# DART: Disambiguation-Aware Reasoning for Video-guided Machine Translation

**Boyu Guan**[1,2], **Chuang Han**[1,2], **Yang Zhao**[1,2]*, **Chengqing Zong**[1,2]*
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{guanboyu2022, hanchuang2025}@ia.ac.cn, {yang.zhao, cqzong}@nlpr.ia.ac.cn

## Abstract

Video-guided Machine Translation (VMT) seeks to enhance translation quality by incorporating contextual information derived from paired short video clips. However, many VMT samples are text-sufficient; even when visual information is needed, only minimal cues are required. Aiming to tackle these issues, we propose a novel framework **DART** (**D**isambiguation-**A**ware **R**easoning for Video-guided Machine **T**ranslation). Reinforcement learning is used to incorporate multimodal large language models' multimodal reasoning into VMT. The model dynamically switches between text-only processing and multimodal integration, contingent on the necessity of visual disambiguation. Furthermore, we present **TVRF** (**T**ranslation-oriented **V**ideo **R**elevance **F**iltering), a systematic pipeline for constructing training data based on multimodal relevance to translation. This pipeline filters samples where video information is translation-relevant, mitigating training collapse caused by video-irrelevant data in conventional VMT. Experimental results show that our approach improves multimodal information utilization in VMT, yielding gains in both translation quality and computational efficiency.

## 1 Introduction

Video-guided Machine Translation (VMT) is a rapidly emerging task at the intersection of multimodality and machine translation (Shen et al., 2024; Feng et al., 2025b). The VMT task takes as input an 8-10 second video clip paired with a text segment, typically from subtitles or video descriptions. Its goal is to leverage video context to improve translation quality, especially for ambiguous source text (Wang et al., 2019).

Multimodal Large Language Models (MLLMs) have rapidly advanced in cross-modal understanding (OpenAI, 2024a). Building on this, test-time scaling enables Large Multimodal Reasoning Models (LMRMs) to leverage increased inference computation for improved downstream performance (OpenAI, 2024b; Li et al., 2025). Human translation in VMT is inherently a reasoning process. Translators first determine whether the source sentence contains ambiguities that require video context. If not, text-only translation suffices; otherwise, multimodal video information is used to disambiguate and produce the final translation.


Figure 1: Comparison of existing LMRMs (top) and DART (bottom) for VMT. LMRMs apply verbose and inefficient reasoning to all samples, whereas DART adapts to input ambiguity, translating directly when unambiguous and selectively using **multimodal cues** for **disambiguation**. **Green** and **red** indicate correct and incorrect translations, respectively.

However, as illustrated in Figure 1, directly applying LMRMs to the VMT task leads to three critical issues. (i) **Inefficient and counterproductive reasoning traces**. LMRMs adopt a uniform step-by-step reasoning strategy across VMT inputs of varying ambiguity, from text-clear utterances (e.g., "oHHHHH!") to video-dependent cases (e.g., "It's hitting his yacht"). This mechanical reasoning paradigm has been shown to incur 40× inference overhead and to harm translation quality (Wu et al., 2025b; Zebaze et al., 2025). (ii) **Neglect of visual context in disambiguation**. Experiments in Section 5 and Appendix E demonstrate that the lengthy

*Equal corresponding authors.



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7772

reasoning trajectories of these LMRMs are heavily biased toward text-only decomposition, rarely prompting the model to reason about or attend to the video information when it is actually required. (iii) **Dataset bias induced visual underutilization**. Moreover, our empirical analysis in Section 5.3 shows that approximately 69% of samples in existing VMT datasets are text-sufficient. This dataset bias encourages models to underutilize video information, which in turn degrades performance on cases where visual context is essential (Yang et al., 2022; Kang et al., 2023).

Inspired by this observation, we propose **DART** (**D**isambiguation-**A**ware **R**easoning for Video-guided Machine **T**ranslation) to mitigate overly long reasoning trajectories and the neglect of disambiguating visual cues. After training with Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL), DART adapts its translation strategy to input semantic ambiguity. For unambiguous cases, it explicitly determines that no video information is required and translates directly from the source sentence. For ambiguous or challenging cases, DART first identifies the most relevant multimodal cue from *people*, *objects*, *actions*, *OCR*, *spatial relations*, and *pointing gaze*. It then explicitly verbalizes this cue to ground the model's attention in the video before performing the translation. To address visual underutilization caused by dataset bias, we propose TVRF (**T**ranslation-oriented **V**ideo **R**elevance **F**iltering), a robust pipeline for training data construction. TVRF serves as an efficient filter that allows DART to learn to distinguish between different sample types during training, thereby optimizing its translation strategy based on the necessity of visual context.

