoT-SFT (48.78 Disambi-Inst.) already exceeds both, although it is trained on VIDA-Base. The subset-level results (Appendix F) show that the remaining difficulty is most pronounced on VIDA-CollN, where both models reach at most 16.02 Disambi-Inst., close to the vanilla 7B/8B models (12.16) and far below CoT-SFT (27.45–32.55).

Larger-Scale CoT-SFT Evaluation. On InternVL3-14B, CoT-SFT improves Disambi-Term/Inst. by 5.50/7.41 points over SFT and achieves the highest COMET, while SFT retains the highest BLEU. This reproduces the pattern observed at the 7B/8B scale: SFT better fits the surface form of the references, whereas CoT-SFT improves semantic adequacy and disambiguation. As at the smaller scale, the gains come mainly from the OOD subsets (+9.30 and +20.00 Disambi-Inst. on VIDA-Sent and VIDA-CollN; Appendix F), indicating that the benefit of CoT-SFT extends beyond the 7B/8B scale.

6 Conclusion

We introduce VIDA, an MMT dataset for visually-dependent word- and sentence-level ambiguities, and propose Disambiguation-Centric Metrics that directly measure disambiguation accuracy based on an LLM-as-a-judge classifier. Evaluations with stronger recent LVLMs show that VIDA remains challenging, especially for collective nouns. Our comparison of SFT and CoT-SFT shows that SFT better fits the surface patterns of the training references, whereas CoT-SFT improves visually grounded disambiguation. The gains are robust on VIDA-CollN across both models, while the effect on VIDA-Sent is model-dependent; the disambiguation benefit also holds for InternVL3-14B.

Limitations

Although **VIDA** targets highly visually dependent ambiguities, the dataset remains modest in scale compared to large-scale translation corpora and is currently limited to English–Chinese translation. In addition, the reasoning traces used to supervise CoT-SFT are synthetically constructed based on manually designed patterns, which may not reflect natural model reasoning. Although we include stronger vanilla baselines and a 14B experiment, the CoT-SFT evaluation still covers a limited number of open-weight model families. Future work includes expanding the dataset to additional language pairs, validating the CoT-SFT findings on a broader range of LVLMs, and moving beyond fixed templates toward dynamic or adaptive reasoning generation, for instance through reinforcement learning or preference-based optimization methods that let the model adapt its reasoning depth to the difficulty of the ambiguity.

Ethics Statement

The proposed **VIDA** dataset is curated from publicly available data sources and inherits their research-use license; the images do not introduce additional personally identifying information beyond the underlying datasets. Annotators were informed of the task goals before annotation. The dataset is publicly released with usage guidelines to support research on multimodal machine translation and ambiguity resolution.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. 2025. Qwen look again: Guiding vision-language reasoning models to re-attention visual information. *arXiv preprint arXiv:2505.23558*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

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, Brussels, Belgium. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Google DeepMind. 2026. Gemma 4 model card.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Zheng Cui. 2022. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2862–2872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Danyang Liu, Fanjie Kong, Xiaohang Sun, Dhruva Patil, Avijit Vajpayee, Zhu Liu, Vimal Bhat, and Najmeh

Sadoughi. 2025. Detect, disambiguate, and translate: On-demand visual reasoning for multimodal machine translation with large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1559–1570, Albuquerque, New Mexico. Association for Computational Linguistics.

Xinyu Ma, Xuebo Liu, Derek F. Wong, Jun Rao, Bei Li, Liang Ding, Lidia S. Chao, Dacheng Tao, and Min Zhang. 2024. 3AM: An ambiguity-aware multimodal machine translation dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1–13, Torino, Italia. ELRA and ICCL.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

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

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents.

Qwen Team. 2026b. Qwen3.6-Plus: Towards real world agents.

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

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ru Wang, Selena Song, Liang Ding, Shixiang Shane Gu, Mingming Gong, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. 2024a. MMA: Benchmarking multimodal large language model in ambiguity contexts.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 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, Online. Association for Computational Linguistics.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 37:92012–92035.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Positioning Against Prior Resources

Existing resources address visual disambiguation from different angles, but they differ substantially from VIDA in task format and validation target. MLT focuses on localized lexical translation, where the model resolves an ambiguous source word with sentence and image context rather than generating a full translation. AmbigCaps and M^3-AmbigCaps use open-ended caption translation, but their ambiguity coverage is concentrated on gender or pronoun choices derived from back-translation. CoMMuTE provides a controlled contrastive setting by

Resource	Generation format	Ambiguity focus	Ambiguity evaluation
MLT (Lala and Specia, 2018)	localized lexical translation	lexical choice	rule-based lexical matching
AmbigCaps / M^3-AmbigCaps (Li et al., 2021; Guo et al., 2022)	open-ended caption translation	primarily gender/pronoun ambiguity	rule-based matching over lexical variants
CoMMuTE (Futeral et al., 2023)	contrastive candidate ranking	lexical/gender ambiguity	contrastive accuracy over predefined candidates
3AM (Ma et al., 2024)	open-ended MMT generation	mainly word-level ambiguity	general MT and ambiguity-oriented evaluation
MMA (Wang et al., 2024a)	VQA-style disambiguation	sentence-level ambiguity	VQA-style answer accuracy
VIDA	open-ended MMT generation	lexical, sentence, pragmatic, collective noun	LLM-judge semantic span accuracy

