Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

Language Technology

MASTERTHESIS

Ambiguity Resolution in Multimodal Machine Translation

Jingheng Pan

Field of Study: Wirtschaftsinformatik
Matriculation No.: 7662185
1st Examiner: Prof. Dr. Chris Biemann, Universität Hamburg
2nd Examiner: Xintong Wang, Universität Hamburg

Language Technology
Department of Informatics
Faculty of Mathematics, Informatics and Natural Sciences

Universität Hamburg
Hamburg, Germany

A thesis submitted for the degree of

Master of Science (M. Sc.)

Masters's Thesis submitted by: Jingheng Pan

Date of Submission: 23.09.2025

Supervisor(s):
Prof. Dr. Chris Biemann, Universität Hamburg
Xintong Wang, Universität Hamburg

Committee:
1st Examiner: Prof. Dr. Chris Biemann, Universität Hamburg
2nd Examiner: Xintong Wang, Universität Hamburg

Universität Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Language Technology

Affidavit

!!! CHANGE THIS TEMPLATE ACCORDING TO YOUR REQUIREMENTS !!!

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe.

I hereby declare in lieu of an oath that I have written this thesis for the Master's degree programme in Computer Science independently and have not used any aids other than those specified - in particular no Internet sources not named in the list of sources. All passages taken verbatim or in spirit from publications are labelled as such. I further certify that I have not previously submitted the thesis in another examination procedure.

23.09.2025	Jingheng Pan
Date	Signature (Jingheng Pan)

Acknowledgements

Personal

I would like to sincerely thank Prof. Dr. Chris Biemann. Attending his course two years ago sparked my deep interest in Natural Language Processing (NLP). I am also grateful for his support in allowing me to join the Language Technology Group as a research assistant, which provided me with the opportunity to engage with cutting-edge research in NLP.

My deepest gratitude also goes to Xintong Wang, both an advisor and a friend. He guided me into the fields of NLP and LVLM research, taught me how to conduct scientific work, and offered invaluable support in both my academic and personal life.

I am especially grateful to my family, whose unwavering encouragement and financial support have given me the confidence to complete my studies and to embark on the upcoming PhD journey. I am thankful to my roommates, who have provided tremendous help in my daily life throughout my three years in Hamburg. Finally, I would like to thank my girlfriend, who has been a constant source of emotional strength and encouragement during moments of uncertainty.

Abstract

Multimodal machine translation (MMT) aims to improve translation quality by incorporating visual information. However, prior studies in MMT suggest that the gains of multimodal models over language-only models are often marginal, raising the core question of whether models truly exploit visual cues in the translation process. With the advent of Large Vision Language Models (LVLMs), this work revisits this question through the lens of ambiguity resolution in MMT, to directly assess whether LVLMs effectively leverage visual information during translation.

Addressing the role of visual cues in translation through ambiguity resolution necessitates datasets with sufficient instances that are irresolvable from text alone but resolvable with visual information. This consideration motivates the first research question *RQ1: Do existing datasets sufficiently support multimodal machine translation disambiguation?* The analysis in this work reveals fundamental limitations of current resources, motivating the construction of the **VIDA (Visually-Dependent Ambiguity)** dataset, a high-quality dataset curated via a three-stage semi-automatic pipeline. The **VIDA** dataset specifically targets visually dependent instances and comprises three subsets that cover diverse disambiguation scenarios, ranging from word-level to sentence-level ambiguities.

Beyond suitable datasets, evaluating whether LVLMs exploit visual cues for disambiguation further requires appropriate evaluation metrics to assess if models truly resolve ambiguities in translation, this work raises *RQ2: Are standard translation metrics adequate for assessing disambiguation performance?* The analysis shows that both lexical- and semantic-level metrics fall short in capturing disambiguation accuracy. In response, this work proposes **Disambiguation-Centric Metrics** which directly measure whether models correctly resolve ambiguous expressions, complementing standard translation metrics that primarily reflect overall translation quality rather than targeted disambiguation.

Having established both a suitable dataset and appropriate evaluation metrics for MMT disambiguation, this work next addresses *RQ3: Do LVLMs effectively utilize visual information for disambiguation?* This work address RQ3 by comparing LVLMs with their language-only backbone models on the VIDA test set using both standard translation metrics and **Disambiguation-Centric Metrics**. The results show modest gains on standard metrics but substantial and consistent improvements on **Disambiguation-Centric Metrics**, confirming that visual input is effectively leveraged to resolve ambiguity and highlighting the necessity of **Disambiguation-Centric Metrics**.

Building on the findings from RQ3, this work further explores how to enhance the capability of LVLMs in leveraging visual information for MMT disambiguation, and presents **Disambiguation-Driven Chain-of-Thought Supervised Fine-Tuning (DDCoT-SFT)**. This training strategy combines a synthetic **Disambiguation-Driven Chain-of-Thought (DDCoT)** template with CoT-based supervised fine-tuning to internalize explicit, visually grounded reasoning for MMT disambiguation. Experimental results show that **DDCoT-SFT** yields stronger semantic adequacy and higher disambiguation

accuracy than conventional SFT settings, particularly on out-of-distribution subsets and the aggregated All-Test set, highlighting superior generalizability beyond the training domain.

Finally, this work evaluates the impact of synthetic versus native reasoning traces in training for MMT disambiguation. Specifically, the **DDCoT-SFT** model, trained with structured **DDCoT** traces, is compared against the same backbone fine-tuned with unstructured native traces extracted from a reasoning model. Results show that **DDCoT-SFT** consistently outperforms the native-CoT fine-tuned model across datasets and metrics, indicating that synthetic reasoning traces—concise, structured, and tailored to the MMT task—offer clearer supervision and yield superior performance over the unstructured and often excessively long native reasoning traces.

In summary, this work contributes the **VIDA** dataset and **Disambiguation-Centric Metrics** as foundational resources, demonstrates that LVLMs can effectively leverage visual cues when appropriately evaluated, and introduces **DDCoT-SFT** as a reasoning-based fine-tuning strategy to strengthen visual utilization, thereby providing both essential resources and a novel perspective for advancing MMT disambiguation research.

Contents

1 Introduction

Multimodal machine translation (MMT) extends conventional neural machine translation (NMT) by incorporating visual information alongside text to improve translation quality (Lala and Specia, 2018; Yao and Wan, 2020). The ability to leverage visual information opens up important applications where text alone is often insufficient. For example, vision–language translation (Wang et al., 2025) is an important application, where accurate recognition of in-image text (e.g., street signs, product labels, or advertisements) and contextually grounded translation are crucial. Another emerging application is Multimodal Sentiment Chat Translation (Liang et al., 2022; Shen et al., 2024), where translation systems leverage both dialogue history and visual context to not only ensure semantic accuracy but also preserve sentiment polarity in bilingual conversations. These applications highlight the importance of effectively integrating textual and visual modalities, which has been the central focus of MMT research.

Early works (Calixto et al., 2017; Huang et al., 2016; Specia et al., 2017; Yin et al., 2020; Yao and Wan, 2020) focused on optimizing model architectures to effectively integrate visual and textual representations, achieving superior performance compared to language-only models. With the emergence of Large Vision Language Models (LVLMs) (Bai et al., 2025; Zhu et al., 2025), recent studies (Gao et al., 2025; Lu et al., 2025) have demonstrated the substantial potential of LVLMs in MMT research. For example, Gao et al. (2025) report state-of-the-art results by enabling deeper text–image interactions, while Lu et al. (2025) show that an LVLM-based translation agent significantly outperforms prior baselines in subtitle and general translation tasks.

Although LVLMs demonstrate impressive performance on MMT benchmarks, the core question remains unresolved: *do LVLMs truly and effectively leverage visual information during the translation process?* Prior studies (Elliott, 2018; Wu et al., 2021) questioned whether visual inputs genuinely contribute to performance gains. For instance, Elliott (2018) showed that adversarially replacing images with unrelated ones had little effect on translation results, while Wu et al. (2021) found that performance improvements often stemmed from regularization effects rather than meaningful visual grounding. These findings from prior studies suggest that the role of visual information in MMT remains unclear. This work revisits whether LVLMs truly and effectively leverage visual information in translation, and approaches the question from the perspective

of **ambiguity resolution**, a scenario in which textual context alone is insufficient and visual cues are essential for disambiguation.

Ambiguity in translation can manifest at multiple levels. At the word level, a single word can admit multiple possible meanings or translations when textual context alone is insufficient (Elliott et al., 2017; Bawden et al., 2018). At the sentence level, ambiguity arises from structural complexities, idiomatic usage, or abstract expressions, leading to multiple plausible interpretations even when individual words are unambiguous (J Lee et al., 2023). Figure 1.1 illustrates concrete cases of both word-level and sentence-level ambiguity. In the word-level example, the English word *instruments* can be interpreted either as *musical instruments* or as *scientific equipment*, with the image providing the necessary cue for correct disambiguation. In the sentence-level example, the phrase *top with stickers on it* can refer to either the laptop itself or the table surface, and the visual context clarifies the intended meaning.

Resolving translation ambiguities, specifically those that are **irresolvable from textual context alone** but can be resolved through visual cues, offers a direct assessment of whether LVLMs genuinely exploit visual information. As illustrated in Figure 1.1, both cases cannot be disambiguated without images: with textual context alone, either musical or scientific equipment is a reasonable interpretation of *instruments*, and either the laptop or the table could be described as the *top with stickers on it*. However, once images are provided, the intended meanings become unambiguous, revealing that the correct interpretations are *musical equipment* and *stickers on the laptop*, respectively. Therefore, if a model can successfully resolve such ambiguities with the assistance of images, this success constitutes strong evidence that the visual modality is being effectively leveraged in the translation process.

To examine whether LVLMs genuinely make effective use of visual information in translation from the perspective of ambiguity resolution, it is essential to rely on datasets that contain sufficient instances where disambiguation can only be achieved through visual cues. This consideration leads to the first research question (RQ1): *Do existing datasets sufficiently support MMT disambiguation?* Several datasets have been proposed to explore the disambiguation task. The 3AM dataset (Ma et al., 2024) targets English–Chinese translation and contributes valuable word-level ambiguity cases, particularly enriching Chinese translation scenarios. The MMA dataset (R Wang et al., 2024) primarily focuses on sentence-level ambiguity, aiming to assess whether models can leverage visual context to interpret ambiguous information within sentences. While both datasets provide useful resources, they also present limitations: the 3AM dataset is hindered by issues of data quality, and MMA, being primarily designed for visual question answering, is not fully aligned with MMT scenarios.

To address the lack of suitable data resources, this work curates a new dataset, **VIDA (Visually-Dependent Ambiguity)** dataset, through a rigorous three-stage semi-automatic pipeline. The **VIDA** dataset is characterized by high ambiguity complexity and strong visual dependency. The dataset encompasses both word-level and sentence-level cases that can only be resolved through visual cues, and is organized into three subsets: (i) VIDA-Base, which contains 1,932 samples and primarily targets word-level ambiguities that require visual context for disambiguation. (ii) VIDA-CollN, which includes 256 samples focusing on the disambiguation of collective nouns, where the abstract notion of a group is grounded through associated visual information. (iii) VIDA-Sent, which provides 312 samples and involves more complex, sentence-level semantic ambiguities.

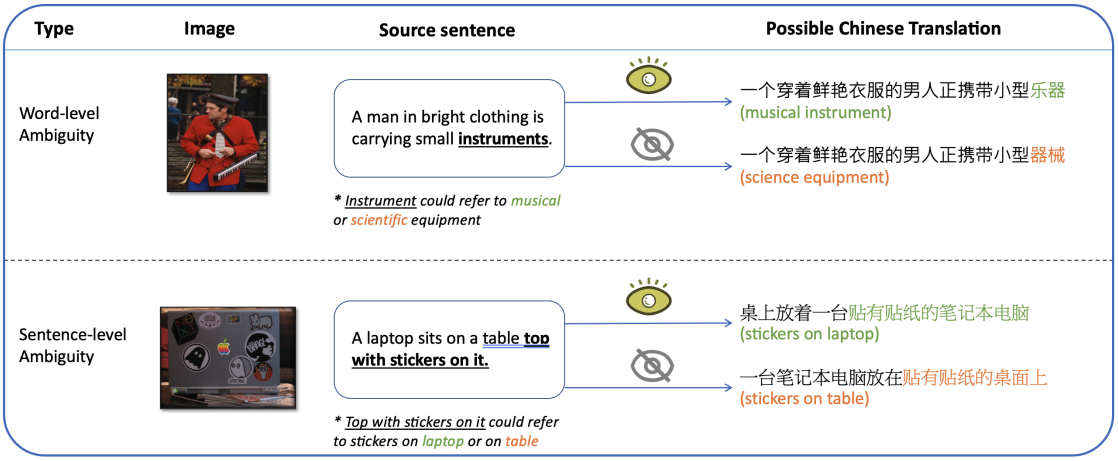

Figure 1.1: Examples of word-level and sentence-level ambiguities in MMT. At the word level, the term instrument may refer to a musical instrument or scientific equipment; at the sentence level, top with stickers on it may describe stickers on a laptop or on a table. Visual context is necessary to select the correct translation.

Establishing the **VIDA** dataset provides a data basis for investigating the central challenge of determining how effectively LVLMs make use of visual information for resolving ambiguity in MMT. However, without appropriate evaluation metrics to determine whether a model's translations successfully resolve ambiguities, it remains unclear whether the improvements of LVLMs stem from genuine disambiguation. Most existing works (Futeral et al., 2023; Ma et al., 2024) rely on standard translation metrics for evaluating MMT performance, which primarily capture overall translation quality, raising doubts about whether these metrics can adequately assess disambiguation performance. Accordingly, the second research question (RQ2) is: *Are standard translation metrics adequate for assessing disambiguation performance?* In addressing RQ2, this work finds that both lexical-level and semantic-level translation metrics, such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020), are not well aligned with the disambiguation task. BLEU prioritizes surface-level n-gram overlap, while COMET emphasizes holistic semantic coherence. Although BLEU and COMET are widely used to evaluate translation quality in NMT and MMT, these metrics are inherently designed to assess overall quality rather than disambiguation accuracy.

In response, this work adopt an LLM-as-a-judge (Gu et al., 2024) approach, where an LLM evaluates whether the ambiguous spans in the source are correctly resolved in the corresponding disambiguated translations and outputs a binary classification. To quantify model performance from these judgments, this work proposes two **Disambiguation-Centric Metrics**, *Disambi-Term* and *Disambi-Inst.*, which are designed to directly assess whether models accurately resolve ambiguous expressions. Specifically, *Disambi-Term* measures the accuracy of individual annotated ambiguous terms across the dataset, while *Disambi-Inst.* applies a stricter sentence-level criterion, counting a prediction as correct only if all ambiguous terms within a sentence are correctly resolved.

With the proposed **Disambiguation-Centric Metrics** serving as the evaluation basis, and the **VIDA** dataset providing the data foundation, this work is now positioned to investigate the role of visual information in supporting disambiguation during translation. Therefore, the third research question (RQ3) is formulated as: *Do LVLMs effectively utilize visual information for disambiguation?* To investigate RQ3, this work

compares three state-of-the-art LVLMs—LLaVA-OneVision-7B (Li et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-8B (Zhu et al., 2025)—with their language-only backbones, serving as text-only baselines on the **VIDA** dataset, evaluated using both standard translation metrics and the proposed Disambiguation-Centric Metrics. The results indicate that more advanced LVLMs are better able to leverage images to improve general translation quality; however, even for these advanced models, the gains captured by standard translation metrics remain modest and do not reveal whether improvements actually come from successful disambiguation. In contrast, Disambi-Term and Disambi-Inst. show substantial and consistent gains across models, directly reflecting the contribution of visual input through its role in resolving ambiguity. Overall, the results highlight (i) the necessity of Disambiguation-Centric Metrics for assessing the contribution of visual information to the disambiguation task in MMT, and (ii) the strong evidence that LVLMs genuinely leverage visual cues to resolve ambiguities during translation.