We conduct experimental evaluations on the video-caption VMT dataset VATEX (Wang et al., 2019) and the video-subtitle VMT dataset TriFine (Guan et al., 2025b). The results demonstrate that DART consistently outperforms existing VMT approaches across multiple translation evaluation metrics, while simultaneously achieving a substantial improvement in inference efficiency.

Our main contributions are summarized as follows:

- We propose DART, the first approach to apply reasoning-aware MLLMs to VMT, which adaptively leverages video information based on sample ambiguity to jointly enhance translation quality and efficiency.
- We introduce TVRF, a video-helpfulness-aware dataset construction pipeline that selectively identifies VMT samples where visual information genuinely contributes to translation, thereby mitigating long-standing data bias in VMT.
- We evaluate DART on different benchmarks, where the results consistently demonstrate its strong efficiency and effectiveness in improving translation quality.
- All code for SHIFT has been publicly released at `https://github.com/BoyuGuan/DART`.

## 2 Related Work

**Video-guided Machine Translation.** Incorporating visual context into machine translation has proven effective for enhancing translation performance (Wang and Xiong, 2021; Futeral et al., 2023; Feng et al., 2025b; Liang et al., 2025c). Early studies were largely motivated by benchmarks such as Multi30K (Elliott et al., 2016) and primarily concentrated on image-guided machine translation (Lin et al., 2020; Wu et al., 2021; Fang and Feng, 2022; Fei et al., 2023; **?**; Zhang et al., 2025c; Yu et al., 2025b; Zhang et al., 2025e; Xiong and Zhao, 2025). These methods leverage visual images to resolve linguistic ambiguities in the source text, thereby improving translation quality (Cheng et al., 2024; Wang et al., 2024; Yang et al., 2024; Khan et al., 2024; Liang et al., 2025b; Futeral et al., 2025; Zhang et al., 2025d; Gao et al., 2025). Recent research has increasingly shifted its attention toward video-guided machine translation, as it allows models to leverage the richer temporal and dynamic multimodal information available in video clips (Gu et al., 2021; Li et al., 2023; Kang et al., 2023; Shurtz et al., 2024; Lv et al., 2025; Zheng et al., 2025). However, rich visual information is a double-edged sword: excessive redundancy can introduce visual noise and substantially increase computational cost (Yang et al., 2022). Consequently, developing efficient strategies for leveraging video content has become a central research focus in VMT task (Guan et al., 2025a).

**Adaptive Reasoning.** With the introduction of the test-time scaling law paradigm (OpenAI, 2024b), large reasoning models have demonstrated strong performance across a wide range of downstream tasks (DeepSeek-AI et al., 2025); however, they also exhibit a tendency to overthink simple problems (Zhang et al., 2025b; An et al., 2025; Yi et al., 2025). Consequently, recent work focuses on