Table 4: Positioning of VIDA relative to prior multimodal disambiguation resources.

asking models to rank predefined translation candidates, which is useful for isolating visual cues but does not evaluate free-form generation. 3AM targets English–Chinese MMT but primarily focuses on word-level ambiguity and includes many instances resolvable from text alone. MMA broadens the phenomenon to sentence-level ambiguity, but is formulated as a VQA task rather than translation. In contrast, VIDA targets open-ended MMT generation, covers word-, sentence-, pragmatic-, and collective-noun ambiguity, and retains only instances whose intended interpretation is verified to require visual evidence.

The evaluation protocols also differ in what they can verify. Contrastive accuracy measures whether a model ranks a predefined correct candidate above distractors, while rule-based matching over lexical variants is tied to expected surface forms. These designs are less suitable for open-ended translation, where correct disambiguation may be expressed through paraphrases or different lexical choices. General MT metrics such as BLEU and COMET remain useful for overall translation quality, but they do not directly verify whether an annotated ambiguous span is resolved correctly, since surface-overlap metrics may penalize valid paraphrases or lexical variation and sentence-level metrics are too coarse-grained for span-level disambiguation. VIDA therefore uses Disambiguation-Centric Metrics to judge span-level semantic resolution while allowing valid paraphrases. Table 4 summarizes the comparison.

Concrete Examples. Table 5 illustrates these differences with examples from each resource. In 3AM, some captions can be translated correctly without the image, while others are paired with an unrelated image. CoMMuTE asks the model to rank two predefined translations rather than generate one, and MMA evaluates visual question answering rather than translation. In contrast, VIDA manually ensures both visual dependency and image–text consistency for every retained instance, requires the model to generate a complete translation, and evaluates whether the image-supported sense is expressed (see Figure 3 for VIDA examples).

B Auxiliary Overall Quality Evaluation

To complement standard MT metrics, we evaluate overall translation quality with Qwen3.6-Plus (Qwen Team, 2026b), a commercial LVLM, as the judge on a 0–100 scale. Table 6 shows that SFT and CoT-SFT are close in overall quality, and CoT-SFT is often comparable or higher despite stronger disambiguation performance. This supports our main interpretation that CoT-SFT's gains reflect improved ambiguity resolution rather than a severe loss in general translation quality.

C Dataset Statistics

C.1 Definition and Detection of Ambiguous Terms

We define an *ambiguous term* as a word or phrase whose translation admits multiple valid candidates when given only the pure text context. Crucially, its true meaning can only be uniquely determined by referencing the paired visual evidence. Such terms are detected during Stage 1 of our curation pipeline (section 3): Qwen-Max and DeepSeek-v3 independently evaluate the text-only source for translation uncertainty, and an instance is retained only when both models reach a consensus that ambiguity exists. The two models then extract the specific ambiguous span and produce an *Ambiguity Rationale* that describes the alternative interpretations and identifies the visual cue required for disambiguation. Concrete examples are illustrated in Figure 3.

C.2 Pipeline Filtering Statistics

The large reduction from the raw candidate pool to the final benchmark reflects the strict quality control in our curation pipeline. We begin with

Resource	Example	Limitation for evaluating visual disambiguation
3AM	"A lot of rolls of toilet paper."	Correctly translatable from text alone; no visual evidence is needed.
3AM	"A praying mantis perched on a branch at reserve."	Paired with an unrelated image (image–text mismatch).
CoMMuTE	"We'll have to get rid of that mole." with two predefined French translations: "Il va falloir enlever ce grain de beauté" (beauty mark) and "Il va falloir se débarrasser de cette taupe" (animal).	The model ranks the two given translations by probability instead of generating a translation.
MMA	Caption "Crying" with an image and a multiple-choice question about why the person is crying (gold answer: "Feeling sad").	Evaluates visual question answering; the caption is not a complete translation input.

Table 5: Examples illustrating the limitations of prior resources for evaluating visual disambiguation in open-ended MMT.

Model	Dataset	Vanilla	SFT	CoT-SFT
InternVL3-8B	All-Test	83.82	84.76	**86.35**
	VIDA-Base-Test	86.06	87.24	**87.25**
	VIDA-Sent	79.31	78.86	**83.57**
	VIDA-CollN	84.26	86.34	**87.70**
Qwen2.5-VL-7B	All-Test	83.92	85.70	**85.87**
	VIDA-Base-Test	86.07	**87.33**	**87.33**
	VIDA-Sent	79.10	82.24	**82.48**
	VIDA-CollN	84.91	86.24	**86.69**

Table 6: Auxiliary overall translation quality judged by Qwen3.6-Plus. Scores are on a 0–100 scale and supplement the main metrics in Table 1.