Building on the finding that LVLMs are capable of leveraging visual information for disambiguation, this work further investigates how their performance on the MMT disambiguation task can be enhanced through specialized fine-tuning strategies. Inspired by the idea of the explicit reasoning paradigm and prior works on Chain-of-Thought Supervised Fine-tuning (CoT-SFT) (Magister et al., 2022; Hsieh et al., 2023; Muennighoff et al., 2025), this work introduces **Disambiguation-Driven Chain-of-Thought Supervised Fine-Tuning (DDCoT-SFT)**, a training strategy that incorporates structured reasoning traces explicitly tailored for resolving translation ambiguities. **DDCoT-SFT** comprises two key components: (i) a disambiguation-oriented reasoning template (**DDCoT**), and (ii) a CoT-SFT procedure that enables models to internalize and apply this reasoning during inference. Specifically, **DDCoT** provides a six-step structured reasoning template that explicitly aligns ambiguous expressions with visual evidence to guide accurate disambiguation. In addition, the CoT-SFT training strategy embeds the structured reasoning patterns of **DDCoT**, enabling the model to internalize the template during inference. As a result, **DDCoT-SFT** performs explicit step-by-step reasoning when resolving translation ambiguities, which not only improves disambiguation accuracy but also provides interpretability and transparency into the model's decision-making process.

Experiments are conducted on the **VIDA** test set, evaluating InternVL3-8B (Zhu et al., 2025) and Qwen2.5-VL-7B (Bai et al., 2025) under three training settings: Vanilla, SFT, and the proposed **DDCoT-SFT**. Evaluation is carried out on three subsets: the in-distribution set (VIDA-Base-Test), the out-of-distribution sets (VIDA-Sent and VIDA-CollN), and the union All-Test set. Performance is assessed with standard translation metrics and the proposed Disambiguation-Centric Metrics, enabling a comprehensive evaluation of overall translation quality as well as disambiguation accuracy. Building on these settings, the experimental results show that **DDCoT-SFT** achieves substantial and consistent improvements on the proposed Disambiguation-Centric Metrics compared to standard SFT, specifically on the OOD subsets and the aggregated All-Test set. The OOD subsets evaluate whether models can generalize disambiguation ability to unseen ambiguity types, while the All-Test set provides a comprehensive assessment across diverse ambiguity cases. These improvements indicate that **DDCoT-SFT** adapts well to varied ambiguity types and demonstrates superior generalization in visually grounded disambiguation.

The experimental results demonstrate that DDCoT-SFT, which leverages a synthetic, structured reasoning approach, is highly effective for MMT disambiguation, presenting an interesting contrast to the findings of Muennighoff et al. (2025), who showed that using unstructured, native reasoning traces yields superior improvements on mathematical reasoning tasks. This contrast raises the question of how such native traces perform in the context of MMT disambiguation. Native reasoning traces are raw, free-form chains of thought generated by a reasoning model without external design or template constraints, whereas synthetic reasoning traces are deliberately crafted under human guidance with explicit task goals. To this end, this work compares synthetic traces (**DDCoT**) with native traces sampled from a strong reasoning model. Under the same backbone and inputs, **DDCoT-SFT** consistently outperforms native-CoT fine-tuned model across datasets and metrics. These results suggest that while native reasoning traces reflect spontaneous reasoning, they are often unstructured and excessively long, which obscures the critical steps required for translation disambiguation. In contrast, **DDCoT** provides concise, structured reasoning explicitly tailored to the MMT task, yielding clearer supervision and superior performance.

The main contributions of this work are summarized as follows:

- This work introduces **VIDA (Visually-Dependent Ambiguity)**, a dataset characterized by high ambiguity complexity and strong visual dependence, curated via a rigorous semi-automatic pipeline and covering both word-level and sentence-level ambiguities.
- To quantitatively evaluate disambiguation performance, **Disambiguation-Centric Metrics** (*Disambi-Term* and *Disambi-Inst.*) are proposed to directly measure a model's ability to resolve ambiguous expressions, complementing standard translation metrics for MMT disambiguation.
- This work further provides empirical evidence that **LVLMs leverage visual information for disambiguation** by comparing LVLMs with their language-only backbone models on the **VIDA** test set. The results validate the necessity of the Disambiguation-Centric Metrics and confirm that LVLMs effectively use visual cues to resolve ambiguities.
- A novel training strategy tailored to the MMT disambiguation task is presented: **DDCoT-SFT (Disambiguation-Driven Chain-of-Thought Supervised Fine-Tuning)**. By guiding models to perform explicit reasoning that aligns ambiguous text with visual evidence, **DDCoT-SFT** yields consistent improvements in disambiguation accuracy across **VIDA** subsets.
- A comparative analysis is conducted between **DDCoT-SFT** model and native CoT-SFT model on the MMT disambiguation task. The results show that **DDCoT-SFT** consistently outperforms the native CoT-SFT model, indicating that using concise, step-wise **DDCoT** reasoning traces as supervision yield more reliable and effective visually grounded disambiguation.

2
Related Work

2.1 Large Vision Language Models

Large Language Models (LLMs) demonstrate remarkable proficiency in various language understanding and generation tasks (Bai, Bai, Chu, et al., 2023; Touvron et al., 2023). These advances have spurred the development of Large Vision Language Models (LVLMs), which extend LLM capabilities to multimodal settings (H Liu et al., 2023; Bai, Bai, Yang, et al., 2023). LVLMs employ a cross-modal fusion module to integrate visual and textual inputs, enabling them to perform a wide range of multimodal tasks, including captioning (Junnan Li et al., 2023), visual question answering (Antol et al., 2015), and translation (Jiaoda Li et al., 2021), in a unified manner.

In the early development of multimodal models, architectures were generally divided into single-tower and dual-tower designs (Fields and Kennington, 2023). Single-tower structures unify visual and textual inputs into the same Transformer encoder for end-to-end joint modeling, as in VisualBERT (LH Li et al., 2019) and UNITER (YC Chen et al., 2020). Formally, the single-tower model can be expressed as a function

$$f_{\text{single}} : (\mathbf{x}_{\text{text}}, \mathbf{x}_{\text{image}}) \mapsto \mathbf{h}, \tag{2.1}$$

where both modalities are concatenated and jointly encoded into a shared representation $\mathbf{h}$. This unified encoding enables unconstrained modality interaction but the quadratic self-attention over concatenated tokens makes the model computationally expensive, and parameter sharing reduces modularity compared to dual-encoder designs (Fields and Kennington, 2023).

In contrast, dual-tower architectures adopt separate encoders for vision and language and align them in a shared representation space, with CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) as prominent examples. In this case, dual-tower architecture computes

$$f_{\text{dual}} : (\mathbf{x}_{\text{text}}, \mathbf{x}_{\text{image}}) \mapsto (\mathbf{h}_{\text{text}}, \mathbf{h}_{\text{image}}), \tag{2.2}$$

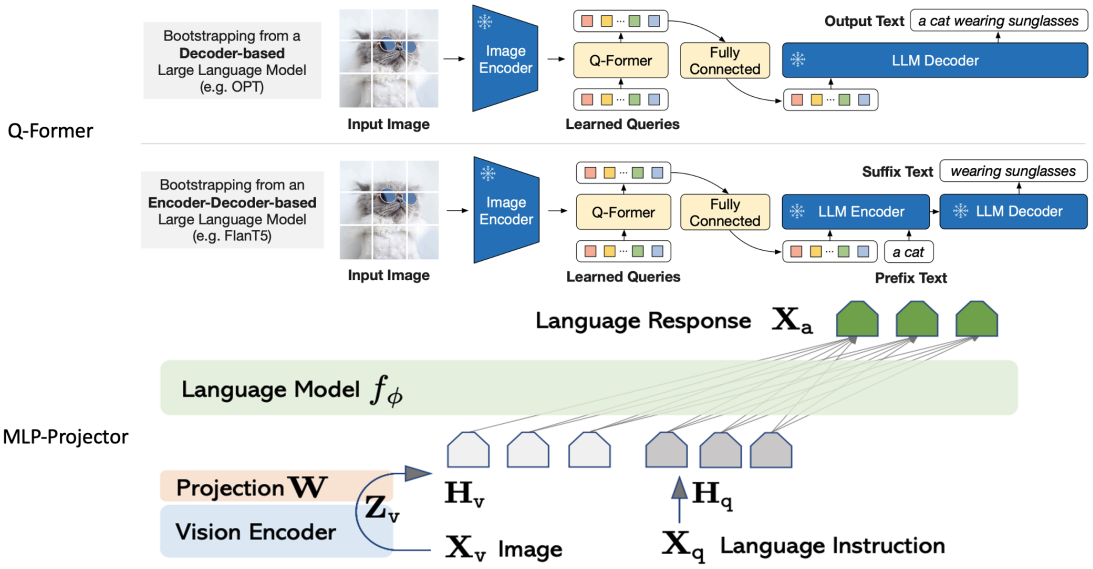

Figure 2.1: Illustration of two representative fusion modules in LVLMs: (Above) Q-Former, which generates learned query tokens from image features, and (Below) MLP-Projector, which directly maps vision embeddings into the LLM embedding space.

and relies on a contrastive objective

$$\mathcal{L}_{\text{contrast}} = -\sum_i \log \frac{\exp(\text{sim}(\mathbf{h}^i_{\text{text}}, \mathbf{h}^i_{\text{image}})/\tau)}{\sum_j \exp(\text{sim}(\mathbf{h}^i_{\text{text}}, \mathbf{h}^j_{\text{image}})/\tau)}, \tag{2.3}$$

to align the two modality-specific embeddings in a joint space. Here, i indexes a matched text–image pair, while j ranges over all images in the mini-batch, so that the denominator contrasts the true pair against both positive ($j = i$) and negative ($j \neq i$) candidates. The function $\text{sim}(\cdot, \cdot)$ denotes a similarity measure such as cosine similarity, and τ is a temperature hyperparameter that controls the sharpness of the softmax distribution.

Dual-tower architectures are more efficient than single-tower to train and easier to scale, but the cross-modal interactions are relatively shallow, limiting the performance on complex reasoning tasks, for which deeper fusion is typically required (Fields and Kennington, 2023). Overall, single- and dual-tower architectures laid an important foundation for subsequent large vision language models, which have progressively evolved toward more sophisticated fusion strategies and stronger multimodal aligning capabilities.

To overcome the limitations of early single- and dual-tower models, subsequent works introduced lightweight alignment modules to better connect visual encoders with LLMs. A representative design is the Q-Former (Junnan Li et al., 2023; Zhu et al., 2023), which employs a query-based Transformer to distill visual features into a compact set of tokens before feeding them into an LLM. Formally, given visual features $\mathbf{Z}_v = g(\mathbf{X}_v)$ extracted from an image $\mathbf{X}_v$, Q-Former generates a compact representation through cross-attention with a set of learnable query tokens $\mathbf{Q}$:

$$\mathbf{H}_v = \text{QFormer}(\mathbf{Z}_v, \mathbf{Q}), \tag{2.4}$$

where $\mathbf{H}_v$ are the distilled visual tokens aligned with the LLM input space.

Another line of work adopts a simpler yet effective **MLP-based projection** to map image features into the embedding space of the LLM. LLaVA (Zhu et al., 2025) exemplifies this approach by projecting visual features into the language model's token space and fine-tuning the alignment through instruction-following data. In this case, the mapping can be implemented as a linear projection:

$$\mathbf{H}_v = W \cdot \mathbf{Z}_v. \tag{2.5}$$

Figure 2.1 illustrates these two representative designs: Q-Former generates learned query tokens through cross-attention with image features, while MLP projectors directly map visual embeddings into the language model space. Both Q-Former and MLP-based projectors enable more effective vision–language interaction while significantly reducing the computational burden compared to direct fusion. Compared with early single- and dual-tower structures, these lightweight alignment designs strike a balance between efficiency and interaction depth, thereby marking an important step toward the scalable integration of vision encoders with pretrained LLMs (H Liu et al., 2024).

More recent efforts have further advanced the research by refining lightweight connectors and strengthening the visual encoder to produce high-fidelity visual features. While both Q-Former and MLP connectors are lightweight, many state-of-the-art LVLMs (Bai et al., 2025; Zhu et al., 2025; Li et al., 2024) increasingly standardize on a simple MLP-based projector to align vision and language embeddings, and pair it with stronger vision backbones to capture fine-grained visual features. For example, Qwen2.5-VL (Bai et al., 2025) employs native-resolution ViTs with spatial–temporal tokenization to capture fine-grained spatial cues. Similarly, InternVL3 (Zhu et al., 2025) follows the ViT–MLP–LLM paradigm but pushes toward leveraging tiling and pixel-unshuffle strategies to preserve high-resolution image details. Compared with earlier models such as BLIP-2 and LLaVA, the modern LVLMs demonstrate a stronger capacity to retain high-resolution visual details while keeping the alignment process efficient.

Despite the impressive advancements of modern LVLMs, many models still face challenges in generating responses that are faithfully grounded in the visual input. One of the most critical issues is hallucination (Xintong Wang et al., 2024; Leng et al., 2024), where models produce textual descriptions inconsistent with the actual visual content, thereby limiting their reliability in real-world applications. **Our work** (Xintong Wang et al., 2024) analyze hallucination in LVLMs and attribute it primarily to language priors and statistical biases, which lead models to generate text weakly grounded in the visual input. To mitigate hallucination, we propose Instruction Contrastive Decoding (ICD), which deliberately amplifies hallucinations via disturbed instructions and employs contrastive decoding to suppress unstable components, thereby reducing hallucinations in both discriminative and generative benchmarks.

Another challenging task is Vision–Language Translation (VLT), which requires accurate recognition of in-image text and contextually grounded translation into the target language (Wang et al., 2025). **Our work** (Wang et al., 2025) highlight major obstacles of current research on VLT including low-quality datasets with OCR noise and culturally inconsistent references, strong OCR dependency in existing models, and unreliable evaluation metrics under varying text density in the image. To address these three issues, we introduce AibTrans, a human-verified multilingual dataset, and propose a Density-Aware Score (DA Score) for fairer evaluation, together with balanced multilingual fine-tuning to improve cross-lingual performance.

This work also points to another challenge for LVLMs, namely Multimodal Machine Translation (MMT), which raises the question of whether LVLMs truly leverage visual information during the translation process. More details on the task will be provided in the next section.

2.2 Multimodal Machine Translation

Multimodal machine translation (MMT) is an increasingly important area of research that seeks to improve translation quality by leveraging modalities beyond text (Elliott et al., 2016). Compared with conventional neural machine translation (NMT) (Bahdanau et al., 2014), MMT faces additional challenges, as the model must not only align source and target languages but also effectively integrate visual information, which requires accurate cross-modal grounding as well as robustness to irrelevant or noisy visual cues.

A variety of approaches have been proposed for multimodal machine translation, focusing on how to effectively integrate visual information with textual representations. Early work (Calixto et al., 2017; Huang et al., 2016) explored attention-based architectures that allow the model to selectively attend to both textual and visual features. For example, the doubly-attentive decoder of Calixto et al. (2017) extends the neural machine translation framework with dual attention over source words and image regions, while Huang et al. (2016) introduce an attention-based MMT model that jointly learns alignments across modalities.