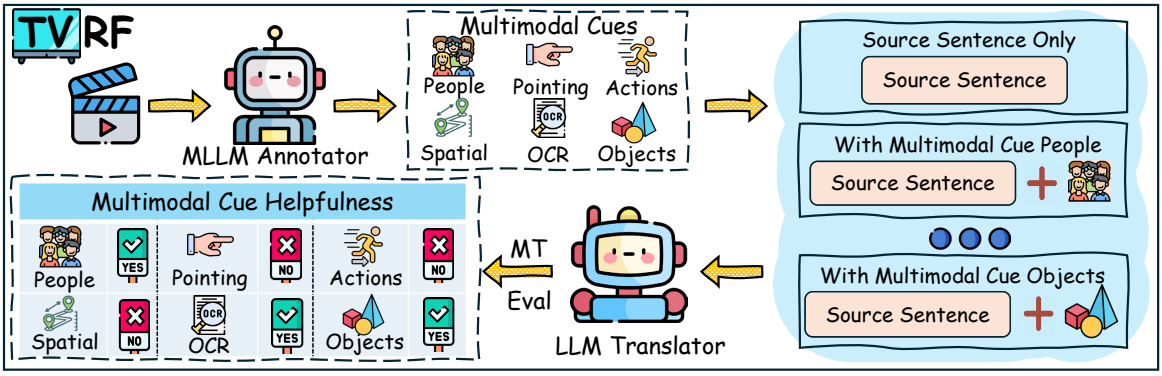


Figure 2: Overview of the TVRF framework. For each VMT data instance, TVRF determines whether video-based multimodal cues aid translation. An MLLM extracts and verbalizes multimodal cues, which are supplied to an LLM alongside the source sentence for translation, while the baseline uses the source sentence alone. Quantitative comparisons between these settings evaluate the impact of multimodal cues on translation quality.

optimizing the trade-off between reasoning performance and efficiency (Jiang et al., 2025; Fang et al., 2025; Liang et al., 2025a; Lu et al., 2025; Zhang et al., 2025a). This line of research has advanced from shortening reasoning trajectories to adaptive thinking, enabling models to dynamically decide whether and how deeply to reason (Ma et al., 2025; Yi et al., 2025; An et al., 2025; Lin et al., 2025b; Chen et al., 2025b; Jian et al., 2025; Guo et al., 2025; Wu et al., 2025a). The paradigm has also been extended to multimodal settings, offering a critical solution to the inherently high computational overhead (Lin et al., 2025a; Sun et al., 2024; Xiao and Gan, 2025; Yang et al., 2025b).

However, these approaches are not directly applicable to VMT. LMRMs utilize decomposition-based reasoning suited for logic-heavy tasks; however, research suggests this paradigm is counterproductive for translation (Wu et al., 2025b; Zebaze et al., 2025). Furthermore, current LMRMs fail to adaptively exploit fine-grained multimodal signals or recognize when textual context alone suffices for VMT (Guan et al., 2025a).

## 3 Method

We propose the TVRF pipeline to alleviate dataset bias and the DART framework to rectify inefficient reasoning and overlooked visual context in disambiguation. We describe these methodologies in Sections 3.1 and 3.2.

### 3.1 TVRF

A large fraction of VMT data is text-sufficient, not requiring video input. This data imbalance biases MLLMs toward consistently ignoring video information when processing VMT samples. To address this issue, we propose TVRF, which retains only samples where multimodal information is beneficial, enabling stable and effective training. The overall framework is illustrated in Figure 2.

Directly asking an MLLM to determine whether the video is helpful for translation yields poor and unstable performance, as this capability is not explicitly learned during either pre-training or post-training. TVRF adopts a fundamentally different annotation and filtering strategy. VMT videos often include rich multimodal signals, the majority of which are unrelated to translation. Based on statistical analyses from preliminary experiments and existing works, we summarize the multimodal cues in video that are potentially beneficial for translation into six categories: *people*, *objects*, *actions*, *OCR*, *spatial relations*, and *pointing gaze*.

$$\mathcal{C} = \{c^{people}, c^{object}, c^{action}, c^{ocr}, c^{spatial}, c^{pointing}\} \quad (1)$$

Let the dataset be $\mathcal{D} = (v_i, x_i, y_i)_{i=1}^{N}$, where $v_i$ denotes the video clip associated with the $i$-th instance, $x_i$ is the source sentence, and $y_i$ is the reference translation. Because these six multimodal cues frequently appear in large-scale training tasks for MLLMs, such models are able to annotate them with high accuracy. For each instance $v_i$ and each cue category $c^k \in \mathcal{C}$, an MLLM annotator $\mathcal{A}$ assigns a natural-language label to the corresponding cue.

$$c_i^k = \mathcal{A}(v_i) \quad (2)$$

Where $c_i^k$ may be an empty string when the corresponding cue is absent in video.