26,452 raw image–text pairs, retain 14,993 after image–text matching and text normalization, and then keep only 2,500 instances after ambiguity detection and dual-LLM verification. The two models show strong bidirectional agreement on the ambiguity decision: 94.38% of ambiguous candidates flagged by Qwen-Max are independently confirmed by DeepSeek-v3, and 87.50% in the reverse direction, indicating that the detected ambiguities are robust rather than model-specific artifacts. Subsequent stages refine translations and annotations without discarding additional instances. All 2,500 retained instances are then human-verified by two native Chinese annotators, with substantial inter-annotator agreement across the three review criteria (average Cohen's $\kappa = 0.8024$; Table 8). The filtering funnel is summarized in Table 7.

Stage	Remaining	Filtered Out
Initial raw pairs	26,452	–
After image–text filtering	14,993	11,459
After ambiguity verification	2,500	12,493
Human verification coverage	2,500 (100%)	0

Table 7: Filtering funnel for constructing VIDA.

Criterion	Cohen's κ
Disambiguation	0.8033
Fluency	0.7938
Preservation	0.8102
Average	**0.8024**

Table 8: Inter-annotator agreement for human verification.

C.3 Human Verification Protocol

LLMs are used to screen the source pairs and to draft disambiguated translations, because processing all candidates manually from scratch would be prohibitively costly. The dual-LLM agreement in Stage 1 reduces reliance on a single model, but it does not determine the final ground truth. In Stage 3, the two annotators do not merely accept or reject LLM outputs; they actively rewrite translations that fail any of the three criteria. When one annotator corrects a translation, the other annotator re-verifies the correction. Disagreements are first discussed between the two annotators, and cases that remain unresolved are referred to an additional native Chinese speaker for adjudication.

C.4 LLM Configurations

We use GPT-4o (`gpt-4o-2024-11-20`) with temperature 1.0 and top_p 1.0. Qwen-Max (`qwen-max-2025-01-25`) and DeepSeek-V3 are accessed through DashScope with the default decoding parameters (temperature/top_p of 0.7/0.8 for Qwen-Max and 0.7/0.6 for DeepSeek-V3); no dated snapshot is available for DeepSeek-V3. The maximum number of output tokens is not set for any of these models.

C.5 VIDA Subset Statistics

The rigorous pipeline outlined in section 3 results in the construction of **VIDA** (**Vi**sually-**D**ependent **A**mbiguity), comprising 2,500 visually dependent instances across three subsets: **VIDA-Base** contains 1,932 samples curated from 3AM and focuses primarily on word-level ambiguities; **VIDA-CollN**

Model	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
InternVL3-8B vs. Qwen2.5-7B									
LVLM	48.04	41.95	32.98	40.29	86.63	58.47	84.49	50.86	39.81
↑ *vision* ↑	*6.88*	*6.48*	*2.26*	*-5.83*	*2.03*	*6.63*	*2.51*	**8.03**	**8.54**
LLM	41.16	35.47	30.72	46.12	84.60	51.84	81.98	42.83	31.27
Qwen2.5-VL-7B vs. Qwen2.5-7B									
LVLM	47.85	41.67	34.12	42.74	86.56	58.88	84.83	50.08	39.81
↑ *vision* ↑	*4.27*	*3.59*	*0.89*	*-1.76*	*1.76*	*5.97*	*1.73*	**7.25**	**12.18**
LLM	41.16	35.47	30.72	46.12	84.60	51.84	81.98	42.83	31.27

Table 9: Ablation results without images on **All-Test**. The middle row reports the gain from adding the correct visual input over the text-only baseline.

Model	Image Setting	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
InternVL3-8B	Correct image	48.04	41.95	32.98	40.29	86.63	58.47	84.49	50.86	39.81
	Random image	44.51	38.79	32.12	41.89	85.53	54.67	82.49	42.39	30.51
Qwen2.5-VL-7B	Correct image	47.85	41.67	34.12	42.74	86.56	58.88	84.83	50.08	39.81
	Random image	43.18	37.71	32.30	43.82	85.18	53.83	82.26	42.81	32.08

Table 10: Ablation results with random images on **All-Test**. Replacing the paired image with a random one substantially degrades both standard MT metrics and disambiguation-centric metrics.

contains 256 collective-noun cases, also curated from 3AM, where visual context concretizes an otherwise abstract group; and **VIDA-Sent** contains 312 samples adapted from MMA with sentence-level semantic ambiguities. A complete statistical summary is provided in Table 11.

Subset	Ambiguity	Size	Avg. Len.	Avg. Ambi.
VIDA-Base	Word-level	1,932	11.12	1.78
VIDA-Sent	Sentence-level	312	6.00	1.00
VIDA-CollN	Word-level	256	10.08	1.20

Table 11: Statistical summary of VIDA subsets.

Table 12 further reports the distribution of ambiguity types in each subset. The types follow the four categories used in Stage 1 ambiguity detection (Prompt 1 in Appendix I) and describe the source of ambiguity: *lexical* (multiple senses of a word or phrase), *syntactic* (multiple possible structures), *pragmatic* (context-dependent interpretations), and *cultural/background* (reliance on external cultural or background knowledge). Since an instance can involve more than one type, the labels are multi-label and their counts may exceed the number of instances.

Subset	Size	Lexical	Syntactic	Pragmatic	Cultural
VIDA-Base	1,932	1,309	755	858	163
VIDA-Sent	312	77	64	145	26
VIDA-CollN	256	253	5	14	2

Table 12: Distribution of ambiguity types in VIDA subsets (multi-label).