However, several studies (Elliott, 2018; Wu et al., 2021) critically examine whether visual information truly contributes to translation quality. Elliott (2018) performed adversarial evaluations by replacing input images with unrelated ones, showing that many systems remained largely unaffected, thus questioning the sensitivity of MMT models to visual signals. Similarly, Wu et al. (2021) revisited the role of visual context with interpretable model designs, revealing that performance gains often stem from regularization effects rather than genuine use of visual cues.

Formally, this distinction can be expressed by comparing a translation model that generates the target translation $\mathbf{t}$ from the source sentence $\mathbf{s}$:

$$P(\mathbf{t} \mid \mathbf{s}), \tag{2.6}$$

with a model that additionally conditions on the paired image $\mathbf{v}$:

$$P(\mathbf{t} \mid \mathbf{s}, \mathbf{v}). \tag{2.7}$$

If visual information is genuinely exploited, then conditioning on $\mathbf{v}$ should alter the predictive distribution, i.e.,

$$P(\mathbf{t} \mid \mathbf{s}, \mathbf{v}) \not\equiv P(\mathbf{t} \mid \mathbf{s}), \tag{2.8}$$

These studies highlight the need for careful analysis when attributing improvements to multimodal integration and raise a critical question: *do LVLMs truly and effectively leverage visual information during translation?* This motivates the present work, which investigates MMT from the perspective of ambiguity resolution, a scenario where visual cues are indispensable for disambiguation.

Ambiguity often emerges in translation when the same source expression can be understood in different ways, and it is precisely in such cases that visual context can be

most helpful (Shen et al., 2024). At the word level, ambiguity arises from polysemous words, morphology (Shen et al., 2024). For example, the English word *bank* could denote either a financial institution or the side of a river; similarly, a form like *book* can be interpreted as a noun or as a verb. Beyond individual words, ambiguity also manifests at the sentence level, where structural or pragmatic factors lead to multiple possible readings (Berzak et al., 2016). A classic case is syntactic attachment, as in *I saw the man with a telescope*, where the prepositional phrase could describe either the man or the act of seeing. Pragmatic phenomena like irony, sarcasm, or context-dependent pronouns add yet another layer of complexity, requiring knowledge that goes beyond text alone.

To support the research of ambiguity resolution in MMT, several datasets exist. The Multimodal Lexical Translation (MLT) dataset (Lala and Specia, 2018) specifically targets lexical ambiguity, providing fine-grained annotations where image information is essential for selecting the correct translation. Building on this line, Jiaoda Li et al. (2021) introduced the Ambiguous Captions (AmbigCaps) dataset, which constructs gender-related ambiguities through back-translation and requires visual cues for correct disambiguation. Futeral et al. (2023) proposed the CoMMuTE benchmark, which provides contrastive pairs of ambiguous sentences and corresponding images, enabling fine-grained evaluation of whether models genuinely use visual information in disambiguation. More recently, 3AM (Ma et al., 2024) offers an ambiguity-aware multimodal benchmark in the English–Chinese translation setting, explicitly designed to evaluate how models exploit visual context to resolve ambiguous expressions.

However, while 3AM extends the scope beyond European languages and contributes valuable word-level ambiguity cases in Chinese translation scenarios, it suffers from data quality issues that restrict its reliability for disambiguation analysis. This limitations of data quality, motivate the construction of my proposed VIDA (Visually-Dependent Ambiguity) dataset, which features higher ambiguity complexity and stronger visual dependency. Notably, all the aforementioned benchmarks evaluate disambiguation indirectly by relying on standard translation metrics such as BLEU (Papineni et al., 2002) or COMET (Rei et al., 2020), which assess the similarity between the system output and reference translations as a proxy for translation quality. This evaluation paradigm raises an important question, corresponding to RQ2 of this work: *Are standard translation metrics adequate for assessing disambiguation performance?* A more detailed discussion will be provided in section 4.2.

2.3 Chain-of-Thought Reasoning

Large language models (LLMs) often struggle with tasks that require multi-step reasoning, such as arithmetic, logic, and multi-hop question answering (S Wang et al., 2024; Boye and Moell, 2025). To enhance the logical thinking ability, Chain-of-Thought (CoT) prompting was introduced by Wei et al. (2022), where models are encouraged to generate intermediate reasoning steps before producing the final answer. This approach significantly improves performance on complex reasoning tasks, and later work showed that even simple prompts such as "Let's think step by step" can elicit reasoning chains in a zero-shot setting (Kojima et al., 2022). A series of extensions followed, including self-consistency (Xuezhi Wang et al., 2022) and Tree-of-Thoughts (Yao et al., 2023), which improve robustness by aggregating multiple reasoning paths or exploring tree-

structured search spaces. However, all these methods remain prompt-based and operate only at inference time.

Beyond prompt-based methods, recent advances have focused on training dedicated reasoning models that are explicitly optimized for multi-step reasoning. Reinforcement learning strategies such as GRPO (Shao et al., 2024) have been applied to align models toward generating coherent and verifiable reasoning steps, while models like DeepSeek-R1 (Guo et al., 2025) and OpenAI's o1 series (Jaech et al., 2024) are trained with specialized objectives and large-scale reasoning traces, achieving much stronger reasoning performance than standard instruction-tuned LLMs. Importantly, the native reasoning traces generated by these reasoning models have also been leveraged for supervised fine-tuning (SFT) of smaller models.

Formally, given training pairs (x_i, y_i), standard SFT minimizes

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{N} \log P_\theta(y_i \mid x_i), \tag{2.9}$$

where the model is directly optimized to predict the target output y_i from input x_i. By contrast, approaches that incorporate reasoning traces augment each pair with a chain of thought generated by a reasoning model,

$$r_i^{\text{nat}} \sim P_{\mathcal{M}_{\text{Reason}}}(r \mid x_i) \tag{2.10}$$

and optimize

$$\mathcal{L}_{\text{SFT+CoT}} = -\sum_{i=1}^{N} \log P_\theta(r_i^{\text{nat}}, y_i \mid x_i), \tag{2.11}$$

thus encouraging the model to reproduce not only the final output but also the intermediate reasoning steps.

Works such as R1-Distill (H Zhao et al., 2025) and s1 (Muennighoff et al., 2025) demonstrate that exposing compact models to such native reasoning traces—with as few as 1K carefully curated examples—can significantly boost their reasoning performance, particularly on math-heavy benchmarks, highlighting a growing trend of using reasoning models and their native reasoning traces as supervision for distillation.

Nevertheless, native reasoning traces often produce excessively long chains of thought (Sui et al., 2025). While such extended reasoning paths can be beneficial for logic-intensive domains such as mathematics (Muennighoff et al., 2025), they also increase inference latency and cause models to "overthink" in tasks that do not inherently require complex reasoning (She et al., 2025), such as machine translation task. To address the drawbacks of native traces, an alternative is to employ synthetic reasoning chains-manually designed or automatically generated traces that are shorter, more controllable, and tailored to specific tasks. In this case, the reasoning trace is constructed via a deterministic mapping under a CoT template $\mathcal{T}_{CoT}$:

$$r_i^{\text{syn}} = g(x_i, y_i; \mathcal{T}_{CoT}), \tag{2.12}$$

where g denotes a generation function that takes the input–output pair (x_i, y_i) and produces a structured reasoning trace following template $\mathcal{T}_{CoT}$.

Early works (Magister et al., 2022; Hsieh et al., 2023) have demonstrated the effectiveness of fine-tuning on synthetic reasoning traces as a means of transferring reasoning

capabilities. For example, Magister et al. (2022) fine-tuned small models on teacher-generated reasoning traces, significantly boosting performance on arithmetic and symbolic reasoning. Hsieh et al. (2023) extended this idea by incorporating teacher rationales in distillation, enabling compact models to outperform larger ones with fewer samples.

In the context of using synthetic reasoning traces in translation task, J Wang et al. (2025) show that constructing synthetic reasoning chains via a multi-agent pipeline and training MT models on them leads to substantial improvements, particularly for challenging literary texts. For multimodal machine translation, D Liu et al. (2025) introduce synthetic disambiguation rationales into the training process, demonstrating that explicitly modeling a reasoning step for visual disambiguation yields superior handling of ambiguous inputs and improved translation quality. These findings from prior work on translation and MMT indicate that synthetic reasoning traces, compared with native CoT, are more effective for task-specific applications.

Motivated by the idea of synthetic reasoning traces in translation task, this work proposes **DDCoT**, a synthetic CoT specifically designed for disambiguation in MMT, which guides models to perform explicit reasoning that aligns ambiguous text with visual evidence. Furthermore, this work systematically compare the proposed synthetic CoT (DDCoT) with native CoT. More details are provided in chapter 5 and section 6.4.

3
Dataset Curation

3.1 Limitations of Current Disambiguation Datasets

A central objective of this work is to determine whether LVLMs genuinely exploit visual information in translation through the lens of ambiguity resolution. Advancing this objective requires datasets that contain visually dependent instances where textual context alone is insufficient but the image provides decisive evidence. This motivates RQ1: *Do existing datasets sufficiently support multimodal machine translation disambiguation?* To address RQ1, this section investigates the limitations of existing multimodal disambiguation datasets.

Two representative datasets currently available for multimodal disambiguation research are 3AM (Ma et al., 2024) and MMA(R Wang et al., 2024). The 3AM dataset targets English-to-Chinese translation scenarios and primarily focuses on word-level ambiguity, but suffers from overall poor data quality with the following two issues:

- **Data Integrity and Noise**: Including extensive image-text mismatches, grammatical errors in English source texts, chaotic punctuation usage, frequent noise words, etc., which compromise the reliability of model training and evaluation.

- **Insufficient ambiguity**: Not all samples exhibit ambiguities that depend on multimodal context; rather, some ambiguities are linguistically resolvable and can be accurately translated without any visual support. The presence of such samples reduces the dataset's sensitivity to the contribution of visual information, thereby limiting its representativeness for MMT disambiguation research.

Additionally, MMA is a benchmark constructed in Visual Question Answering (VQA) format, primarily focusing on sentence-level ambiguity to assess whether models can leverage visual context to interpret ambiguous information within sentences. The benchmark achieves this by pairing a single question with two different images that suggest divergent interpretations. Although this task relates to semantic ambiguity, it is not tailored for translation scenarios and thus cannot be directly applied to evaluating ambiguity resolution in multimodal machine translation task.

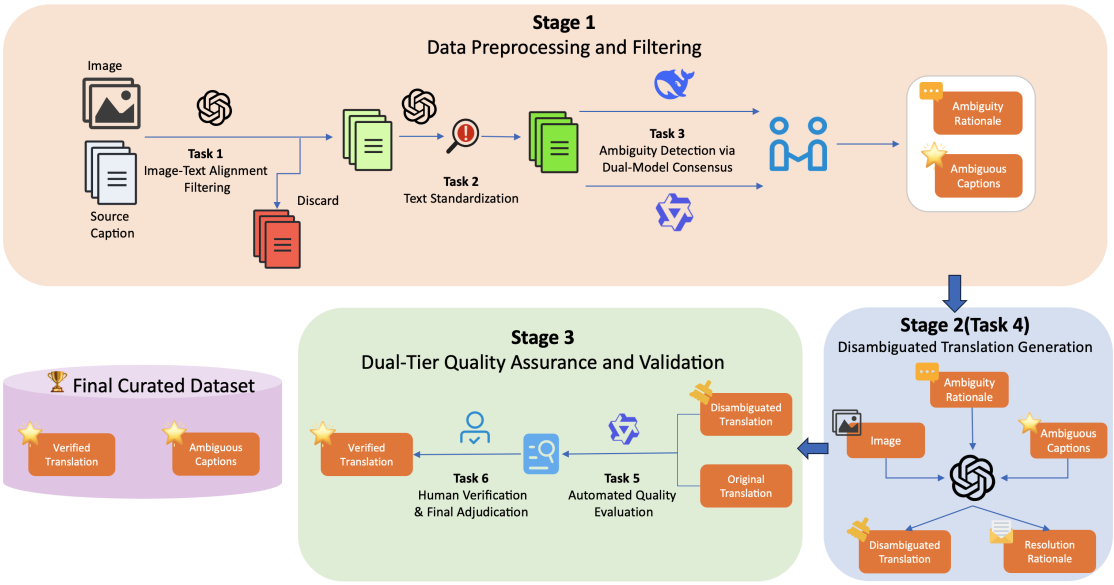

Figure 3.1: Dataset Curation Pipeline

3.2 Dataset Curation Pipeline

To address the limitations discussed in section 3.1, I curated a new dataset, **VIDA** (**V**isually-**I**ndependent **A**mbiguity) dataset, through a rigorous, three-stage semi-automatic pipeline, as illustrated in Figure 3.1. The pipeline is specifically designed to extract visually dependent ambiguities from large-scale data and to produce high-fidelity disambiguated translations.

3.2.1 Source Dataset and Pre-filtering for MMA

The process began with two source dataset: the 3AM dataset, containing 26,000 English-Chinese parallel sentences, and the MMA dataset, with an initial 521 VQA samples. While the 3AM dataset could be directly used in the pipeline, the MMA dataset required a specific pre-filtering step to align its VQA-style content with the translation-focused objectives.

For the MMA dataset, I identified and removed samples that function as standard VQA tasks, which rely on finding direct visual cues for an answer. In these cases, a simple question (e.g., "Who holds the crown in this scenario?") becomes ambiguous only because of the image content, not the language itself. Since this work focuses on resolving linguistic ambiguity through visual context, these samples were filtered out, preserving 256 entries where the ambiguity originates in the text.

3.2.2 Stage 1: Data Preprocessing and Filtering

The primary objective of this initial stage is to obtain an image–text aligned and textual-error-free collection of ambiguous source captions.

Task 1: Image-Text Alignment Filtering. To reduce noise from mismatched image-text pairs, I first conduct a data cleaning process. Specifically, GPT-4o (Hurst et al., 2024)

is employed to determine whether the source caption is semantically aligned with the corresponding image. Pairs identified as inconsistent are discarded, ensuring that the subsequent processing is performed on a corpus of aligned image-text samples.

Task 2: Text Standardization. To prevent textual errors from interfering with subsequent ambiguity detection, the source captions are standardized using GPT-4o. This process corrects grammatical mistakes, spelling errors, and punctuation issues.

Task 3: Ambiguity Detection via Dual-Model Consensus. The goal of this task is to isolate captions with visually-dependent ambiguities, removing those that are either unambiguous or whose ambiguities can be resolved linguistically without visual support. To achieve this goal, I employ two models for cross-checking, retaining only captions on which they reach consensus. This mitigates single-model bias and ensures that only captions consistently judged as ambiguous are retained, thereby improving the precision and reliability of the dataset.

- **Parallel Independent Detection**: I utilize two distinct Large Language Models, Qwen-Max (Team, 2024) and DeepSeek-v3 (Liu et al., 2024), to independently analyze each standardized caption. Each model is prompted to assess whether the English sentence, considered in the context of translation into Chinese and without access to any visual information, contains ambiguous expressions that could yield multiple valid translations.

- **Consensus Filtering**: A caption is retained only if both models concur in identifying it as ambiguous. Such samples, along with their model-generated "Ambiguity Rationale", are preserved for the dataset.