A strong text-only LLM $\mathcal{T}$ is employed as the translation backbone to evaluate the impact of incorporating different multimodal cues on translation performance. A source sentence only baseline

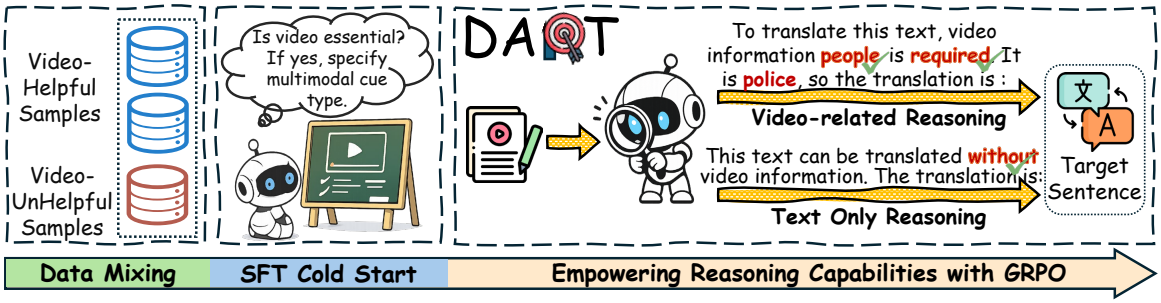


Figure 3: Schematic of the DART training workflow. The pipeline begins with SFT cold-start using TVRF-curated data to initialize the VMT reasoning format in the MLLM. This is followed by GRPO-based optimization to enhance reasoning depth. Crucially, we implement dual-path reward functions tailored to the presence or absence of multimodal cues, allowing the model to adaptively refine its reasoning logic.

translation $\hat{y}_i^0$ is first produced; each multimodal cue $c^k \in \mathcal{C}$ is then integrated with the source sentence and its description through a fixed template $\oplus$ to generate a cue-conditioned translation $\hat{y}_i^k$.

$$\hat{y}_i^0 = \mathcal{T}(x_i), \qquad \hat{y}_i^k = \mathcal{T}(x_i \oplus c_i^k) \tag{3}$$

Translation quality is evaluated using a metric function $m(x, \hat{y}, y)$, where higher values indicate better translation quality. For each instance $i$ and cue type $k$, the cue benefit label is defined by thresholding the score gain over the text-only baseline:

$$b_i^k = \mathbb{I}\left[m\left(x_i, \hat{y}_i^k, y_i\right) - m\left(x_i, \hat{y}_i^0, y_i\right) > \delta\right] \tag{4}$$

$\mathbb{I}[\cdot]$ equals 1 if the condition holds, and 0 otherwise. The binary cue-activation vector $\mathbf{b}_i = (b_i^1, \ldots, b_i^k)^\top \in \{0,1\}^K$ summarizes which cue types yield measurable improvements under the chosen metric and threshold $\delta$. The necessity of auxiliary visual information for a sample is determined by the sparsity of its cue-activation vector $\mathbf{b}_i$. Samples are categorized into two distinct sets based on the $L_0$-norm:

$$S_i = \begin{cases} \mathcal{D}_{\text{multimodal}}, & \text{if } \|\mathbf{b}_i\|_0 > 0 \\ \mathcal{D}_{\text{text-only}}, & \text{if } \|\mathbf{b}_i\|_0 = 0 \end{cases} \tag{5}$$

Subsequently, a new reference $z_i$ is constructed by fusing the optimal reasoning path with the translation content. For text-only instances $i \in \mathcal{D}_{\text{text-only}}$, $z_i$ concatenates an explicit "no video required" indicator with the ground-truth translation $y_i$. In contrast, for multimodal instances ($i \in \mathcal{D}_{\text{multimodal}}$), $z_i$ combines the identity and natural-language description of the most effective cue $c_i^*$ with $y_i$. Here, $c_i^*$ is determined by selecting the cue corresponding to the maximum value of $m(x_i, \hat{y}_i^k, y_i)$.

$$z_i = \begin{cases} \text{Expl}_{\text{text}} \oplus y_i, & i \in \mathcal{D}_{\text{text-only}} \\ \text{Expl}_{\text{cue}-c^*} \oplus \text{Desc}_{\text{cue}-c^*} \oplus y_i, & i \in \mathcal{D}_{\text{multimodal}} \end{cases} \tag{6}$$

Details of TVRF's extraction and generation pipeline are in Appendix A.