C.6 Visual Dependence

To further verify that VIDA instances require the correct visual context, we evaluate models under two ablated settings on **All-Test**: (i) **Text-only**, where the paired image is removed and replaced by the corresponding text-only backbone, and (ii) **Random Image**, where the correct image is replaced with a randomly sampled image from the test set. Results are shown in Table 9 and Table 10. In both cases, performance drops consistently on both standard MT metrics and the proposed disambiguation-centric metrics. These degradations clearly demonstrate the visual dependence of VIDA: the ambiguous text in our dataset must rely on the correct image information to be successfully disambiguated, and performance drops markedly when the model is given no image or a random image.

We further repeat both ablations with the stronger LVLMs from Section 5.3. For these models, the text-only setting removes the image from the same model. As shown in Table 13, removing or replacing the image reduces All-Test Disambi-Inst. by 8.3–10.4 points for both models. The drop is particularly large on VIDA-Sent, e.g., from 57.05 to 35.58 Disambi-Inst. when the image is removed for Qwen3.5-27B. This confirms that the visual dependence of VIDA is not specific to smaller models.

Model	Image	COMET	Disambi-Term	Disambi-Inst.
Qwen3.5-27B	Correct	86.38	57.25	46.47
	None	84.61	49.60	38.10
	Random	84.61	47.95	36.09
Gemma-4-26B-A4B	Correct	85.53	54.86	44.46
	None	83.96	48.38	36.18
	Random	83.57	46.48	34.35

Table 13: Visual-dependence ablation with stronger LVLMs on **All-Test**. *None* removes the image from the same model; *Random* replaces it with a randomly sampled image.

D Judge Reliability and Training Details

D.1 Judge Reliability

We further validate the reliability of the Judge from two complementary perspectives. First, we perform **Cross-Model Judge Verification** by re-evaluating the exact same translation outputs from Qwen2.5-VL-7B (CoT-SFT) on **All-Test** with a second Judge based on LLaMA3.1-8B (Grattafiori et al., 2024). As shown in Table 14, the scores remain highly consistent across judge backbones. Second, we measure **Human-Judge Alignment** on a randomly sampled 20% subset of VIDA (500 instances) and compute Cohen's Kappa between expert human judgments and the Qwen-based Judge. Table 15 shows strong agreement across all subsets, indicating that the Judge captures semantic disambiguation resolution rather than model-specific preferences.

Judge	Disambi-Term	Disambi-Inst.
Qwen-Judge	55.51	46.08
LLaMA-Judge	55.72	47.69

Table 14: Cross-model judge verification on the same Qwen2.5-VL-7B (CoT-SFT) predictions.

Dataset	Cohen's Kappa
VIDA-Base-Test	0.8828
VIDA-Sent	0.8932
VIDA-CollN	0.9023
Average	**0.8928**

Table 15: Human-Judge alignment measured by Cohen's Kappa on a 500-instance sample.

D.2 Judge Training Details

The Judge is fine-tuned from Qwen3-8B as a binary classifier for span-level disambiguation verification. Each training instance contains the source caption, a candidate translation, the annotated ambiguous span, and the gold interpretation of that span; the model predicts whether the ambiguity is correctly resolved in the translation. Gold disambiguated translations are used as positive examples. Negative examples are drawn from candidate translations produced during the curation pipeline that fail to resolve the annotated ambiguity. To reduce false negatives, we further filter out semantically equivalent candidates with an auxiliary LLM-based screening step before training.

We fine-tune the Judge for 10 epochs with learning rate 1×10^{-5} and batch size 8. LoRA is used with rank 8, alpha 32, and dropout 0.1. The judge data contain 1,499 positive and 1,497 negative examples, which are sampled at a positive-to-negative ratio of 2:1 during training. We use a random seed of 42. This contrastive setup encourages the Judge to focus on disambiguation correctness rather than superficial lexical overlap, which is crucial for handling valid paraphrases and lexical variation in open-ended translation. All experiments were conducted on one NVIDIA A100 GPU.

E Statistical Testing Details

The OOD subsets are relatively small (312 instances in VIDA-Sent and 256 in VIDA-CollN), so differences in disambiguation accuracy between training settings may partly reflect sampling variation. We therefore test whether the improvements of CoT-SFT over SFT on these subsets are statistically reliable. Since both settings are evaluated on the same instances, we use paired analyses.

Disambi-Inst. Under each setting, each instance yields a binary correct/incorrect outcome, so the two settings form paired binary observations. We estimate the 95% confidence interval of the accuracy difference with a paired bootstrap over instances (10,000 resamples), which resamples each instance together with both of its outcomes and makes no distributional assumption about the difference. To test significance, we use the McNemar test, which is designed for paired binary outcomes: instances on which both settings are correct or both are incorrect carry no information about their difference, so the test relies only on the discordant instances where exactly one setting is correct. We use its exact (binomial) version because the number of discordant instances on the small OOD subsets is limited, in which case the chi-squared approximation can be inaccurate.