3.2.3 Stage 2: Disambiguated Translation Generation

Task 4: Disambiguated Translation Generation. As the core stage of the pipeline, this task produces high-quality, disambiguated translations for each filtered source caption. To achieve this, GPT-4o is guided with a structured tripartite input: (1) the *Cleaned Ambiguous Caption* as the source text, (2) the *Image* as the essential visual context, and (3) the *Ambiguity Rationale* from Task 3, explicitly directing the model to the specific point of ambiguity that must be resolved. The output is twofold: *Disambiguated Translation* that accurately resolves the ambiguity, and *Resolution Rationale* explaining how visual information was utilized to resolve the ambiguity.

For instance, using the word-Level ambiguity case illustrated in Figure 3.2, when presented with the ambiguous caption "*Two trunks stacked next to an open door on a sidewalk*", the model is also given the rationale that "trunks" could refer to suitcases or trees. By observing the accompanying image of two leather suitcases, GPT-4o correctly generates the disambiguated Chinese translation for "suitcases" and a Resolution Rationale explaining that the visual context ruled out the alternative meaning. This demonstrates how the triplet input guides the model to a precise, visually-grounded translation.

Figure 3.2: Examples of Word-Level, Sentence-Level, and Collective Noun ambiguities from the dataset. Each panel demonstrates how the visual context is used to resolve the issue described in the Ambiguity Rationale (Ambiguous) and derive a correct, disambiguated translation (Ref).

3.2.4 Stage 3: Dual-Tier Quality Assurance and Validation

This final stage implements a rigorous quality control process that incorporates both a "Model-as-a-Judge" strategy and a human-in-the-loop approach to ensure the reliability and overall quality of the dataset.

Task 5: Automated Quality Evaluation In this task, Qwen-Max is prompted to compare the disambiguated translations from Task 4 against the original translations and determine which translation is superior by considering two key dimensions:

- **Semantic Preservation**: whether the translation faithfully conveys the intended meaning of the source text.
- **Fluency**: the linguistic quality of the translated text, including grammatical correctness and naturalness.

The evaluation reports from this task identify cases in which GPT-4o's disambiguated translation is judged inferior, and these are flagged for subsequent human verification.

Task 6: Human Verification and Final Adjudication To guarantee the ultimate quality of the dataset, human verification is conducted only for cases in which the automated judge deems the new translation inferior to the translation from the original dataset. These selected cases are forwarded to a expert annotator, who then adjudicate by either selecting the superior translation or providing a corrected version based on a holistic evaluation.

Notably, during the human verification phase of Task 6, I identified a distinct category of samples containing **collective nouns** that GPT-4o from Task 3 contextualizes into

Subset	Ambiguity Focused	Size	Avg. Length (words)	Avg. Ambi. Terms
VIDA-Base	Word-Level	1,932	11.12	1.7826
VIDA-Sent	Sententence-Level	312	6.00	1.00
VIDA-CollN	Word-Level	256	10.08	1.20

Table 3.1: Statistical summary of **VIDA** subsets

specific entities based on visual cues, such as rendering "third-party person" as 摊贩 (street vendor), as illustrated in Figure 3.2. While this produces semantically precise translations, Qwen-Max from Task 5 flagged these as over-translations due to lexical divergence from the source text. I contend that such vision-guided semantic specification is both necessary and appropriate in multimodal translation, as literal translations often yield incomplete or unnatural outputs. Consequently, I designated these samples as the **Collective Noun Subset**.

3.3 VIDA: A New Dataset for Multimodal Machine Translation Disambiguation

The rigorous pipeline outlined in section 3.2 results in the construction of a new dataset, **VIDA** (**V**isually-**D**ependent **A**mbiguity). In total, **VIDA** comprises 2,500 instances specifically curated to feature high ambiguity complexity and visual dependency. The dataset comprehensively covers both word-level and sentence-level ambiguities and is organized into the following three subsets:

- **VIDA-Base**: Curated from the 3AM dataset, this subset contains **1,932 samples**, primarily focusing on word-level ambiguities that require visual context for resolution.
- **VIDA-CollN (Collective Noun Subset)**: This specialized subset consists of **256 samples** focusing on the disambiguation of collective nouns, where the abstract nature of the group is made concrete by the associated visual information.
- **VIDA-Sent**: Adapted from the MMA dataset, this subset provides **312 samples**. These instances tend to exhibit more complex, sentence-level semantic ambiguities that necessitate a holistic understanding of the image for correct interpretation and translation.

A complete statistical summary of these subsets is provided in Table 3.1. For each subset, the column *Ambiguity* specifies the primary ambiguity focus (word-level, sentence-level, or mixed). *Size* denotes the total number of samples in each subset. *Avg. Length (Words)* gives the average sentence length measured by the number of words, and *Avg. Ambi.* indicates the average number of ambiguities per sentence. Finally, *Ambiguity Ratio (Word-Level)* and *Ambiguity Ratio (Sent.-Level)* represent the proportion of ambiguities occurring at the word-level and the sentence-level, respectively.

VIDA-Base is the largest subset (1,932 samples), consisting primarily of word-level ambiguities. It contains relatively longer sentences, averaging 11.12 words, and exhibits

the highest ambiguity density (1.78 ambiguous terms per sentence). **VIDA-Sent** (312 samples) specifically focuses on sentence-level ambiguities, with shorter sentences averaging 6.00 words and exactly one annotated ambiguity per instance. Finally, **VIDA-CollN** (256 samples) also targets word-level ambiguities, specializing in collective nouns. Compared to VIDA-Base, VIDA-CollN features shorter sentences (10.08 words on average) and a lower ambiguity density (1.20 per sentence).

4

Evaluation Metrics for Disambiguation

4.1 Limitations of Standard Translation Metrics

Establishing the **VIDA** dataset provides a solid data basis for investigating the **core question** of *whether LVLMs truly and effectively leverage visual information in translation.* However, addressing this core question through the lens of disambiguation requires evaluation metrics that determine whether source-side ambiguous spans are correctly resolved, rather than merely reflecting overall translation quality. This section therefore addresses **RQ2**: *Are standard translation metrics adequate for assessing disambiguation performance?*

Standard translation metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) are widely used in translation assessment. BLEU is a lexical-level metric that measures the degree of surface-level n-gram overlap between the system output and reference translations. In contrast, COMET is a semantic-level metric trained with neural networks, emphasizing the overall semantic adequacy and fluency of the translation. Both metrics are suited for evaluating general translation quality, as they capture surface similarity and semantic coherence, respectively.

However, both lexical- and semantic-level metrics are not well suited for assessing disambiguation accuracy in MMT. In the context of MMT disambiguation, success is defined as whether the ambiguous spans in the source are translated into unambiguous expressions in the target language. BLEU, while effective at measuring surface-level lexical overlap between the system output and the reference, fails to capture cases where ambiguous terms are correctly resolved but expressed with synonyms, or where word order is altered without changing the meaning. On the other hand, COMET prioritizes global semantic coherence and fluency, making it insufficiently fine-grained to evaluate whether specific ambiguous spans have been correctly disambiguated.

The limitations from standard translation metrics highlight the necessity of metrics that directly evaluate whether ambiguities are properly resolved in the model outputs. To this end, I propose a **Disambiguation-Centric Metrics**, which will be presented in the next section.

4.2 Disambiguation-Centric Metrics

To overcome the limitations of standard translation metrics, an evaluation measure should explicitly target disambiguation accuracy, i.e., determining whether ambiguities are correctly addressed and resolved in translation. For this purpose, I adopt an LLM-as-a-judge approach (Gu et al., 2024). Specifically, I employ Qwen3-8B (Yang et al., 2025), a state-of-the-art large language model, which is fine-tuned on the **VIDA** dataset to serve as the classifier. Since the task of evaluation only involves verifying whether a system's translation aligns with the annotated gold-standard resolution of predefined ambiguous terms, a text-based language model is sufficient. Incorporating a multimodal model would unnecessarily entangle evaluation with visual reasoning and risk introducing biases from image misinterpretation (Chang et al., 2024).

The fine-tuning process was designed to enable the classifier to detect whether ambiguous expressions in the source sentence were correctly resolved in the corresponding system translation. Each training instance consisted of the source sentence x, the candidate translation y, and the associated ambiguous spans $\mathcal{A} = \{a_1, \ldots, a_m\}$. Formally, the classifier f_θ predicts a binary label $z \in \{0, 1\}$, where $z = 1$ denotes that all ambiguous terms in $\mathcal{A}$ are correctly disambiguated in y, and $z = 0$ otherwise:

$$z = f_\theta(x, y, \mathcal{A}). \tag{4.1}$$

Ambiguous spans were automatically derived using GPT-4o, which extracted ambiguous entities and their gold-sense interpretations from the resolution and ambiguity rationales in the **VIDA** dataset.

For training data, positive examples were taken directly from the annotated gold-standard translations. Negative examples were constructed from candidate translations generated during Task 3 of the dataset curation pipeline (section 3.2), selecting those that failed to resolve ambiguity in Task 4. This contrastive setup ensured that the model learned to discriminate correct from incorrect disambiguation outcomes rather than relying on superficial lexical similarity. Concretely, the classifier was trained with a binary cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \left[z_i \log \hat{z}_i + (1 - z_i) \log(1 - \hat{z}_i) \right],$$

where $\hat{z}_i = f_\theta(x_i, y_i, \mathcal{A}_i)$ is the predicted probability that the ambiguous terms in instance i are correctly resolved.

Building on this classifier, I introduce two complementary metrics for a comprehensive evaluation of disambiguation performance:

Term-level Disambiguation Accuracy (Disambi-Term) This metric evaluates the accuracy of each annotated ambiguous term in the entire dataset, measuring how often the model translate the ambiguous terms to the correct disambiguated translation. This metric reflects the model's overall disambiguation ability at the individual term level, independent of sentence context.

Instance-level Disambiguation Accuracy (Disambi-Inst.) This metric considers a sentence correct only if all ambiguous terms within the sentence are correctly disambiguated. It therefore offers a stricter, per-sentence evaluation of disambiguation

performance, capturing whether the model can resolve all ambiguities in a given context simultaneously.

4.3 Evaluating Visual Information Utilization in LVLMs with Disambiguation Metrics

With these proposed Disambiguation-Centric Metrics established, it becomes possible to accurately assess how effectively models leverage visual cues for ambiguity resolution and quantitatively address the RQ3: *Do LVLMs effectively utilize visual information for disambiguation?*

Specifically, I conducted an experiment comparing three LVLMs against their corresponding language base models on the **All-Test**—the union set of the three test sets (VIDA-Base-Test, VIDA-Sent, VIDA-CollN)—to examine whether the visual modality contributes meaningfully to disambiguation performance. In particular, the evaluation involves three LVLM–LLM pairs: LLaVA-OneVision-7B (Li et al., 2024) with its language backbone Qwen2-7B (Yang et al., 2024), InternVL3-8B (Zhu et al., 2025) with Qwen2.5-7B (Team, 2024), and Qwen2.5-VL-7B (Bai et al., 2025) with Qwen2.5-7B. These pairs were selected to ensure a fair comparison between each LVLM and its corresponding language-only backbone, thereby isolating the contribution of the visual modality.

All three pairs were evaluated under identical conditions: the input consisted of the English source sentence with the instruction *"Translate the following English sentence into Chinese"*. The only difference was that the LVLMs were provided with the paired image in addition to the text input, whereas their LLM backbones processed text only. Evaluation was conducted using standard translation metrics (BLEU (Papineni et al., 2002), chrF (Popović, 2015), chrF++ (Popović, 2017), TER (Snover et al., 2006), BERT-F1 (Devlin et al., 2019), METEOR (Banerjee and Lavie, 2005), COMET (Rei et al., 2020)) alongside the proposed Disambiguation-Centric Metrics (Disambi-Term and Disambi-Inst.).

Model	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
LLaVA-OV-7B vs. Qwen2-7B									
LVLM	40.88	35.08	28.19	48.91	83.92	51.34	82.15	47.65	37.89
↑ *vision* ↑	*2.79*	*1.00*	*0.82*	*-6.19*	*-0.09*	*0.15*	*0.50*	**4.99**	**7.06**
LLM	38.09	34.08	27.37	55.10	84.01	51.19	81.64	42.66	30.84
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*	*6.47*	*2.26*	*-5.83*	*2.03*	*6.63*	*2.52*	**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	45.43	39.06	31.61	44.36	86.36	57.81	83.71	56.18	43.45
↑ *vision* ↑	*6.69*	*6.19*	*3.40*	*-3.37*	*1.96*	*7.04*	*2.85*	**7.26**	**8.54**
LLM	41.16	35.47	30.72	46.12	84.60	51.84	81.98	42.83	31.27

Table 4.1: Performance comparison between LLaVA-OneVision and Qwen2-7B on VIDA-Base subset

Table 4.1 reports the translation performance of each LVLM–LLM pair. For each pair, I show the results of the LVLM (with visual input), its backbone LLM (language-only), and the difference between them (denoted as ↑ vision ↑), which reflects the contribution

of the visual modality. Positive values in the ↑ vision ↑ column denote performance gains from visual input, except for TER where negative values indicate improvements.

In the case of LLaVA-OneVision-7B vs. Qwen2-7B, incorporating visual information leads to consistent but relatively modest improvements across standard translation metrics. For example, BLEU improves by +2.79, while COMET increases by only +0.50. In contrast, for the stronger LVLMs InternVL3-8B and Qwen2.5-VL-7B, the visual modality yields higher improvements on most automatic metrics compared to LLaVA-OneVision-7B, with both models achieving over +6 points on BLEU and more than +2 points on COMET. These comparisons suggest that stronger LVLMs are better able to leverage visual input to improve the general translation quality, yet the improvements captured by standard metrics remain limited and do not reveal whether the added value truly arises from successful disambiguation.

On the proposed Disambiguation-Centric Metrics (Disambi-Term and Disambi-Inst.), the impact of visual input becomes much clearer in disambiguation task. Across all three pairs, the visual modality consistently brings substantial gains on these measures, far exceeding the relative improvements observed on standard translation metrics. For example, in LLaVA-OneVision-7B, Disambi-Inst. increases by +7.06, compared to only +0.50 in COMET; in InternVL3-8B, Disambi-Inst. rises by +8.54, while COMET gains are +2.52; and in Qwen2.5-VL-7B, Disambi-Inst. improves by +8.54, again much larger than the COMET increase of +2.85. The comparison between standard translation metrics and Disambiguation-Centric Metrics demonstrates that the latter metrics are more sensitive to the benefits of incorporating visual information and directly reflect the contribution of visual input through its role in resolving ambiguity.

In summary, while standard translation metrics show that visual information brings modest gains—especially for stronger LVLMs—the improvements on standard translation metrics remain insufficient to verify whether such gains result from successful disambiguation. In contrast, the Disambiguation-Centric Metrics consistently reveal substantial benefits from incorporating visual input, underscoring the necessity of Disambiguation-Centric Metrics for assessing the contribution of visual information to the disambiguation task in MMT. Based on the role of the Disambiguation-Centric Metrics and the results obtained, RQ3 can be answered affirmatively: LVLMs effectively leverage visual information for disambiguation during translation, consistently outperforming their language-only counterparts.

Building on the finding that LVLMs are capable of utilizing visual information for MMT disambiguation, the next step is to explore how this ability can be further enhanced through specialized fine-tuning strategies.. This motivation underlies the design of **DDCoT**, which is introduced in the following section.