### 3.2 DART

Figure 3 depicts the training workflow for DART. A large portion of VMT samples fall into $\mathcal{D}_{\text{text-only}}$, accounting for approximately 69% of the data. Directly training on this skewed distribution may bias the model toward a degenerate text-only policy that systematically ignores visual information. To alleviate this issue, DART introduces an imbalance-aware sampling strategy. Specifically, training batches are constructed using a predefined mixing coefficient $\alpha \in (0, 1)$.

$$\alpha \cdot \mathcal{D}_{\text{multimodal}} + (1 - \alpha) \cdot \mathcal{D}_{\text{text-only}} \tag{7}$$

Explicit supervision templates are designed for both subsets to manifest the model's latent decision on visual necessity. For text-only samples ($|\mathbf{b}_i|_0 = 0$), the target response declares the absence of visual requirements before providing the translation. Conversely, for multimodal samples ($|\mathbf{b}_i|_0 > 0$), the response confirms visual necessity and conditions the translation on the salient cue set $\mathcal{K}_i$, paired with corresponding textual descriptions. This mechanism forces explicit visual-dependency reasoning and grounds the translation in the most influential visual evidence.

DART employs a two-stage training strategy. For initial alignment with VMT-oriented reasoning patterns, a Supervised Fine-Tuning (SFT) stage is performed as a cold start on $\mathcal{D}_{\text{SFT}} = \{(x_i, v_i, z_i)\}_{i=1}^N$. To further strengthen multimodal reasoning through task-specific rewards, the model is refined via Group Relative Policy Optimization (GRPO) on $\mathcal{D}_{\text{RL}} = \{(x_i, v_i, z_i)\}_{i=1}^M$ (Shao et al., 2024). The objective maximizes the advantage $A(\hat{z})$ while maintaining stability via a reference

policy $\pi_{\theta_{\text{old}}}$. The advantage $A(\hat{z}_i)$ is obtained by standardizing rewards across the $G$ sampled outputs for each input, capturing a candidate translation's quality relative to the group mean.

$$A(\hat{z}_i) = \frac{r(\hat{z}_i, z_i) - \frac{1}{G}\sum_{j=1}^{G} r(\hat{z}_j, z_i)}{\text{std}(\{r(\hat{z}_1, z_i), \ldots, r(\hat{z}_G, z_i)\})} \tag{8}$$

Where $r(\cdot)$ denotes the DART reward function evaluating the sampled translation against the reference.

$$\begin{aligned} r(\hat{z}_i, z_i) = & \, w_{\text{qual}} r_{\text{qual}}(\hat{z}_i, z_i) \\ & + w_{\text{logic}} r_{\text{logic}}(\hat{z}_i, z_i) \\ & - r_{\text{len}}(\hat{z}_i) \end{aligned} \tag{9}$$

Where $w_{\text{qual}}$ and $w_{\text{logic}}$ balance translation fidelity, logical consistency, and output length. $r_{\text{quality}}$ is defined based on the COMET (Rei et al., 2022) score to incentivize high-fidelity translations in the translation component of $\hat{z}_i$. $r_{\text{logic}}$ is used to incentivize VMT-oriented reasoning behavior in the model. For sample $i$ in $\mathcal{D}_{\text{text-only}}$, the reward $r_{\text{logic}}$ is a binary indicator.

$$r_{\text{logic}} = \mathbb{I}[\text{text-only} \in \hat{z}_i] \tag{10}$$

For each sample $i$ in $\mathcal{D}_{\text{multimodal}}$, the reward $r_{\text{logic}}$ consists of two components: the accuracy of the model in determining whether video information is required, and the similarity between the multimodal cue types used in the reasoning process and the most effective multimodal cue $c_i^*$.

$$S = \lambda + (1 - \lambda)\,\text{sim}(\hat{c}_i, c_i^*) \tag{11}$$

$$r_{\text{logic}} = \mathbb{I}[\text{multimodal} \in \hat{z}_i] \times S \tag{12}$$

Here, $\lambda$ is a weighting coefficient, and $\hat{c}_i$ denotes the predicted multimodal cues derived from $\hat{z}_i$. $sim()$ is used to compute textual similarity. This structure ensures the model is penalized for both hallucinating visual needs and failing to utilize relevant visual cues accurately. To avoid degradation in translation quality and inference efficiency, $r_{\text{len}}$ penalizes excessively long outputs.

$$r_{\text{len}} = \begin{cases} 0 & \text{if } |\hat{z}_i| \leq L_{limit} \\ e^{\alpha(|\hat{z}_i| - L_{limit})} - 1 & \text{if } |\hat{z}_i| > L_{limit} \end{cases} \tag{13}$$

Where $L_{\text{limit}}$ denotes the output length limit.