Disambi-Term. The annotated terms within the same instance share the same source sentence, im-

Model	Dataset	Setting	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
Qwen3.5-27B	All-Test	Vanilla	48.14	41.69	33.95	40.70	87.51	58.88	86.38	57.25	46.47
	VIDA-Base-Test	Vanilla	50.84	44.85	36.89	37.60	89.01	63.32	87.32	68.53	55.79
	VIDA-Sent	Vanilla	48.66	43.14	38.95	39.53	86.66	57.39	87.32	57.05	57.05
	VIDA-CollN	Vanilla	37.95	32.69	24.57	49.39	85.16	50.67	83.11	19.93	12.50
Gemma-4-26B-A4B	All-Test	Vanilla	45.82	39.74	34.07	42.39	86.86	57.31	85.53	54.86	44.46
	VIDA-Base-Test	Vanilla	48.76	43.20	37.88	38.97	88.44	61.98	86.63	65.19	52.33
	VIDA-Sent	Vanilla	44.37	38.94	34.88	44.38	85.49	55.28	85.72	53.21	53.21
	VIDA-CollN	Vanilla	36.49	31.32	23.49	49.61	84.95	49.23	82.84	22.22	16.02
InternVL3-14B	All-Test	Vanilla	49.93	43.84	37.50	38.12	87.18	59.56	84.90	50.28	39.67
		SFT	**51.23**	**45.08**	**38.60**	**36.95**	87.67	60.48	85.75	52.97	42.63
		CoT-SFT	49.55	43.19	34.95	39.34	**87.69**	**60.76**	**86.20**	**58.47**	**50.04**
	VIDA-Base-Test	Vanilla	55.53	48.83	42.55	33.28	89.38	66.06	86.35	60.71	47.59
		SFT	**57.25**	**50.51**	**44.06**	**31.72**	**90.00**	**67.39**	**87.27**	63.36	51.03
		CoT-SFT	53.31	46.47	37.89	36.04	88.94	64.98	86.42	**63.75**	**51.90**
	VIDA-Sent	Vanilla	44.93	39.57	35.43	42.25	84.72	53.88	84.79	47.44	47.44
		SFT	45.08	39.49	35.55	**41.84**	84.99	54.22	85.50	51.28	51.28
		CoT-SFT	**46.39**	**40.51**	**36.28**	42.28	**86.46**	**57.95**	**87.12**	**60.58**	**60.58**
	VIDA-CollN	Vanilla	38.78	33.55	25.16	47.44	85.18	51.73	81.74	18.36	12.16
		SFT	39.76	34.44	25.83	46.68	85.62	52.42	82.61	20.00	12.94
		CoT-SFT	**41.99**	**36.36**	**27.59**	**45.62**	**86.36**	**54.58**	**84.59**	**38.69**	**32.94**

Table 16: Full results of stronger recent LVLMs under vanilla inference (top) and of **InternVL3-14B** under Vanilla, SFT, and CoT-SFT settings (bottom). In the bottom block, **Box** highlights the best standard metrics and **Box** highlights the best disambiguation metrics within each dataset.

age, and model output, and are therefore not independent. To account for this dependence, we use an instance-level cluster bootstrap that resamples instances together with all of their annotated terms.

Multiple Comparisons. Because we perform four comparisons (two models $\times$ two OOD subsets), we additionally apply a Bonferroni correction, giving an adjusted significance threshold of $\alpha = 0.05/4 = 0.0125$.

Notes on Table 2. All differences Δ in Table 2 are computed from unrounded scores and may therefore differ slightly from the differences of the rounded values in Table 1. Since each VIDA-Sent instance contains exactly one annotated span, its term- and instance-level results coincide.

F Full Results of Stronger LVLMs and InternVL3-14B

Table 16 reports subset-level results for the models in Section 5.3. At the subset level, the stronger LVLMs improve over the vanilla 7B/8B models mainly on VIDA-Base-Test and VIDA-Sent, whereas VIDA-CollN remains difficult (12.50 and 16.02 Disambi-Inst.). For InternVL3-14B, CoT-SFT is comparable to SFT on the in-domain VIDA-Base-Test (51.90 vs. 51.03 Disambi-Inst.), while yielding large gains on VIDA-Sent (60.58 vs. 51.28) and VIDA-CollN (32.94 vs. 12.94).

G Chain-of-Thought Supervised Fine-Tuning

We define a task-specific six-step reasoning template that guides models to articulate the alignment between ambiguous expressions and visual evidence. Each synthetic trace is constructed according to the following standardized reasoning template:

1. **Visual Grounding**: Examine the image carefully and identify the visual elements that correspond to key words or phrases in the source sentence. Describe how these elements connect to the text.

2. **Initial Translation**: Generate a preliminary translation based on both the text and the grounded visual evidence.

3. **Ambiguity Check**: Review the initial translation and highlight any terms that remain ambiguous—those whose meanings are unclear or context-dependent when relying on text alone.

4. **Visual Disambiguation**: This step is critical. While visual grounding establishes a mapping between the image and the text, the initial translation can still leave some ambiguities unresolved. The model explicitly revisits the image, not only to strengthen the connection between ambiguous terms and their corresponding visual evidence, but also to refresh its ac-



SRC: He watched the workers on the bridge.

