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

Anonymous ACL submission

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. Experiments with two state-of-the-art LVLMs show that supervised fine-tuning (SFT) improves overall translation quality, while chain-of-thought SFT (CoT-SFT) yields stronger out-of-distribution disambiguation, suggesting that explicit disambiguation guidance improves generalization to diverse ambiguity types.

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 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.

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 that CoT-SFT yields stronger disambiguation under OOD and aggregate evaluation, suggesting better generalization across diverse ambiguity types. 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 further explore a CoT-SFT method by augmenting training with synthetic reasoning traces for MMT disambiguation.

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). In comparison, VIDA targets open-ended MMT generation with broader ambiguity coverage, retaining only cases whose correct interpretation requires visual evidence.

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.

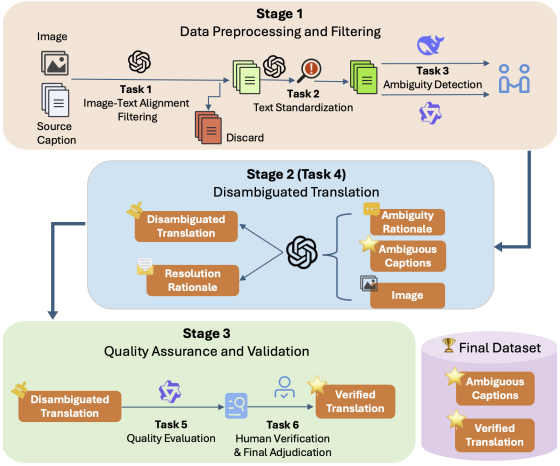

Figure 1: Three-stage VIDA curation pipeline

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.

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 (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 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, while VIDA-CollN and VIDA-Sent are held out as out-of-distribution (OOD) test sets due to limited size.

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*) together with Disambi-Term and Disambi-Inst to evaluate disambiguation accuracy.

Models and Baselines We evaluate two state-of-the-art LVLMs: **Qwen2.5-VL-7B** (Bai et al., 2025), **InternVL3-8B** (Zhu 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 E).

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 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 a commercial LVLM judge and find that CoT-SFT remains comparable to SFT in overall quality (Appendix B). Under Disambiguation-Centric evaluation, CoT-SFT yields the best performance for InternVL3-8B

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.

(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 qualitative analysis in Appendix F.

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. For standard translation metrics, CoT-SFT achieves the highest COMET on both OOD subsets for both models, indicating stronger semantic adequacy on unseen ambiguity types. 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 **+7.5** points for Qwen2.5-VL-7B on both Disambi-Term and Disambi-Inst.). The improvements are substantially larger on VIDA-CollN (over **+15** points for InternVL3-8B and over **+12** points for Qwen2.5-VL-7B), suggesting that CoT-SFT demonstrates stronger generalization beyond the training distribution.

Analysis on All-Test Dataset. All-Test merges all subsets into a single evaluation rather than averaging their scores, reflecting overall performance weighted by subset sizes. 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 balances semantic adequacy and disambiguation accuracy while generalizing better to diverse ambiguity types.

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. Our comparison of SFT and CoT-SFT shows that SFT better fits the surface patterns of the training references, whereas CoT-SFT yields stronger visually grounded disambiguation, especially under distribution shift. This suggests that explicit disambiguation guidance improves generalization to diverse ambiguity types.

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. Our CoT-SFT experiments are also conducted on two mid-sized open-source LVLMs, which leaves open the question of how the strategy transfers to larger or proprietary LVLMs. 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 will be 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.

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*.

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.

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.

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

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.

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 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

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 2: Positioning of VIDA relative to prior multimodal disambiguation resources.

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 2 summarizes the comparison.

B Auxiliary Overall Quality Evaluation

To complement standard MT metrics, we evaluate overall translation quality with a commercial LVLM judge on a 0–100 scale. Table 3 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.

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 3: Auxiliary overall translation quality judged by a commercial LVLM. Scores are on a 0–100 scale and supplement the main metrics in Table 1.

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 4: Filtering funnel for constructing VIDA.

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 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. Sub-

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

Table 5: Inter-annotator agreement for human verification.

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 7: 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 8: 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.

sequent 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 5). The filtering funnel is summarized in Table 4.

C.3 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** contains 256 collective-noun cases 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 6.