# 4 Experiments

## 4.1 Implementation Details

Balancing performance and computational cost, we employ the MLLM Qwen3-VL-32B-Instruct (Bai et al., 2025) to pre-extract multimodal cues and the LLM Qwen3-30B-A3B-Instruct (Yang et al., 2025a) to generate final translations under different input configurations. The implementation details of TVRF can be found in Appendix A. Considering the trade-off between computational cost and performance, we adopt Qwen3-VL-4B-Instruct as the base model for DART. The $\alpha$ in Equation 7 is set to 0.6, indicating that 60% of the training samples require video information to assist translation. RL training is conducted using the verl framework (Sheng et al., 2025). The full training details are given in Appendix B.

## 4.2 Data

We employ TVRF to construct 36K training instances from the training split of the TriFine dataset (Guan et al., 2025b) for SFT, and 10K instances for RL. We conduct our evaluation on the test sets of the following two datasets. (i) **TriFine Dataset**: A large-scale VMT dataset featuring approximately 2.4M sentence pairs aligned with 10-second video clips. While the general test sets contain 7,000 samples per direction, the ambiguity test set consists of 1,001 cases specifically designed to require video context for accurate translation. (ii) **VATEX Dataset** (Wang et al., 2019): This video description dataset comprises 25,991 training videos. As the official test set of VATEX is unavailable, we adhere to the common setup (Kang et al., 2023) of bisecting the 3,000-video validation set into distinct validation and testing subsets.

## 4.3 Evaluation Metrics

To ensure consistency with prior work and enable fair comparison, we adopt standard machine translation evaluation metrics: BLEU[1] (Papineni et al., 2002; Post, 2018), COMET[2] (Rei et al., 2022) and BLEURT[3] (Sellam et al., 2020). In addition, we further report end-to-end samples per second (SPS) to assess processing speed on the VMT task.

## 4.4 Baselines

The baselines are grouped into three categories. Detailed descriptions and deployment details of the baselines are provided in Appendix G.

(i) **Traditional VMT methods.** Transformer (Vaswani et al., 2017), TVE, CVE (Shurtz et al., 2024), and FIAT (Guan et al., 2025b) are selected

[1]https://github.com/mjpost/sacrebleu
[2]https://huggingface.co/Unbabel/wmt22-comet-da
[3]https://github.com/lucadiliello/bleurt-pytorch