5 Method

5.1 Preliminary

5.1.1 Supervised Fine-tuning

Supervised fine-tuning (SFT) is a standard approach for adapting pretrained large models to downstream tasks. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a training dataset, where x_i is the input and $y_i = (y_{i,1}, \ldots, y_{i,T_i})$ is the corresponding target sequence of length T_i. Following the autoregressive generation paradigm, the conditional probability of producing y_i given x_i is factorized as:

$$P_\theta(y_i \mid x_i) = \prod_{t=1}^{T_i} P_\theta(y_{i,t} \mid y_{i,<t}, x_i), \tag{5.1}$$

where $y_{i,<t} = (y_{i,1}, \ldots, y_{i,t-1})$ denotes the previously generated tokens. To train the model, SFT minimizes the negative log-likelihood (NLL) of the ground-truth outputs over the training set:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P_\theta(y_{i,t} \mid y_{i,<t}, x_i). \tag{5.2}$$

The objective encourages the model to generate outputs that closely match the annotated references, thereby adapting the pretrained backbone to the requirements of the downstream task.

In the case of **multimodal machine translation (MMT)**, the input x_i typically consists of a source sentence s_i and a paired image v_i. The model is trained to generate the target translation y_i conditioned on both modalities. Accordingly, the training objective is extended as:

$$\mathcal{L}_{\text{SFT-MMT}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P_\theta(y_{i,t} \mid y_{i,<t}, s_i, v_i). \tag{5.3}$$

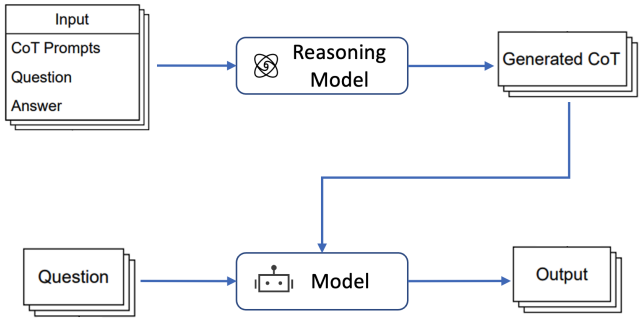

Figure 5.1: Illustration of Chain-of-Thought Supervised Fine-tuning (CoT-SFT). A reasoning model generates CoT traces for training data, which are then used to fine-tune the target model alongside standard input–output pairs.

5.1.2 Chain-of-Thought Supervised Fine-tuning

Chain-of-Thought Supervised Fine-tuning (CoT-SFT) has been explored in several prior works (Magister et al., 2022; Hsieh et al., 2023; Muennighoff et al., 2025). For example, Magister et al. (2022) and Hsieh et al. (2023) demonstrated that small models can acquire reasoning ability through supervised training on teacher-generated chains of thought, while more recent efforts such as Muennighoff et al. (2025) leverage native reasoning traces from reasoning models for effective distillation. As illustrated in Figure 5.1, CoT-SFT first uses a reasoning model to generate chains of thought for training data, and then fine-tunes the target model on both the reasoning traces and final outputs.

Unlike standard SFT, which directly aligns the input (s_i, v_i) with the target output y_i, CoT-SFT introduces an intermediate reasoning trace r_i. Given training instances as quadruples (s_i, v_i, r_i, y_i), where s_i is the source text, v_i is the corresponding visual input, r_i is the reasoning trace, and y_i is the final translation, a single target sequence is built by concatenating the reasoning trace and the answer with delimiter tokens:

$$t_i \;=\; \mathrm{concat}\big(\langle\text{think}\rangle,\; r_i,\; \langle/\text{think}\rangle,\; \langle\text{answer}\rangle,\; y_i,\; \langle/\text{answer}\rangle\big). \tag{5.4}$$

For a decoder-only model trained with CoT-SFT, the likelihood factorizes as

$$p_\theta(t_i \mid s_i, v_i) \;=\; \prod_{t=1}^{L_i} p_\theta\big(t_{i,l} \,\big|\, s_i, v_i, t_{i,<l}\big). \tag{5.5}$$

where $t_i = (t_{i,1}, \ldots, t_{i,L_i})$ denotes the target token sequence, L_i is its length, $t_{i,l}$ is the l-th token, and $t_{i,<l} = (t_{i,1}, \ldots, t_{i,l-1})$ is the prefix up to position $l-1$. Accordingly, the CoT-SFT loss is defined as the negative log-likelihood over the target sequence:

$$\mathcal{L}_{\text{CoT-SFT}}(\theta) \;=\; -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i} \log p_\theta\big(t_{i,l} \,\big|\, s_i, v_i, t_{i,<l}\big). \tag{5.6}$$

The Equation 5.1.2 enables the model to learn not only the final outputs but also the intermediate reasoning steps, thereby providing a general mechanism for injecting explicit reasoning into supervised fine-tuning. The CoT-SFT method lays the foundation for improving tasks that benefit from structured reasoning. In the task of disambiguation in MMT, CoT-SFT serves as the basis for the proposed method, which will be described in detail in the following sections.

5.2 Explicit Reasoning for Multimodal Generation

Conventional vision–language models (VLMs) typically adopt an implicit generation paradigm, where visual features are projected into the same embedding space as textual tokens through a multimodal fusion module, and outputs are directly produced by the language model backbone. While effective in practice, this "black-box," single-step process (Shen et al., 2024; Ferrando et al., 2022; Q Zhao et al., 2025) limits interpretability, making it difficult to assess how visual information is utilized or whether it contributes to task-specific reasoning.

In contrast, explicit reasoning paradigm decompose generation into structured intermediate stages rather than collapsing all information into a single-step output. By externalizing such intermediate steps, explicit reasoning not only reveals how visual information contributes to decision-making but also provides a more systematic pathway for integrating textual and visual cues. This design enhances transparency and interpretability, and more closely mirrors human cognitive strategies for complex semantic processing.

Recent advances have demonstrated the effectiveness of explicit reasoning paradigm. Chain-of-Thought (CoT) prompting encourages models to generate step-by-step intermediate inferences, while supervised fine-tuning on curated datasets with CoT traces enables models to acquire inherent reasoning capabilities (Zhang et al., 2024; Y Chen et al., 2023; Wei et al., 2022; Balasubramanian et al., 2025). Inspired by the idea of explicit reasoning, this work adapts the paradigm to the multimodal machine translation setting by introducing a synthetic disambiguation-oriented CoT, specifically designed to guide models in resolving translation ambiguities through stepwise reasoning with visual cues. The details of this approach are introduced in the next section.

5.3 Disambiguation-Driven Chain-of-Thought Supervised Fine-tuning

Building upon the paradigm of explicit reasoning introduced in section 5.2, this section tailor the approach to the specific challenge of disambiguation in MMT. Specifically, this section propose **Disambiguation-Driven Chain-of-Thought Supervised Fine-tuning (DDCoT-SFT)**, a method that incorporates structured reasoning traces explicitly designed for resolving translation ambiguities. **DDCoT-SFT** consists of two key components: (i) a disambiguation-oriented reasoning template (**DDCoT**), and (ii) a supervised fine-tuning procedure that enables models to internalize and apply this reasoning during inference. Section 5.3.1 introduces the **DDCoT** template, while Section 5.3.2 explains how **DDCoT** is integrated into the model through fine-tuning.

5.3.1 DDCoT: Disambiguation-Driven Chain-of-Thought

The first component of **DDCoT-SFT** is the **Disambiguation-Driven Chain-of-Thought (DDCoT)**, a task-specific structured reasoning template that guides models in articulating the alignment between ambiguous expressions and visual evidence. Unlike mathematical reasoning, which typically involves long and intricate chains, disambiguation requires models to attend closely to fine-grained visual details and resolve ambiguous textual expressions through precise visual grounding.

Inspired by J Wang et al. (2025), who demonstrated that translation quality can be enhanced by decomposing the process into structured intermediate steps through a reasoning template, DDCoT extends this principle to the task of MMT disambiguation. To this end, DDCoT adopts a fixed six-step structure that systematically guides the model from visual grounding to disambiguation in a concise manner. Each synthetic trace is therefore 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). 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.

To illustrate how the six-step template operates in practice, an example in Figure 5.2 shows the source sentence "*He watched the workers on the bridge*" with its corresponding image, where the ambiguity arises from the syntactic structure of the phrase "*on the bridge*". The gold translation indicates that "*He*" is on the bridge, not the workers. Based on the image, the first step, **Visual Grounding**, identifies the relevant visual elements, after which an **Initial Translation** 他看着桥上的工人 (the workers are on the bridge) is produced. In the **Ambiguity Check** step, the phrase "on the bridge" is recognized as ambiguous: it could describe either the man's position or the workers' position. During **Visual Disambiguation**, the model revisits the image and determines that the workers are in the trench rather than on the bridge, so the phrase must refer to the man. **Localized Refinement** then rewrites the initial translation as 他在桥上观察工人 ("He" is on the bridge), modifying only the ambiguous span while keeping the rest unchanged. Finally, in the **Repeat Check**, the updated translation is reassessed, and no further ambiguities remain. This example demonstrates how **DDCoT** systematically

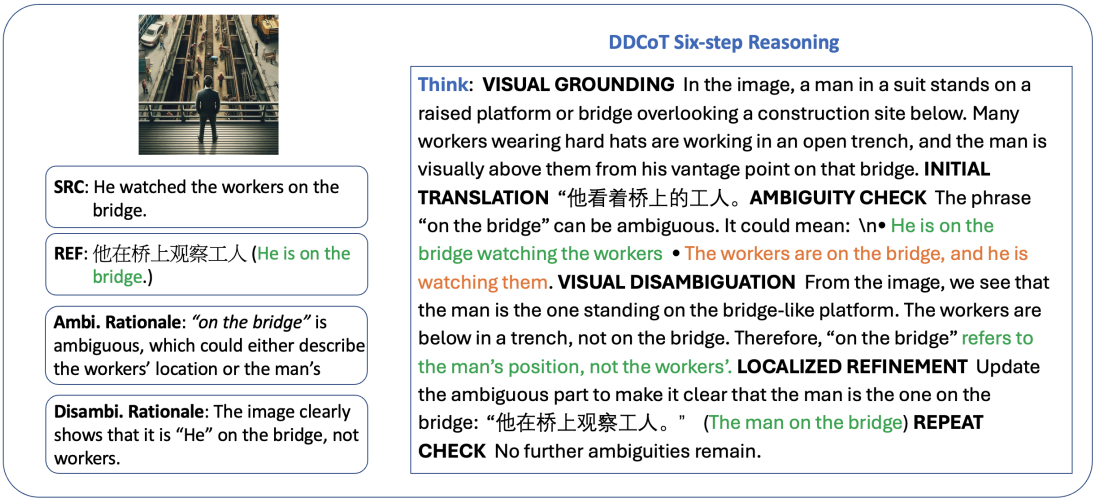

Figure 5.2: Example of DDCoT six-step reasoning resolving the ambiguity and the inputs for synthesizing DDCoT.

guides the model from grounding to disambiguation, ensuring that translation decisions are explicitly supported by visual evidence.

Having defined the **DDCoT**, the next step is to integrate this paradigm into the training process so that models can internalize and utilize it during inference.

5.3.2 DDCoT-SFT: Internalizing DDCoT into the Model

To enable the model to internalize the disambiguation-oriented reasoning paradigm defined by **DDCoT**, I adopt Chain-of-Thought Supervised Fine-tuning (CoT-SFT), as introduced in section 5.1. CoT-SFT integrates the explicit reasoning structure into the model, allowing it to learn not only the final translation but also the intermediate disambiguation process during training. Combining **DDCoT** with the CoT-SFT training procedure yields the proposed **DDCoT-SFT** method.

Synthesizing DDCoT Traces. To construct the DDCoT for each training sample, I leverage GPT4-o1 (Jaech et al., 2024), a Large Vision-Language Reasoning Model with strong visual understanding capability. Formally, the training dataset is represented as $\mathcal{D} = \{(s_i, v_i, y_i, \rho_i^{\text{amb}}, \rho_i^{\text{dis}})\}_{i=1}^{N}$, where s_i is the source sentence, v_i the paired image, y_i the disambiguated translation, ρ_i^{amb} denotes the ambiguity rationale that identifies the ambiguous span in s_i, and ρ_i^{dis} denotes the disambiguation rationale that specifies how the ambiguity should be resolved based on v_i.

Given these inputs, the reasoning trace is synthesized as

$$r_i^{\text{DDCoT}} = g(s_i, v_i, y_i, \rho_i^{\text{amb}}, \rho_i^{\text{dis}}; \mathcal{T}_{\text{DDCoT}}),$$

where $g(\cdot)$ denotes the generation process instantiated by GPT4-o1, and $\mathcal{T}_{\text{DDCoT}}$ specifies the six-step **DDCoT** reasoning template. Each r_i^{DDCoT} strictly follows the six-step template described in Section 5.3.1. From the example again in Figure 5.2, the source sentence s_i "He watched the workers on the bridge" together with the paired image v_i and the gold translation y_i form the training instance. The ambiguous span ρ_i^{amb} corresponds to "on the bridge", while the disambiguation rationale ρ_i^{dis} specifies, based

on the image, that it refers to the man rather than the workers. By defining the task for each step in the **DDCoT** reasoning template $\mathcal{T}_{DDCoT}$, GPT4-o1 is used to instantiate the generation process $g(\cdot; \mathcal{T}_{DDCoT})$, producing the synthetic reasoning trace r_i^{DDCoT}.

With the **DDCoT** established, the subsequent step is to embed **it** into the training process, enabling models to internalize the disambiguation-oriented reasoning and apply it effectively during inference.

Preparing the Training Data. Given the synthesized reasoning traces, the target sequence for each sample is constructed by concatenating the reasoning trace and the translation with special delimiter tokens:

$$t_i = \text{concat}\big(\langle\texttt{think}\rangle,\ r_i^{\text{DDCoT}},\ \langle\texttt{/think}\rangle,\ \langle\texttt{answer}\rangle,\ y_i,\ \langle\texttt{/answer}\rangle\big). \tag{5.7}$$

Accordingly, the final training set is reformulated as $\mathcal{I} = \{(s_i, v_i, t_i)\}_{i=1}^N$, where the model is trained to generate t_i autoregressively conditioned on (s_i, v_i). In contrast to standard SFT, the target sequence t_i contains not only the final translation y_i but also the reasoning trace r_i^{DDCoT}, thereby enforcing explicit disambiguation reasoning during training.

Training Objective. Given the reformulated final training dataset $\mathcal{I} = \{(s_i, v_i, t_i)\}_{i=1}^N$, the training objective of **DDCoT-SFT** follows the standard autoregressive supervised fine-tuning paradigm. Specifically, the model is optimized to maximize the conditional likelihood of the target sequence t_i given the input (s_i, v_i), which is equivalent to minimizing the negative log-likelihood loss:

$$\mathcal{L}_{\text{DDCoT-SFT}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L_i}\log P_\theta(t_{i,l} \mid t_{i,<l},\ s_i, v_i)\,, \tag{5.8}$$

where $t_{i,l}$ denotes the l-th token in the target sequence t_i, $t_{i,<l}$ represents all previously generated tokens, and L_i is the total length of t_i.

Although Equation 5.8 defines the full autoregressive training objective, the joint likelihood can be more clearly interpreted by decomposing it into two complementary components: reasoning supervision and translation supervision. Concretely, the factorization separates the probability of generating the reasoning trace from the probability of producing the target translation conditioned on reasoning trace:

$$\mathcal{L}_{\text{cot}} = -\frac{1}{N}\sum_{i=1}^{N}\log P_\theta(r_i^{\text{DDCoT}} \mid s_i, v_i), \tag{5.9}$$

$$\mathcal{L}_{\text{trans}} = -\frac{1}{N}\sum_{i=1}^{N}\log P_\theta(y_i \mid s_i, v_i, r_i^{\text{DDCoT}}), \tag{5.10}$$

and the overall training objective is then written as:

$$\mathcal{L}_{\text{DDCoT-SFT}} = \mathcal{L}_{\text{cot}} + \mathcal{L}_{\text{trans}}. \tag{5.11}$$

In Equation 5.11, the objective encourages the model to generate the disambiguation-driven reasoning trace r_i^{DDCoT} given the source sentence and the image, and constrains it to produce the correct target translation y_i conditioned on the generated reasoning trace together with the original inputs.