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 6: Statistical summary of VIDA subsets.

C.4 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 7 and Table 8. 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.

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. As shown in Table 9, 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 10 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 9: 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 10: 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 final training set is constructed with a positive-to-negative ratio of 2:1. 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 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 access 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.

F Qualitative Analysis

As discussed in section 5, CoT-SFT exhibits a strong ability to enhance disambiguation performance, particularly on challenging OOD subsets

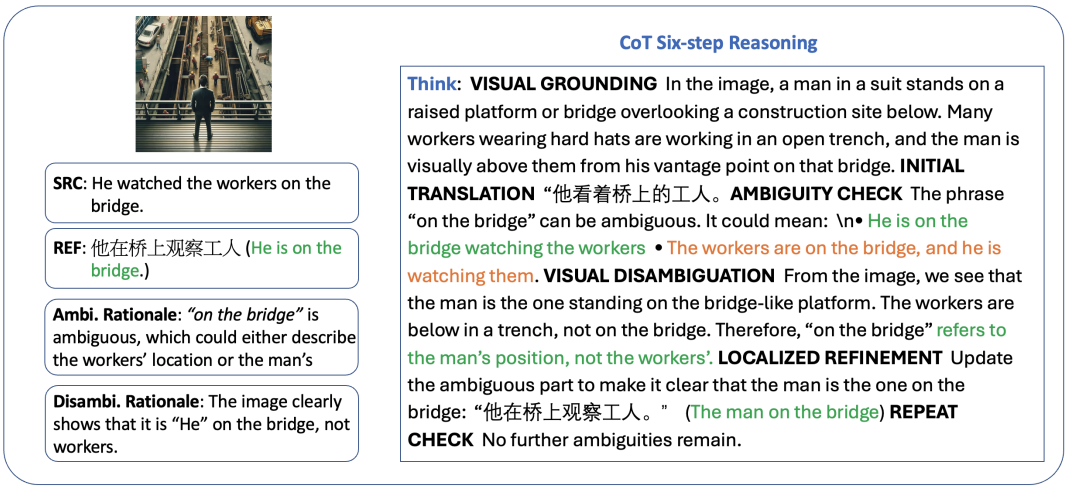

Figure 2: Example of CoT six-step reasoning resolving the ambiguity.

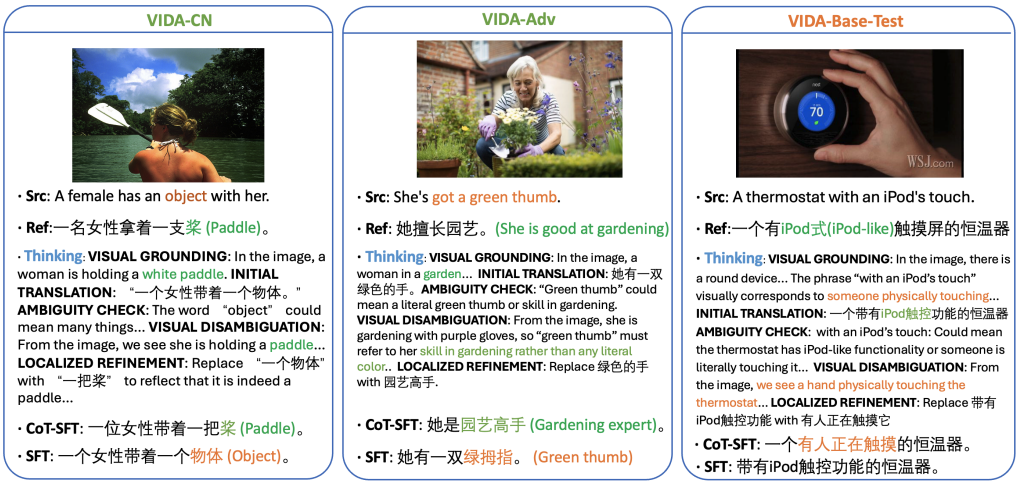

Figure 3: Case study of CoT-SFT vs. SFT

(**VIDA-Sent, VIDA-CollN**). This raises a key question: *how does explicit reliance on visual information shape the model's reasoning?* Figure 3 (left and middle) illustrates two case studies that shed light on this process, showing 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 partially explains the performance drop observed for CoT-SFT on in-domain data.

G 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}
```