| # | Method | TriFine | | | VATEX | Speed |
|---|---|---|---|---|---|---|
| | | General (zh→en) | General (en→zh) | Ambiguity (en→zh) | Test (en→zh) | |
| | | BLEU ↑ / COMET ↑ / BLEURT ↑ | | | | SPS ↑ |
| | ***Traditional VMT Methods*** | | | | | |
| 1 | Transformer | 23.47/71.89/56.42 | 36.19/75.11/54.42 | 29.74/74.32/52.98 | 29.61/ 73.07/53.82 | **75.38** |
| 2 | TVE* | 23.85/72.58/57.20 | 36.55/75.64/54.98 | 30.37/74.45/55.55 | 30.30 / 73.37 / — | 1.30 |
| 3 | CVE* | 23.97/72.60/57.19 | 36.43/75.58/55.29 | 30.28/74.39/55.55 | 29.40 / 73.44 / — | 1.28 |
| 4 | FIAT* | **25.51**/73.59/57.89 | **38.06**/76.48/56.15 | 31.24/75.93/56.32 | 30.75/73.92/55.43 | 0.71 |
| | ***Open-source LLMs and LRMs (Text-only)*** | | | | | |
| 5 | Llama-3.1-8B* | 16.68/72.54/55.78 | 25.11/77.66/57.39 | 24.95/77.14/58.91 | 27.81/78.15/57.95 | 9.21 |
| 6 | Qwen3-4B-Instruct | 17.41/74.56/58.79 | 30.42/78.39/59.31 | 30.05/80.15/61.32 | 29.49/77.87/57.56 | 19.82 |
| 7 | DeepSeek-R1-Distill-Llama-8B | 13.38/71.34/54.52 | 25.20/74.44/55.53 | 23.53/76.08/57.13 | 24.48/74.81/54.12 | 0.78 |
| 8 | DeepSeek-R1-Distill-Qwen-7B | 11.26/70.66/54.24 | 22.07/73.60/53.25 | 20.54/73.06/53.12 | 22.82/72.64/51.20 | 0.67 |
| 9 | Qwen3-4B-Thinking | 17.37/74.54/58.86 | 30.63/77.93/58.93 | 29.98/79.97/61.07 | 29.97/77.94/57.42 | 0.31 |
| | ***Open-source MLLMs and LMRMs ( Video-Text )*** | | | | | |
| 10 | LLaVA-Next-Video-8B* | 12.38/68.65/55.18 | 23.63/73.63/57.26 | 23.66/76.35/58.22 | 25.62/75.45/55.10 | 0.65 |
| 11 | InternVideo2.5-8B* | 19.60/75.55/60.18 | 30.28/77.59/57.85 | 31.49/80.25/61.41 | 30.09/78.25/58.04 | 0.72 |
| 12 | MiniCPM-V-4.5-8B | 22.01/75.98/60.59 | 27.61/78.38/59.07 | 31.92/80.71/61.98 | 28.73/78.44/57.52 | 0.51 |
| 13 | Video-R1-7B | 19.59/75.36/60.08 | 32.02/79.15/60.02 | 32.89/81.40/62.97 | **31.46**/78.82/58.83 | 0.07 |
| 14 | Qwen3-VL-4B-Thinking | 18.01/74.83/59.17 | 29.50/77.90/58.72 | 31.26/81.18/62.43 | 27.97/78.57/58.14 | 0.06 |
| 15 | + Self-reasoning | 17.72/74.92/59.52 | 28.54/78.48/59.34 | 29.94/81.51/62.50 | 25.50/78.38/57.41 | 0.06 |
| 16 | Qwen3-VL-4B-Instruct | 19.86/75.38/60.12 | 31.82/78.99/60.22 | 32.54/81.52/62.92 | 29.76/78.77/58.63 | 3.19 |
| 17 | + Self-reasoning | 18.11/75.15/59.80 | 31.53/78.85/60.19 | 33.13/81.56/63.18 | 29.45/79.04/58.96 | 2.96 |
| 18 | + SHIFT | 20.96/75.81/60.90 | 33.25/79.31/60.62 | 33.67/81.79/63.05 | 30.32/79.05/59.03 | 1.25 |
| 19 | + **DART (Ours)** | 25.21/**77.04/61.51** | 34.22/**80.21/61.73** | **34.60/82.91/64.27** | 31.21/**79.91/59.65** | 1.93 |

Table 1: Main results on TriFine (Chinese–English general and ambiguity subsets) and VATEX. All results are averaged over three random seeds; statistical significance ($p < 0.01$) verifies robustness. SPS (Samples Per Second) denotes the average end-to-end inference speed. For each dataset and metric, the best score is highlighted in **bold**. Rows marked with * are reported from Guan et al. (2025a).

as representative traditional VMT methods. They all adopt Transformer-based architectures.

(ii) **Open-source LLMs and LRMs.** We selected the widely used Llama-3.1-8B (Grattafiori et al., 2024) and with Qwen3-4B-Instruct (Yang et al., 2025a) as baseline LLMs. DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025), and Qwen3-4B-Thinking are selected as baseline LRMs.

(iii) **Open-source MLLMs and LMRMs.** We evaluate representative video-capable MLLMs including Qwen-3-VL-4B-Instruct (Bai et al., 2025), LLaVA-Next-Video (Zhang et al., 2024), InternVideo-2.5-Chat-8B (Wang et al., 2025), and MiniCPM-V 4.5 (Yu et al., 2025a), as well as strong LMRMs such as Qwen-3-VL-4B-Thinking, R1-OneVision-7B (Feng et al., 2025a), and Video-R1-7B (Feng et al., 2025a). We also evaluate SHIFT (Guan et al., 2025a), an effective MLLM-based VMT framework.