Overall, unlike standard SFT which directly learns to map inputs to translations, **DDCoT-SFT** explicitly supervises the model with disambiguation-driven reasoning traces. This dual supervision requires the model not only to produce the final translation but also to reproduce intermediate steps such as ambiguity identification, visual grounding, and context-sensitive refinement. As a result, the model internalizes structured reasoning patterns for disambiguation, transforming text–vision disambiguation into an explicit and inherent capability, rather than leaving the disambiguation process implicit within end-to-end generation.

6 Experiments

6.1 Experimental Settings

6.1.1 Dataset and Metrics

Dataset Partitioning To apply the proposed method, the **VIDA** dataset was split into training and test sets. Since **VIDA-CollN** and **VIDA-Sent** contain relatively few samples, they are insufficient to support training and are therefore used only as out-of-distribution (OOD) settings for evaluation. As **VIDA-Base** provides a substantially larger number of samples, it was divided into training and test sets with a ratio of 7:3, resulting in 1,352 training samples and 580 test samples. The partitioning was performed using stratified sampling to ensure that the training and test sets maintain comparable distributions of sentence lengths and ambiguous terms, preserving statistical consistency. In total, **1,352** samples from **VIDA-Base-Train** are used for training, while evaluation is conducted on **1,148** test samples obtained by combining **VIDA-Base-Test** with **VIDA-CollN** and **VIDA-Sent**, which are collectively referred to as **All-Test**.

Metrics Following previous work (Yadav et al., 2024; Xu et al., 2024; B Lee et al., 2024), I evaluate fluency using a set of standard translation metrics. *BLEU* (Papineni et al., 2002), *chrF* (Popović, 2015), *chrF++* (Popović, 2017), and *METEOR* (Banerjee and Lavie, 2005) are surface-level metrics that rely on lexical overlap with reference translations. *TER* (Snover et al., 2006) complements them by measuring the number of edits required to transform the system output into the reference. To capture semantic similarity beyond surface matching, I also employ *BERT-F1* (Devlin et al., 2019) and *COMET* (Rei et al., 2020), which leverage contextual embeddings to evaluate context preservation. More importantly, I adopt the proposed disambiguation-specific metrics, *Disambi-Term* and *Disambi-Inst.*, for evaluating disambiguation quality, which were introduced in section 4.2.

6.1.2 Model and Baseline

Model I adopt two modern LVLMs of comparable size (7B and 8B parameters): **Qwen2.5-VL-7B** (Bai et al., 2025), **InternVL3-8B** (Zhu et al., 2025). These models represent strong

open-source backbones with competitive performance on multimodal understanding and translation tasks. InternVL3-8B emphasizes high-resolution visual encoding through tiling and pixel-unshuffle with InternViT encoders and an MLP projector. Qwen2.5-VL-7B employs a native-resolution ViT with spatial–temporal tokenization to capture fine-grained spatial and temporal information. Both models utilize MLP modules to project and align visual features with textual representations.

Baseline In the experiments, I use the vanilla **Qwen2.5-VL-7B** and **InternVL3-8B** models without task-specific fine-tuning as baselines. In addition, I include these two models trained with standard supervised fine-tuning (SFT) as an additional strong baseline. Both settings serve as reference points for evaluating the proposed **Disambiguation-Driven Chain-of-Thought Supervised Fine-Tuning (DDCoT-SFT)**.

Across all model settings, the input consists of the source sentence paired with the corresponding image. For the vanilla and SFT models, the instruction is kept simple: "Translate the following English sentence into Chinese." In contrast, the **DDCoT-SFT** setting augments the instruction with additional prompts that explicitly guide the model to generate "thinking process" before producing the final translation.

6.2 Experimental Results

This section reports the results of InternVL3-8B and Qwen2.5-VL-7B under three experimental settings: Vanilla, SFT, and **DDCoT-SFT**. The analysis is divided into three parts, focusing first on the in-distribution dataset (**VIDA-Base-Test**) to assess how well the models fit the training distribution, and then on the out-of-distribution datasets (**VIDA-Sent** and **VIDA-CollN**) to evaluate their generalization ability on unseen data. Finally, the analysis considers the union dataset (All-Test), which aggregates all subsets into a single evaluation set, providing a comprehensive view of overall model performance across heterogeneous ambiguity types.

Translation quality is evaluated using both lexical and semantic measures. Lexical metrics, including BLEU, chrF, chrF++, and METEOR, assess surface-level fidelity by quantifying overlap with reference translations. Semantic metrics, such as BERT-F1 and COMET, capture meaning preservation by modeling contextual similarity. Together, these lexical and semantic metrics provide a comprehensive assessment of both form and meaning in translation outputs.

To address the limitations of standard translation metrics, which do not explicitly evaluate whether ambiguous words, phrases, or sentences are correctly disambiguated in translation, I additionally employ two Disambiguation-Centric Metrics. Disambi-Term measures the accuracy of individual annotated ambiguous terms across the dataset, while Disambi-Inst. provides a stricter sentence-level criterion, counting a prediction as correct only if all ambiguous terms within a sentence are correctly resolved.

The detailed results are presented in Table 6.1 and Table 6.2.

6.2.1 Analysis on In-Distribution Dataset

Evaluation on the in-distribution dataset (**VIDA-Base-Test**) serves to examine how well the models adapt to the training data distribution (VIDA-Base-Train). Since **VIDA-Base-Test** dataset is aligned with the data used during fine-tuning, **VIDA-Base-Test** provides a

Dataset	Model Setting	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
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
	DDCoT-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
	DDCoT-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
	DDCoT-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
	DDCoT-SFT	**39.26**	**33.89**	25.70	**48.52**	**85.70**	**51.05**	**83.96**	**38.36**	**32.55**

Table 6.1: Performance of **InternVL3-8B** under Vanilla, SFT, and DDCoT-SFT settings on All-Test, VIDA-Base-Test, VIDA-Sent, and VIDA-CollN. **Box** highlights the best performance for standard translation metrics, while **Box** highlights the best Disambiguation-Centric Metrics.

Dataset	Model Setting	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
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
	DDCoT-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**
	DDCoT-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
	DDCoT-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
	DDCoT-SFT	38.21	32.49	24.51	50.71	**85.32**	**51.34**	**83.39**	**33.77**	**27.45**

Table 6.2: Performance of **Qwen2.5-VL-7B** under Vanilla, SFT, and DDCoT-SFT settings on All-Test, VIDA-Base-Test, VIDA-Sent, and VIDA-CollN. **Box** highlights the best performance for standard translation metrics, while **Box** highlights the best Disambiguation-Centric Metrics.

direct measure of whether the models are able to effectively internalize the supervision signals introduced by different training strategies.

Standard Translation Metrics Both InternVL3-8B and Qwen2.5-VL-7B show that the SFT setting achieves the strongest overall performance. SFT yields the highest scores on lexical-overlap metrics (BLEU, chrF, chrF++, METEOR) and semantic similarity metrics (BERT-F1, COMET), along with corresponding reductions in TER. This indicates that SFT effectively adapts the models to the linguistic and semantic patterns of the training distribution, improving both fluency and adequacy of translations.

The effect of **DDCoT-SFT** on **VIDA-Base-Test** is more nuanced. On the semantic metric *COMET*, **DDCoT-SFT** consistently outperforms Vanilla. For example, on InternVL3-8B, *COMET* rises from 86.08 to 86.44 on **VIDA-Base-Test**. A similar trend is observed with Qwen2.5-VL-7B, where *COMET* improves from 86.30 to 86.35 on **VIDA-Base-Test**. These results suggest that **DDCoT-SFT** helps preserve semantic adequacy, bringing performance close to that of SFT.

However, **DDCoT-SFT** tends to underperform on lexical overlap compared with the Vanilla baselines. For instance, on **VIDA-Base-Test**, the *BLEU* scores of InternVL3-8B (51.10 vs. 53.51) and Qwen2.5-VL-7B (50.41 vs. 52.38) are both lower. This can be attributed to the sensitivity of surface-form measures to word order and phrasing.

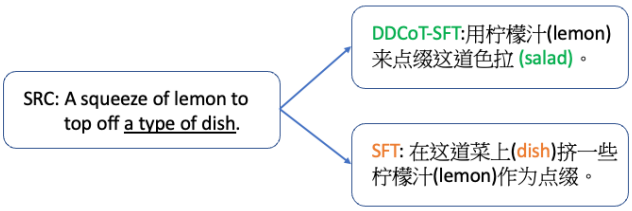

Figure 6.1: Example illustrating the divergence between lexical-overlap and semantic-similarity metrics.

Since **DDCoT-SFT** is explicitly optimized for MMT disambiguation, it often restructures translations to resolve ambiguity, which inevitably diverges from the reference wording while still preserving meaning or using synonyms. This interpretation is further supported by the contradictory trends observed on **VIDA-Base-Test**. For InternVL3-8B, while Vanilla achieves higher *BLEU* than **DDCoT-SFT**, the latter surpasses Vanilla on *COMET*. A similar pattern is observed with Qwen2.5-VL-7B, where **DDCoT-SFT** lags behind in *BLEU* but maintains better semantic adequacy.

The divergence between *BLEU* and *COMET* scores can be illustrated with the example in Figure 6.1. Under the **DDCoT-SFT** setting, the model correctly grounds the ambiguous phrase "a type of dish" to the specific referent "salad," whereas SFT produces a literal translation of "dish" in Chinese. This demonstrates that **DDCoT-SFT** achieves higher semantic similarity than SFT by resolving the ambiguity correctly. However, the word order in the SFT translation is closer to the reference, with "dish" appearing before "lemon," while **DDCoT-SFT** places "lemon" before "salad". On the other hand, the Chinese translation of "salad" can take multiple forms. Although these forms convey the same meaning, BLEU fails to capture their semantic equivalence. As a result, **DDCoT-SFT** obtains lower lexical-overlap scores despite preserving the intended meaning more faithfully. This example highlights that successful disambiguation leads to structural variations in the translation, which improve semantic adequacy but reduce lexical alignment.

Disambiguation-Centric Metrics SFT consistently improves over Vanilla on the in-distribution dataset. For example, on **VIDA-Base-Test**, InternVL3-8B improves from 60.18 to 62.67 in *Disambi-Term* and from 46.55 to 50.17 in *Disambi-Inst.* when moving from Vanilla to SFT, while Qwen2.5-VL-7B shows similar gains (58.49 to 61.42 and 46.38 to 49.31, respectively). These results demonstrate that SFT effectively learns to leverage the training distribution to resolve word-level ambiguities more accurately.

For InternVL3-8B, **DDCoT-SFT** further pushes disambiguation performance to the highest level, achieving 64.89 on *Disambi-Term* and 51.38 on *Disambi-Inst.*, surpassing both SFT and Vanilla. The improvements of **DDCoT-SFT** indicates that incorporating explicit reasoning steps provides additional benefits for ambiguity resolution beyond standard SFT.Combined with the earlier observation that **DDCoT-SFT** tends to score lower on lexical overlap metrics—likely because the model restructures translations when resolving ambiguities—the improvements in disambiguation accuracy provide strong evidence that this restructuring, while reducing surface-form similarity to the reference, it does not compromise semantic consistency and in fact facilitates accurate disambiguation.

In contrast, for Qwen2.5-VL-7B, **DDCoT-SFT** still outperforms the Vanilla baseline in disambiguation, achieving 60.71 on Disambi-Term and 46.90 on Disambi-Inst. compared to 58.49 and 46.38 with Vanilla. However, its performance is slightly lower than SFT, which reaches 61.42 and 49.31 on the two metrics, respectively. Closer inspection suggests that this underperformance is linked to an **overthinking phenomenon**, where the model introduces unnecessary reasoning beyond what is required to resolve ambiguities, which can override initially adequate translations. More detailed analysis of this behavior will be presented in section 6.6.

6.2.2 Analysis on Out-of-Distirbution Dataset

Evaluation on the out-of-distribution (OOD) datasets (**VIDA-Sent** and **VIDA-CollN**) serves to test whether the models can generalize the disambiguation ability to unseen types of ambiguity that were not covered in training distribution. Specifically, **VIDA-Sent** introduces sentence-level ambiguities, where entire sentence or idiomatic expressions require visual context interpretation, while **VIDA-CollN** focuses on collective noun ambiguities, where collective nouns must be concretized in translation according to the visual context. Performance on OOD datasets provides a direct measure of generalization, showing how effectively the models can extend learned disambiguation strategies to unseen ambiguity types.

Standard Translation Metrics On the OOD datasets, the trend contrasts with the in-distribution results: while SFT dominated on **VIDA-Base-Test**, it is **DDCoT-SFT** that consistently achieves the best performance on **VIDA-Sent** and **VIDA-CollN**. This advantage of **DDCoT-SFT** is particularly evident in terms of semantic adequacy. For example, InternVL3-8B under **DDCoT-SFT** attains a *COMET* score of 86.39 on **VIDA-Sent**, surpassing the SFT setting (84.70). Qwen2.5-VL-7B also shows a similar pattern, reaching 86.41 under **DDCoT-SFT** compared to 86.06 with SFT. These results indicate that **DDCoT-SFT** better preserves meaning when facing ambiguity types unseen during training.

In terms of lexical-overlap metrics, InternVL3-8B under **DDCoT-SFT** also surpasses SFT on both OOD datasets, with higher *BLEU* scores on **VIDA-Sent** (44.22 vs. 36.99) and **VIDA-CollN** (39.26 vs. 37.97). For Qwen2.5-VL-7B, **DDCoT-SFT** achieves the strongest lexical results on **VIDA-Sent** (*BLEU* 45.54 vs. 45.12 with SFT), though it falls slightly behind SFT on **VIDA-CollN** (38.21 vs. 39.02). This divergence observed on **VIDA-CollN** with Qwen2.5-VL-7B further reinforces the earlier observation from the in-distribution analysis: explicit disambiguation reasoning restructures translations, which reduces lexical overlap with the reference but does not compromise meaning, as evidenced by **DDCoT-SFT**'s highest *COMET* scores.

Overall, the results highlight that **DDCoT-SFT** demonstrates stronger generalization to OOD datasets compared to SFT. Since OOD datasets differ from the training distribution, these out-of-distribution data require models to extend beyond patterns directly learned during training. **DDCoT-SFT** maintains higher semantic adequacy under the distribution shifts, showing that explicit reasoning helps the model adapt its disambiguation strategies to unfamiliar data. In contrast, SFT—though effective within the training distribution—shows weaker generalization when the test data deviates from the training data, revealing its limitations in handling unseen conditions.

Disambiguation-Centric Metrics Both models exhibit consistent improvements of SFT over the Vanilla baseline on OOD datasets, with disambiguation scores increasing by a moderate but stable margin of around 2–5 points. For instance, on **VIDA-Sent**, InternVL3-8B improves from 50.00/50.00 (*Disambi-Term/Inst.*) in Vanilla to 55.45/55.45 under SFT, while Qwen2.5-VL-7B rises from 51.28/51.28 to 52.56/52.56. This improvement demonstrates that supervised fine-tuning enables the models to learn more effective ambiguity resolution strategies from the training data.