REF: 他在桥上观察工人 (He is on the bridge.)

Ambi. Rationale: *"on the bridge"* is ambiguous, which could either describe the workers' location or the man's

Disambi. Rationale: The image clearly shows that it is "He" on the bridge, not workers.

CoT Six-step Reasoning

Think: **VISUAL GROUNDING** In the image, a man in a suit stands on a raised platform or bridge overlooking a construction site below. Many workers wearing hard hats are working in an open trench, and the man is visually above them from his vantage point on that bridge. **INITIAL TRANSLATION** "他看着桥上的工人。 **AMBIGUITY CHECK** The phrase "on the bridge" can be ambiguous. It could mean: \n• He is on the bridge watching the workers • The workers are on the bridge, and he is watching them. **VISUAL DISAMBIGUATION** From the image, we see that the man is the one standing on the bridge-like platform. The workers are below in a trench, not on the bridge. Therefore, "on the bridge" refers to the man's position, not the workers'. **LOCALIZED REFINEMENT** Update the ambiguous part to make it clear that the man is the one on the bridge: "他在桥上观察工人。" (The man on the bridge) **REPEAT CHECK** No further ambiguities remain.

Figure 2: Example of CoT six-step reasoning resolving the ambiguity.

cess to visual information while mitigating the risk of visual token attention decay during long-sequence generation (Xing et al., 2024; Chu et al., 2025) and hallucination (Wang et al., 2024b). Through this re-examination, the model is better guided to ground its disambiguation decisions in the most relevant visual cues.

5. **Localized Refinement**: Update only the ambiguous parts of the initial translation while keeping the rest unchanged. This constraint prevents unnecessary modifications to the sentence structure and helps maintain overall translation fluency.

6. **Repeat Check**: Reassess the updated translation. If ambiguities remain, iterate steps 3–5 until the translation is fully disambiguated.

An example is provided in Figure 2.

H Error Analysis

H.1 Quantitative Analysis of In-Domain Regressions

Model	Cases	Overthinking	Refinement	Grounding	Other
Qwen2.5-VL-7B	62	54 (87.1%)	5 (8.1%)	2 (3.2%)	1 (1.6%)
InternVL3-8B	64	55 (85.9%)	5 (7.8%)	2 (3.1%)	2 (3.1%)
Total	**126**	**109 (86.5%)**	**10 (7.9%)**	**4 (3.2%)**	**3 (2.4%)**

Table 17: Error types of in-domain regression cases on VIDA-Base-Test, i.e., instances where SFT resolves all annotated ambiguities but CoT-SFT does not. Percentages are computed within each row.

Unlike on the OOD subsets, CoT-SFT does not consistently outperform SFT on the in-domain VIDA-Base-Test (Table 1): it is better for InternVL3-8B but slightly worse for Qwen2.5-VL-7B. To understand this gap, we analyze where the two settings disagree.

Instance-Level Disagreements. Among the 580 test instances, SFT resolves all annotated ambiguities while CoT-SFT does not in 64 instances for InternVL3-8B and 62 for Qwen2.5-VL-7B. Conversely, CoT-SFT is correct while SFT is not in 71 and 48 instances, respectively. The in-domain difference between the two settings is therefore the net result of both gains and regressions, and the regressions are frequent enough to offset most of the gains.

Error Taxonomy. To identify the causes of these regressions, we manually analyze all 126 cases where SFT is correct but CoT-SFT is not, inspecting the CoT-SFT reasoning trace together with the image and the source sentence. Following the order of the reasoning template (Appendix G), we assign each case to one of four mutually exclusive error types:

- **Overthinking**: a correct initial interpretation or translation is later overwritten by an unnecessary or unsupported inference.
- **Refinement error**: the interpretation and reasoning remain correct, but the final localized refinement introduces an omission, unnecessary restructuring, or semantic change.
- **Incorrect visual grounding**: the model misidentifies an object, action, attribute, or

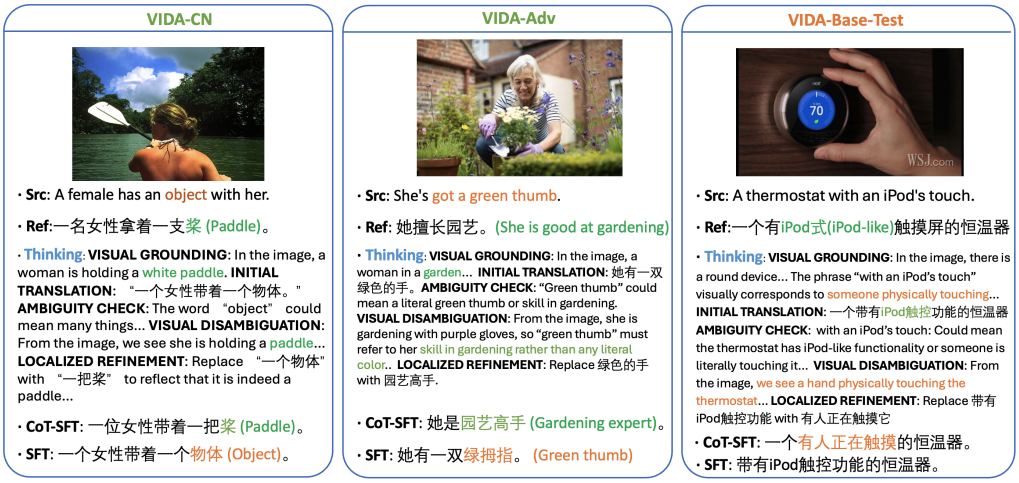


Figure 3: Case study of CoT-SFT vs. SFT

relation in the image, leading to an incorrect interpretation or translation.