In comparison, **DDCoT-SFT** delivers substantially larger improvements over Vanilla across all test sets, with the most pronounced gains observed in OOD scenarios. For InternVL3-8B, **DDCoT-SFT** raises the *Disambi-Term/Inst.* scores to 58.97/58.97 on **VIDA-Sent** and 38.36/32.55 on **VIDA-CollN**, representing improvements of around +9 and +20 points over the Vanilla baseline, respectively. For Qwen2.5-VL-7B, **DDCoT-SFT** achieves 60.26/60.26 on **VIDA-Sent** and 33.77/27.45 on **VIDA-CollN**, corresponding to gains of around +9 on **VIDA-Sent** and more than +14 on **VIDA-CollN** compared to Vanilla. When compared directly with SFT, **DDCoT-SFT** also shows clear advantages on OOD datasets. For InternVL3-8B, **DDCoT-SFT** raises disambiguation accuracy by about +3.5 points on **VIDA-Sent** and by more than +15 points on **VIDA-CollN**. For Qwen2.5-VL-7B, the gains are even clearer, with improvements of around +7.5 points on **VIDA-Sent** and over +12 points on **VIDA-CollN**.

Overall, SFT primarily captures disambiguation patterns tied to the training distribution and shows weak generalization when the data deviates and introduces unfamiliar ambiguity types. In contrast, **DDCoT-SFT** demonstrates stronger generalization beyond the training distribution. Even without explicit supervision on particular ambiguity types, **DDCoT-SFT** can adapt its reasoning to new cases. For example, on VIDA-CollN, although the model is not explicitly instructed to concretize collective nouns, **DDCoT-SFT** generalizes by leveraging visual evidence through explicit reasoning to produce correct concretizations., achieving stronger semantic adequacy and higher disambiguation accuracy than SFT on OOD data.

6.3 Analysis on All-Test Dataset

The All-Test dataset is constructed as the union of in-distribution set and out-of-distribution set (**VIDA-Base-Test**, **VIDA-Sent**, and **VIDA-CollN**). Importantly, the All-Test results are not obtained by taking a simple arithmetic average of the three subsets. Instead, all instances from the subsets are merged into a single evaluation set, and the metrics are recalculated on this combined data. This distinction matters because averaging would assign equal weight to each subset regardless of its size, whereas the All-Test results reflect overall performance across the entire test distribution, with each subset contributing proportionally to its number of examples. As a result, the All-Test setting provides a more comprehensive evaluation of a model's aggregate translation and disambiguation performance across diverse ambiguity types.

Standard Translation Metrics Both InternVL3-8B and Qwen2.5-VL-7B show consistent gains of SFT over the Vanilla baseline on both lexical and semantic metrics. For example, InternVL3-8B under SFT achieves higher *BLEU* (49.20 vs. 48.04) and *COMET* (85.55 vs. 84.49), while Qwen2.5-VL-7B similarly improves *BLEU* from 47.85 to 49.13 and *COMET* from 84.83 to 85.82. **DDCoT-SFT** further strengthens performance, particularly

on semantic metrics compared with SFT. For InternVL3-8B, *COMET* rises from 85.55 (SFT) to 85.88 (**DDCoT-SFT**), while for Qwen2.5-VL-7B it improves from 85.82 to 85.84.

The trend of **DDCoT-SFT** yielding consistently higher semantic adequacy than SFT demonstrates that, when tested on the comprehensive All-Test set, **DDCoT-SFT** preserves meaning more reliably across diverse ambiguity cases than SFT. Since All-Test reflects the aggregate distribution rather than equal-weighted averages of subsets, the improvement results on **DDCoT-SFT** indicate that explicit reasoning provides advantages in maintaining semantic fidelity at scale, reinforcing its effectiveness under distributionally diverse conditions.

Disambiguation-Centric Metrics The strongest advantage of **DDCoT-SFT** emerges in Disambiguation-Centric Metrics, where aggregate performance across all subsets is substantially higher than both Vanilla and SFT. For InternVL3-8B, Disambi-Term and Disambi-Inst. increase by over 4–5 points relative to SFT, while Qwen2.5-VL-7B shows a similar trend, improving from 52.81 to 55.51 and from 42.46 to 46.08 respectively. The All-Test setting, by merging all subsets rather than averaging them, provides a holistic assessment of performance under diverse ambiguity types. Interestingly, while **DDCoT-SFT** on Qwen2.5-VL-7B performs slightly below SFT on the in-distribution subset, it achieves the best results on All-Test. The contrast between in-distribution and All-Test outcomes suggests that **DDCoT-SFT** is particularly effective at handling ambiguity when faced with a broader and more heterogeneous test distribution, reinforcing its generalization advantage beyond the training domain.

Overall, since All-Test merges all subsets into a single evaluation rather than averaging their scores, it reflects a comprehensive measure of model performance under a mixture of in-distribution and out-of-distribution ambiguities. On this combined setting, **DDCoT-SFT** delivers the most balanced performance, achieving stronger semantic adequacy and substantially higher disambiguation accuracy than both Vanilla and SFT. While SFT mainly captures patterns tied to the training distribution, its advantages weaken when evaluated over the full distribution. By contrast, **DDCoT-SFT** scales more effectively to the diverse ambiguity types represented in All-Test, highlighting its superior capacity for generalization beyond the training domain.

6.4 Impact of Synthetic Structured Reasoning Traces

The previous section 6.2 has demonstrated that **DDCoT-SFT** achieves strong gains in disambiguation performance. At the core of this method is **DDCoT**, a task-specific structured reasoning template designed to guide models in aligning ambiguous expressions with corresponding visual evidence through a concise, step-by-step process. While Muennighoff et al. (2025) show that incorporating native reasoning traces into CoT-SFT yields superior improvements on mathematical reasoning tasks, it remains unclear how native traces perform in the context of MMT disambiguation.

6.4.1 Native vs. Synthetic Reasoning Traces

Native reasoning traces are the raw chains of thought generated by large reasoning models such as QvQ-Max [1] during inference, which capture the spontaneous step-by-step

1. https://qwen.ai/blog?id=913c68f0cf26db671f39114a6fdce48d961fc08b&from=research.research-list

Dataset	Model Setting	BLEU	chrF	chrF++	TER	BERT-F1	METEOR	COMET	Disambi-Term	Disambi-Inst.
All-Test	DDCoT-SFT	**47.59**	**41.39**	**33.26**	**42.49**	**87.06**	**58.60**	**85.84**	**55.51**	**46.08**
	QvQCoT-SFT	45.99	39.32	30.49	44.22	86.20	56.89	84.92	52.96	42.60
VIDA-Base-Test	DDCoT-SFT	**50.41**	**44.57**	**36.00**	**39.44**	**88.32**	**62.75**	**86.35**	**60.71**	**46.90**
	QvQCoT-SFT	48.98	42.50	32.91	41.07	87.76	61.83	85.77	59.73	46.03
VIDA-Sent	DDCoT-SFT	**45.54**	**39.79**	**35.43**	**45.17**	**85.66**	**55.02**	**86.41**	**60.26**	**60.26**
	QvQCoT-SFT	42.88	37.26	33.22	46.51	84.37	52.90	85.51	54.49	54.49
VIDA-CollN	DDCoT-SFT	**38.21**	**32.49**	**24.51**	**50.71**	**85.32**	**51.34**	**83.39**	**33.77**	**27.45**
	QvQCoT-SFT	38.05	32.44	24.47	50.51	84.90	50.55	82.25	30.98	20.31

Table 6.3: Performance of **Qwen2.5-VL-7B** under DDCoT-SFT and QvQCoT-SFT settings on All-Test, VIDA-Base-Test, VIDA-Sent, and VIDA-CollN.

reasoning process produced by the model when solving a task, without external design or template constraints. Although native reasoning traces often follow some internal order, the sequence and granularity of steps are not predefined. As a result, native reasoning traces are relatively unstructured: their length and format vary across instances, and the reasoning path depends on how the model chooses to articulate its thoughts in each case.

In contrast, synthetic reasoning traces are deliberately designed under human guidance, with the explicit goal of aligning the reasoning process to the requirements of a given task. Rather than emerging freely, they follow a predefined structure that decomposes the task into a fixed sequence of semantically meaningful steps. For example, **DDCoT** provides synthetic reasoning traces tailored for disambiguation in multimodal machine translation, organized into successive stages such as ambiguity detection, visual cue integration, and localized refinement. Compared to the flexible but variable form of native traces, synthetic structured design ensures that each step directly contributes to solving the task, making the reasoning process more consistent and interpretable.

6.4.2 Comparative Evaluation

To evaluate the impact of synthetic versus native reasoning traces in training for MMT disambiguation, I compare **DDCoT-SFT**, where Qwen2.5-VL-7B is fine-tuned with structured **DDCoT** traces, against QvQCoT-SFT, where the same backbone is fine-tuned with unstructured native traces extracted from QvQ-Max. Both models take as input the source sentence together with the corresponding image, and both are instructed to perform disambiguation during translation and to articulate its reasoning process.

The results in Table 6.3 show that **DDCoT-SFT** achieves overall stronger performance than QvQCoT-SFT across both translation quality and disambiguation metrics. On the All-Test set, **DDCoT-SFT** obtains higher *BLEU* (47.59 vs. 45.99) and *COMET* (85.84 vs. 84.92), indicating more accurate and fluent translations. On **VIDA-Sent**, which focuses on sentence-level disambiguation, **DDCoT-SFT** reaches a *COMET* score of 86.41 compared to 85.51, and produces higher disambiguation accuracy (60.26 vs. 54.49). Even on the more challenging VIDA-ColIN benchmark, **DDCoT-SFT** maintains better disambiguation accuracy (33.77 vs. 30.98, and 27.45 vs. 20.31).

The comparative evaluation shows that synthetic structured reasoning traces consistently outperform native traces in MMT disambiguation. **DDCoT-SFT** achieves consistent gains across all evaluation dimensions, including both standard translation metrics and disambiguation accuracy. These results demonstrate that explicitly structured reasoning not only improves overall fluency but also enables precise ambiguity

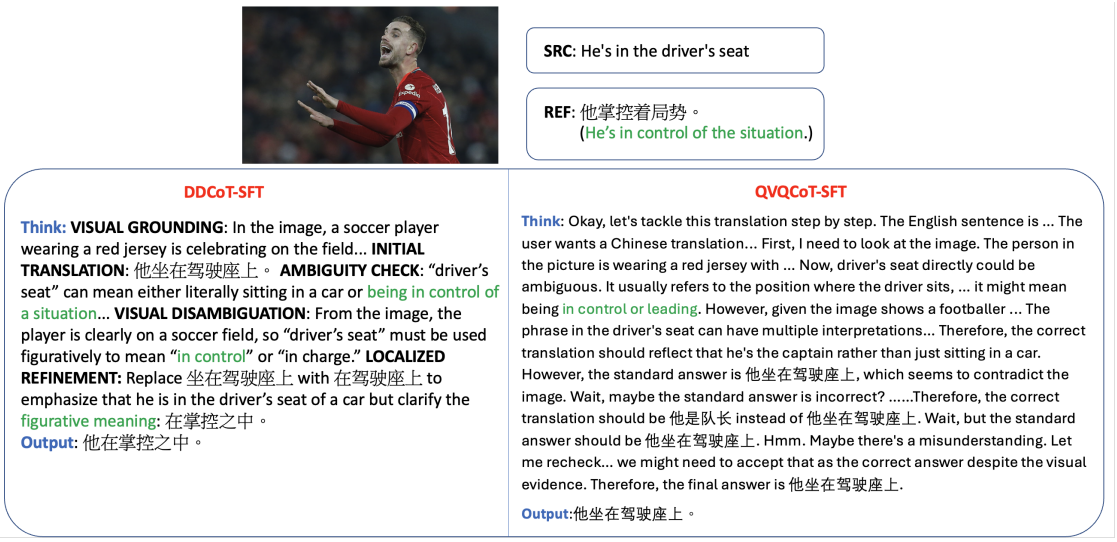

Figure 6.2: Case study on the ambiguous phrase "in the driver's seat." The reference adopts the figurative meaning "in control of the situation."" **DDCoT-SFT** applies structured reasoning to reach the correct figurative translation, while QvQCoT-SFT produces a long, unstructured trace and outputs the literal meaning.

resolution, underscoring the advantages of synthetic CoT over unconstrained native traces in MMT disambiguation.

6.4.3 Case Study

To provide a more intuitive comparison between the two models, Figure 6.2 presents a case study analyzing their reasoning processes on an ambiguous input. The English phrase "in the driver's seat" can be interpreted literally as "sitting in the driver's seat of a car" or figuratively as "being in control of the situation." The reference translation adopts the figurative meaning.

For **DDCoT-SFT**, the reasoning process follows a well-defined structure. The model first grounds the visual context (a soccer player celebrating on the field), then explicitly checks for ambiguity, and finally applies visual disambiguation to conclude that the figurative interpretation is more appropriate. A localized refinement step further adjusts the expression to match the figurative sense, producing the correct output 他在掌控之中 (He's in control of the situation).

By contrast, **QvQCoT-SFT** generates an unstructured and excessively long reasoning trace. Although it identifies the ambiguity and considers both literal and figurative interpretations, its reasoning is meandering and repetitive, filled with backtracking markers such as "wait" and repeated re-evaluations. Despite acknowledging the figurative possibility, the model ultimately adheres to the literal translation 他坐在驾驶座上 (He's in the driver's seat), which misaligns with the intended meaning.

Summarizing all results, both the quantitative evaluation and the case study highlight the limitations of native **QvQCoT** traces. While they reflect spontaneous reasoning, the traces are relatively unstructured and often excessively long, which obscures the critical steps for translation disambiguation. In contrast, **DDCoT** provides concise and structured reasoning tailored to the disambiguation in MMT task, yielding clearer supervision and superior performance across metrics as well as qualitative analysis.

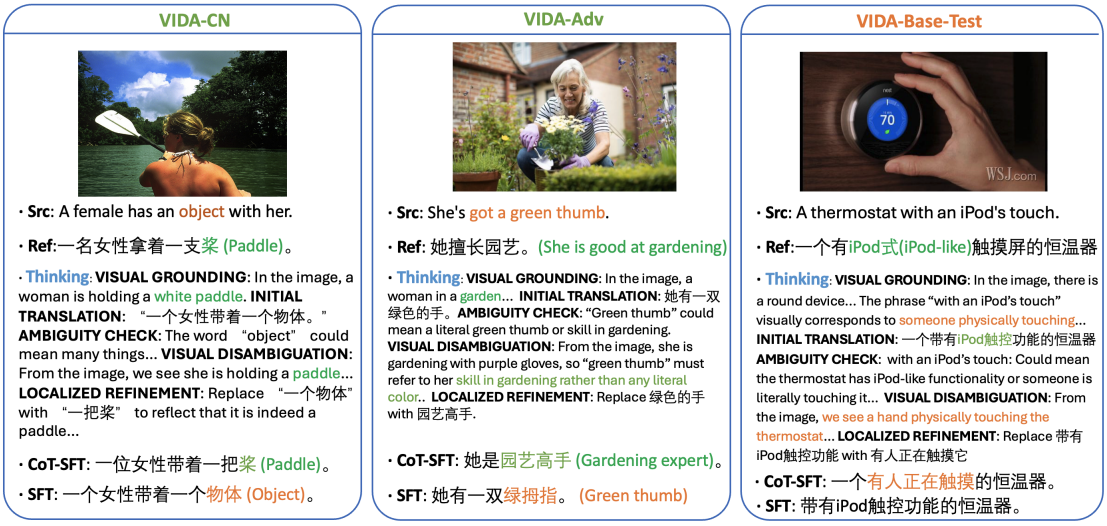

Figure 6.3: Case study of DDCoT-SFT vs. SFT

6.5 Qualitative Analysis

As discussed in section 6.2, **DDCoT-SFT** exhibits a strong ability to enhance disambiguation performance, particularly on challenging OOD subsets (**VIDA-Sent**, **VIDA-CollN**). This raises a key question: *how does explicit reliance on visual information shape the model's reasoning?* Figure 6.3 (left and middle) illustrates two case studies that shed light on this process, showing how **DDCoT-SFT** aligns ambiguous terms with visual evidence in **VIDA-CollN** and **VIDA-Sent**.