- **Other translation errors**: residual lexical, grammatical, omission, or realization errors that are not attributable to the preceding categories.

Results. As shown in Table 17, overthinking accounts for 86.5% of the regressions (109/126), with similar rates for Qwen2.5-VL-7B (87.1%) and InternVL3-8B (85.9%). Refinement errors account for 7.9%, while incorrect visual grounding (3.2%) and other translation errors (2.4%) are rare. The low proportion of grounding errors indicates that the regressions are rarely caused by failing to perceive the image. Instead, the model usually reaches a correct interpretation at an early step but abandons it later in the reasoning process.

Interpretation. This pattern explains why the benefit of CoT-SFT is smaller in-domain than under distribution shift. VIDA-Base-Test shares the word-level ambiguity types of the training data, and SFT already fits the reference translation patterns for many of these instances, so a correct translation can often be produced directly. The fixed six-step template nevertheless requires the model to re-examine every instance, and each additional step provides an opportunity to overwrite an adequate initial translation with spurious evidence, such as irrelevant visual details or an overly literal reading of an idiomatic expression (see the example in Appendix H.2). On the OOD subsets, where direct translation fails more often, the same re-examination is more likely to correct errors than to introduce them. These findings suggest that adapting the reasoning depth to the difficulty of the ambiguity, for example, by stopping after the ambiguity check when no unresolved ambiguity is detected, is a promising direction for reducing overthinking.

H.2 Qualitative Examples

The OOD results in section 5 raise a key question: *how does explicit reliance on visual information shape the model's reasoning?* Figure 3 (left and middle) illustrates how CoT-SFT aligns ambiguous terms with visual evidence in **VIDA-CollN** and **VIDA-Sent**.

The **VIDA-CollN** example (left of Figure 3) illustrates the collective noun ambiguity: the source sentence contains the ambiguous noun "object", which requires a concrete translation ("paddle"). The SFT model, without reasoning, outputs the literal "object," which fails to capture the intended meaning. In contrast, the CoT-SFT shows that the model first generates an initial translation (物体), maintaining the literal meaning. During the ambiguity check, the model detects that "object" is ambiguous. In the subsequent visual disambiguation step, it grounds the word to the image and identifies that the woman is holding a paddle. Finally, in localized refinement, the model updates the translation to "paddle", producing the correct disambiguated output.

The **VIDA-Sent** example (middle of Figure 3) demonstrates sentence-level ambiguity where an idiomatic expression could be misunderstood literally. The phrase "got a green thumb" could be interpreted literally or idiomatically. The SFT

model again produces a literal Chinese rendering of the thumb color. In contrast, the CoT-SFT first provides a literal initial translation (绿色的手). Through visual disambiguation, it recognizes from the image that the woman is gardening, and therefore refines the output to "gardening expert", correctly capturing the idiomatic meaning.

Although CoT-SFT can improve disambiguation accuracy, we observe that it may also introduce *overthinking* on relatively straightforward inputs, which can degrade translation quality. In our reasoning template, this behavior often arises after the model produces an adequate initial translation: subsequent reasoning steps may overwrite the initial output by injecting unnecessary or spurious reasoning, e.g., overusing irrelevant visual details or interpreting idiomatic expressions too literally, thereby leading to flawed revisions.

The right panel of Figure 3 illustrates an overthinking case. The phrase *"iPod's touch"* should be interpreted as *"iPod-like touch screen"*. The model first provides a reasonable image description and recognizes the intended interpretation during ambiguity checking. However, in the later disambiguation step, it over-interprets the phrase by incorrectly linking it to *"someone physically touching"* mentioned in the grounding step, rather than the relevant cue about the product feature. As a result, the model revises an initially adequate interpretation into an incorrect final translation. This example suggests that excessive reasoning can override correct early hypotheses and illustrates the dominant in-domain error type quantified in Appendix H.1.

I Prompts

This section lists the prompts used in VIDA construction and disambiguation evaluation.

Prompt 1: Stage 1 Ambiguity Detection

```
You are a professional text analysis and disambiguation system specialized in English-to-Chinese translation.
Your task is to:
1. Determine whether a given English caption contains any ambiguity when interpreted without any additional
     context or images.
2. If ambiguity exists, explain:
   - The type of ambiguity (lexical, syntactic, pragmatic, or cultural/background).
   - The reason it is ambiguous (how multiple interpretations can arise).
   - Potential different Chinese translations reflecting these interpretations.
3. If no ambiguity exists, respond with that conclusion.

Ambiguity Definition:
- Lexical: a word or phrase has multiple meanings (e.g., "bank" = financial institution vs. river bank).
- Syntactic: the sentence structure permits multiple interpretations (e.g., "I saw the man with a telescope").
- Pragmatic: the context or speaker's intention is unclear (e.g., "Don't put all your eggs in one basket" may
     be literal or metaphorical).
- Cultural/Background: specialized or culturally dependent knowledge is required (e.g., "Break a leg!").

Input:
- A single English caption (no images, no external context).

Output Requirements:
Return a JSON object with the following structure:
{
  "has_ambiguity": <true | false>,
  "ambiguities": [
    {
      "type": "<lexical | syntactic | pragmatic | cultural-background>",
      "explanation": "<why it is ambiguous>",
      "possible_chinese_translations": [
        "<translation option 1>",
        "<translation option 2>"
      ],
      "ambiguous_terms": ["<term1>"]
    }
  ]
}
If the caption is unambiguous:
{ "has_ambiguity": false, "ambiguities": [] }

Important Notes:
- Only consider the caption's text itself (no external context or visual cues).
- If multiple ambiguity types exist, list them all.
- The "possible_chinese_translations" field should illustrate how each interpretation would be rendered
     differently in Chinese.
- Be concise but clear.

Example:
Caption: "Jaguar is approaching rapidly."
{
  "has_ambiguity": true,
  "ambiguities": [
    {
      "type": "lexical",
      "explanation": "The word 'Jaguar' can refer to an animal or a car brand.",
      "possible_chinese_translations": [
        "A jaguar is approaching rapidly.",
        "A Jaguar car is approaching rapidly."
      ],
      "ambiguous_terms": ["Jaguar"]
    }
  ]
}
Now, please analyze the next English caption according to these instructions.
```

Prompt 2: Stage 1 Consensus Merging

```
You are a professional data processing assistant skilled in understanding English ambiguities and structuring
     annotation data. Given ambiguity annotations from two independent models (Qwen-Max and DeepSeek-v3) for
     the same English caption, merge them into a single structured record.

Processing Rules:
1. Merge ambiguities by type:
   - Group ambiguity entries from both qwen_ambi and v3_ambi by their "type" field (lexical, syntactic,
     pragmatic, cultural/background).
   - If two ambiguities from different models share the same type and describe the same underlying issue,
     merge them.
   - If ambiguities differ substantially even under the same type, keep them separate.
2. Merging Details:
   - explanation: combine the explanations from both sources into a single concise paragraph.
   - translations: union the translation candidates from both sources, removing exact duplicates.
3. Extract Ambiguous Terms:
   - From the original English sentence (the "en" field), extract the literal word(s) or phrase(s) that cause
     each ambiguity.
   - Save them into a new field "ambiguous_terms" (a list).
   - Terms must be taken literally from the original English sentence.

Output Format:
[
  {
    "type": "lexical",
    "explanation": "<combined explanation>",
    "translations": [
      "<merged translation 1>",
      "<merged translation 2>"
    ],
    "ambiguous_terms": ["<term1>", "<term2>"]
  },
  {
    "type": "syntactic",
    "explanation": "<combined explanation>",
    "translations": [...],
    "ambiguous_terms": [...]
  }
]
Begin processing based on these instructions.
```

Prompt 3: Stage 2 Disambiguation Translation

```
You are an English-to-Chinese multimodal translation expert.

Input:
1. A single English caption (text).
2. One image showing the real-world scene the caption describes.
3. A list of ambiguity notes generated by a previous model.

Your task:
- Look at BOTH the text and the image.
- Disambiguate the caption and produce the most accurate, fluent Chinese translation.
- Briefly state which ambiguity was resolved by the visual evidence.

Output JSON (Chinese UTF-8):
{
  "translation_zh": "<final Chinese translation>",
  "resolved_ambiguity": "<brief explanation of which type of ambiguity was resolved and how the image
     disambiguated it>"
}

Rules:
- If the caption is actually unambiguous even without the image, set "resolved_ambiguity" to "N/A".
- Do not describe the image; focus on producing the best translation.
- Use concise, standard-register Simplified Chinese with natural and fluent word order.

User-message template:
Caption: "<english_caption>"
Ambiguity notes:
<merged ambiguity records from Stage 1>
```

Prompt 4: Disambiguation Judge

```
You are an English-to-Chinese translation reviewer. Your task is to judge whether a given Chinese translation
     correctly conveys a specified "gold sense" (the intended meaning of an ambiguous term in context).
```

```
Inputs:
1. The English source sentence.
2. The Chinese translation under evaluation.
3. The ambiguous term or phrase in the source (ambi_term).
4. The gold sense (gold_sense), expressed in Chinese, describing the meaning we expect the ambiguous term to
     convey in the given context.
5. A reference Chinese translation.

Task:
Based on the gold sense (4) and the reference translation (5), judge whether the Chinese translation under
     evaluation (2) accurately expresses this meaning.
- If yes, return "Correct".
- If the meaning is missing or distorted, return "Incorrect".

Output format (strictly two lines):
Correct/Incorrect, brief reason

English source: {en}
Chinese translation: {output}
Ambiguous term: {ambi_term}
Gold sense: {gold_sense}
Reference translation: {reference_translation}
```