The **VIDA-CollN** example (left of Figure 6.3) illustrates the collective noun ambiguity, which 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 **DDCoT-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 6.3) 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 output of the color of thumb in Chinese. In contrast, the **DDCoT-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.

Together, the two cases highlight the role of structured reasoning in bridging linguistic ambiguity with visual grounding, showing that explicit articulation of reasoning steps supports more reliable disambiguation outcomes.

Model	Overthinking Rate	Pearson
Qwen3-Max	0.84	0.72
GPT-5	0.82	

Table 6.4: Overthinking rates identified by Qwen3-Max and GPT-5, along with Pearson correlation indicating inter-evaluator agreement.

6.6 Analysis of Overthinking in DDCoT

While the quantitative and qualitative analyses confirm the effectiveness of DDCoT for MMT disambiguation, its rigid multi-step structure can also introduce unnecessary complexity. In particular, applying elaborate reasoning to straightforward inputs can cause the model to overthink, leading to degraded outputs. As shown in Table 6.2, **DDCoT-SFT** performs slightly worse than SFT on standard translation metrics in **VIDA-Base-Test**, which is consistent with the overthinking issue. This overthinking phenomenon typically emerges after the second step of generating the initial translation: instead of preserving an adequate initial output, the model introduces unnecessary reasoning, such as incorporating irrelevant visual context or misinterpreting idiomatic expressions literally, resulting in flawed revisions and ultimately poorer translations.

To examine whether the performance drop of **DDCoT-SFT** on standard translation metrics is indeed caused by overthinking, I randomly sampled 100 samples from the results of **VIDA-Base-Test** where **DDCoT-SFT** underperformed SFT in both *BLEU* and *COMET*. These samples were then evaluated by two LLMs, Qwen3-Max (Yang et al., 2025) and GPT-5 [2], to verify the presence of overthinking in the reasoning traces. For each case, the evaluator was given the source sentence, the reference translation, and the full reasoning trace. The assessment followed a two-step procedure: (1) examine the initial translation generated in the second step of the reasoning chain and compare it with the final translation; if the final output diverged further from the reference than the initial one, then (2) check whether subsequent reasoning steps contained unnecessary or excessive interpretations that directly led to the degradation. Cases meeting both conditions were labeled as instances of overthinking.

The overthinking analysis is presented in Table 6.4. Both evaluators, Qwen3-Max and GPT-5, identified a high proportion of overthinking cases, with rates of 0.84 and 0.82 respectively, indicating that the majority of instances where **DDCoT-SFT** underperformed SFT can indeed be attributed to overthinking. Moreover, I use Pearson correlation to validate the agreement of overthinking identification between the two evaluators. The resulting correlation of 0.72 suggests a strong level of agreement between the two evaluators, indicating that the identification of overthinking cases is consistent across two evaluators.

To further illustrate how overthinking traces can affect translation quality, I provide an actual case in the right of Figure 6.3. In this example, the phrase "iPod's touch" should be interpreted as "iPod-like touch screen." During the visual grounding stage, the model correctly describes the image, and in the ambiguity check, it identifies the possible meaning of iPod-like functionality. However, in the visual disambiguation stage, the model begins to over-interpret: instead of leveraging the appropriate visual cue, it associates the ambiguous phrase with the mention of "someone physically touching"

2. https://openai.com/en/index/gpt-5-system-card/

from the grounding step. This misalignment leads the model to revise the initial adequate interpretation into an incorrect final translation. This example highlights the risk of overthinking as a key factor behind the observed performance drop, showing that excessive reasoning can overwrite adequate initial translations and thereby reduce overall performance.

7 Conclusion

This work revisited the core question of whether LVLMs truly and effectively leverage visual information in multimodal machine translation (MMT). By approaching the problem through the perspective of ambiguity resolution, three research questions were addressed. First, this work demonstrated that existing datasets fall short of supporting visually dependent disambiguation, leading to the creation of the **VIDA** dataset, which specifically targets word- and sentence-level ambiguities that can only be resolved through visual cues. Second, the investigation revealed that standard translation metrics such as BLEU and COMET are not aligned with the disambiguation task, motivating the introduction of **Disambiguation-Centric Metrics** (Disambi-Term and Disambi-Inst.) that directly measure whether ambiguous spans are correctly resolved. Third, experiments on the **VIDA** dataset demonstrate that while standard metrics showed limited improvements, the proposed **Disambiguation-Centric Metrics** captured substantial and consistent gains, highlighting the necessity of Disambiguation-Centric Metrics for assessing the contribution of visual information to the disambiguation task in MMT and directly reflecting the role of visual input in resolving ambiguity.

Building on the findings that LVLMs can leverage visual information for MMT disambiguation, this work introduced **Disambiguation-Driven Chain-of-Thought Supervised Fine-Tuning (DDCoT-SFT)**, which integrates a structured **Disambiguation-Driven Chain-of-Thought (DDCoT)** template with CoT-SFT training method. Experiments show that **DDCoT-SFT** achieves stronger semantic adequacy and higher disambiguation accuracy than conventional fine-tuning, particularly on out-of-distribution subsets and the aggregated All-Test set, underscoring its superior generalization to diverse ambiguity types beyond the training distribution. Furthermore, the comparison of the impact of synthetic **DDCoT** traces and native, free-form reasoning traces in training for MMT disambiguation reveals that the **DDCoT-SFT** model consistently outperforms native-CoT fine-tuned model, demonstrating that concise, task-structured reasoning supervision provides clearer guidance and more reliable visually grounded disambiguation than unstructured and excessively long native traces.

Beyond quantitative results, qualitative analysis illustrated how **DDCoT-SFT** systematically aligns ambiguous terms with visual evidence through explicit reasoning steps, enabling accurate resolution of both word-level and sentence-level ambiguities.

At the same time, analysis also revealed a limitation: the rigid multi-step structure could lead to overthinking, where excessive reasoning overwrites already adequate translations and degrades performance. These qualitative analyses highlight both the promise and challenges of structured reasoning for MMT disambiguation, suggesting that future work should explore adaptive reasoning strategies that retain the benefits of explicit visual grounding while mitigating overthinking.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision,* 2425–2433. (Cited on page 6).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473,* (cited on page 9).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609,* (cited on page 6).

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. (Cited on page 6).

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,* (cited on pages 1, 4, 8, 21, 30).

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. 2025. A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models. *arXiv preprint arXiv:2505.23945,* (cited on page 25).

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,* edited by Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, 65–72. Ann Arbor, Michigan: Association for Computational Linguistics, June. (Cited on pages 21, 30).

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* edited by Marilyn Walker, Heng Ji, and Amanda Stent, 1304–1313. New Orleans, Louisiana: Association for Computational Linguistics, June. (Cited on page 2).

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2016. Do you see what i mean? visual resolution of linguistic ambiguities. *arXiv preprint arXiv:1603.08079,* (cited on page 10).

Johan Boye and Birger Moell. 2025. Large language models and mathematical reasoning failures. *arXiv preprint arXiv:2502.11574,* (cited on page 10).

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. *arXiv preprint arXiv:1702.01287,* (cited on pages 1, 9).

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15 (3): 1–45. (Cited on page 20).

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461,* (cited on page 25).

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision,* 104–120. Springer. (Cited on page 6).

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,* (cited on page 26).

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),* 4171–4186. (Cited on pages 21, 30).

Desmond Elliott. 2018. Adversarial Evaluation of Multimodal Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* edited by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 2974–2978. Brussels, Belgium: Association for Computational Linguistics, October. (Cited on pages 1, 9).

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation,* edited by Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, 215–233. Copenhagen, Denmark: Association for Computational Linguistics, September. (Cited on page 2).

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459,* (cited on page 9).

Javier Ferrando, Gerard I Gállego, Belen Alastruey, Carlos Escolano, and Marta R Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. *arXiv preprint arXiv:2205.11631,* (cited on page 25).

Clayton Fields and Casey Kennington. 2023. Vision language transformers: A survey. *arXiv preprint arXiv:2307.03254,* (cited on pages 6 sq.).

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),* edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, 5394–5413. Toronto, Canada: Association for Computational Linguistics, July. (Cited on pages 3, 10).

Yue Gao, Jing Zhao, Shiliang Sun, Xiaosong Qiao, Tengfei Song, and Hao Yang. 2025. Multimodal Machine Translation with Text-Image In-depth Questioning. In *Findings of the Association for Computational Linguistics: ACL 2025,* edited by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, 9274–9287. Vienna, Austria: Association for Computational Linguistics, July. (Cited on page 1).

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594,* (cited on pages 3, 20).

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948,* (cited on page 11).

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301,* (cited on pages 4, 11 sq., 24).

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers,* 639–645. (Cited on pages 1, 9).

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,* (cited on page 14).

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720,* (cited on pages 11, 27).

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning,* 4904–4916. PMLR. (Cited on page 6).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35:22199–22213. (Cited on page 10).

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* (Cited on pages 1, 10).

Beomseok Lee, Hyunwoo Kim, Keon Kim, and Yong Suk Choi. 2024. XDetox: Text Detoxification with Token-Level Toxicity Explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 15215–15226. Miami, Florida, USA: Association for Computational Linguistics, November. (Cited on page 30).

Jaechan Lee, Alisa Liu, Orevaoghene Ahia, Hila Gonen, and Noah A Smith. 2023. That was the last straw, we need more: Are Translation Systems Sensitive to Disambiguating Context? *arXiv preprint arXiv:2310.14610,* (cited on page 2).

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 13872–13882. (Cited on page 8).

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326,* (cited on pages 4, 8, 21).

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,* edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 8556–8562. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, November. (Cited on pages 6, 10).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning,* 19730–19742. PMLR. (Cited on pages 6 sq.).

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557,* (cited on page 6).

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. MSCTD: A multimodal sentiment chat translation dataset. *arXiv preprint arXiv:2202.13645,* (cited on page 1).

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,* (cited on page 15).

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),* edited by Luis Chiruzzo, Alan Ritter, and Lu Wang, 1559–1570. Albuquerque, New Mexico: Association for Computational Linguistics, April. (Cited on page 12).

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* 26296–26306. (Cited on page 8).

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36:34892–34916. (Cited on page 6).

Yichen Lu, Wei Dai, Jiaen Liu, Ching Wing Kwok, Zongheng Wu, Xudong Xiao, Ao Sun, Sheng Fu, Jianyuan Zhan, Yian Wang, et al. 2025. ViDove: A Translation Agent System with Multimodal Context and Memory-Augmented Reasoning. *arXiv preprint arXiv:2507.07306,* (cited on page 1).

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 Multi-Modal Machine Translation Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 1–13. Torino, Italia: ELRA / ICCL, May. (Cited on pages 2 sq., 10, 13).

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching Small Language Models to Reason. *arXiv preprint arXiv:2212.08410,* (cited on pages 4, 11 sq., 24).

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,* (cited on pages 4 sq., 11, 24, 36).

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,* edited by Pierre Isabelle, Eugene Charniak, and Dekang Lin, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July. (Cited on pages 3, 10, 19, 21, 30).

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation,* edited by Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, 392–395. Lisbon, Portugal: Association for Computational Linguistics, September. (Cited on pages 21, 30).

———. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation,* edited by Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, 612–618. Copenhagen, Denmark: Association for Computational Linguistics, September. (Cited on pages 21, 30).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning,* 8748–8763. PmLR. (Cited on page 6).

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),* edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 2685–2702. Online: Association for Computational Linguistics, November. (Cited on pages 3, 10, 19, 21, 30).

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300,* (cited on page 11).

Jianshu She, Zhuohao Li, Zhemin Huang, Qi Li, Peiran Xu, Haonan Li, and Qirong Ho. 2025. Hawkeye: Efficient reasoning with model collaboration. *arXiv preprint arXiv:2504.00424,* (cited on page 11).

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges. *arXiv preprint arXiv:2405.12669,* (cited on pages 1, 10, 25).

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,* 223–231. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, August. (Cited on pages 21, 30).

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. *arXiv preprint arXiv:1710.07177,* (cited on page 1).

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, et al. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419,* (cited on page 11).

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115,* (cited on pages 15, 21).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971,* (cited on page 6).

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025. DRT: Deep Reasoning Translation via Long Chain-of-Thought. In *Findings of the Association for Computational Linguistics: ACL 2025,* edited by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, 6770–6782. Vienna, Austria: Association for Computational Linguistics, July. (Cited on pages 12, 26).

Ru Wang, Selena Song, Liang Ding, Shixiang Shane Gu, Mingming Gong, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. 2024. MMA: Benchmarking Multi-Modal Large Language Model in Ambiguity Contexts. (Cited on pages 2, 13).

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. 2024. Symbolic working memory enhances language models for complex rule application. *arXiv preprint arXiv:2408.13654,* (cited on page 10).

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. In *Findings of the Association for Computational Linguistics ACL 2024,* 15840–15853. (Cited on page 8).

Xintong Wang, Jingheng Pan, Yixiao Liu, Xiaohu Zhao, Chenyang Lyu, Minghao Wu, Chris Biemann, Longyue Wang, Linlong Xu, Weihua Luo, et al. 2025. Rethinking Multilingual Vision-Language Translation: Dataset, Evaluation, and Adaptation. *arXiv preprint arXiv:2506.11820,* (cited on pages 1, 8).

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171,* (cited on page 10).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35:24824–24837. (Cited on pages 10, 25).

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),* edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 6153–6166. Online: Association for Computational Linguistics, August. (Cited on pages 1, 9).

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. (Cited on page 26).

Rongwu Xu, Zian Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, and Han Qiu. 2024. Walking in Others' Shoes: How Perspective-Taking Guides Large Language Models in Reducing Toxicity and Bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 8341–8368. Miami, Florida, USA: Association for Computational Linguistics, November. (Cited on page 30).

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech. In *Findings of the Association for Computational Linguistics: ACL 2024,* edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 13967–13983. Bangkok, Thailand: Association for Computational Linguistics, August. (Cited on page 30).

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,* (cited on pages 20, 40).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. 2024. Qwen2 Technical Report. *ArXiv* abs/2407.10671. (Cited on page 21).

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,* edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 4346–4350. Online: Association for Computational Linguistics, July. (Cited on page 1).

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36:11809–11822. (Cited on page 10).

Kayo Yin, Yu Zhang, and Graham Neubig. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. *arXiv preprint arXiv:2007.08742,* (cited on page 1).

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198,* (cited on page 25).

Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 2025. 1.4 million open-source distilled reasoning dataset to empower large language model training. *arXiv preprint arXiv:2503.19633,* (cited on page 11).

Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Xuyao Zhang, and Dahan Wang. 2025. MCA-LLaVA: Manhattan Causal Attention for Reducing Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2507.09184,* (cited on page 25).

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592,* (cited on page 7).

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,* (cited on pages 1, 4, 8, 21, 30).