## 5 Results and Analysis

### 5.1 Main Results

We report the main experimental results of our method and competing approaches in Table 1 on the video subtitle benchmark TriFine (including the en-zh general test sets and the ambiguity test set) and the video description benchmark VATEX.

As shown in the table, DART achieves the best COMET and BLEURT scores across all four test sets, outperforming all baselines. It also attains top-tier BLEU performance, matching FIAT and surpassing other LLM-based methods. These results demonstrate the effectiveness of DART, which, unlike LMRM-based approaches, delivers consistent gains without incurring significant computational overhead.

Compared to the strongest MLLM-based baseline SHIFT (row 18), DART (row 19) consistently outperforms it across all four test sets, yielding average gains of 1.76 BLEU, 1.03 COMET, and 0.89

| Method | General (zh→en) | General (en→zh) | Ambiguity | VATEX |
|---|---|---|---|---|
| | BLEU ↑ / COMET ↑ / BLEURT ↑ | | | |
| Qwen3-VL-4B-Instruct | 19.86/75.38/60.12 | 31.82/78.99/60.22 | 32.54/81.52/62.92 | 29.76/78.77/58.63 |
| + SFT | 25.04/75.53/58.94 | 34.25/78.67/58.81 | 31.83/79.39/60.48 | 30.49/77.15/56.92 |
| + DART SFT | **25.33**/75.99/60.20 | **34.48**/79.31/60.42 | 32.26/81.19/62.03 | 31.17/78.28/57.49 |
| **+ DART** | 25.21/**77.04**/**61.51** | 34.22/**80.21**/**61.73** | **34.60**/**82.91**/**64.27** | **31.21**/**79.91**/**59.65** |

Table 2: Ablation study on the impact of different stages of DART on the final performance.

| Method | General (zh→en) | General (en→zh) | Ambiguity | VATEX |
|---|---|---|---|---|
| | BLEU ↑ / COMET ↑ / BLEURT ↑ | | | |
| **DART** | **25.21**/77.04/**61.51** | **34.22**/80.21/**61.73** | **34.60**/82.91/**64.27** | **31.21**/79.91/**59.65** |
| w/o $r_{qual}$ | 21.32/73.25/54.69 | 30.01/74.97/57.06 | 31.15/79.93/59.31 | 25.12/76.10/58.10 |
| w/o $r_{logic}$ | 24.79/**77.6**2/60.04 | 32.70/**80.44**/60.03 | 31.95/**83.08**/63.21 | 30.45/**80.16**/56.85 |
| w/o $r_{len}$* | N/A | N/A | N/A | N/A |

Table 3: Ablation results on the impact of different reward designs in the RL stage of DART on final performance. $*$: The variant w/o $r_{len}$ failed to converge due to training instability.

BLEURT. In addition, DART improves processing speed by 54.4%, achieving superior translation quality and inference efficiency.

Compared to FIAT (row 4), the strongest conventional VMT baseline, DART yields average gains of +5.04 COMET and +5.34 BLEURT. Although BLEU remains largely comparable, LLM-based methods generally score lower than Transformer-based VMT models, a trend attributed to increased lexical and syntactic flexibility rather than degraded translation quality (He et al., 2024; Chen et al., 2025a).

A comparison between instruct models and their reasoning variants shows that, for both text-only (rows 5 vs. 7) and multimodal (rows 14 vs. 16) settings, existing reasoning mechanisms are poorly suited to VMT. Extended reasoning sharply increases inference cost while yielding no translation gains and, in some cases, degrading performance.

In addition, a self-reasoning procedure aligned with the proposed multimodal cues was introduced (see Appendix C). However, it yields no significant improvement over the direct video–text baseline (rows 14 vs. 15; rows 16 vs. 17), indicating that prompt engineering alone is insufficient for effective multimodal cue reasoning in VMT and underscoring the necessity of the proposed DART training pipeline.

Additional experimental results and analyses are reported in Appendix E.

### 5.2 Ablation Study

An ablation study was conducted to evaluate each stage of DART training, as shown in Table 2. Incorporating standard SFT markedly improves BLEU on general test sets (e.g., zh→en from 19.86 to 25.04), but yields only marginal gains in COMET and BLEURT, indicating reliance on shallow lexical patterns rather than grounded video–text alignment. Moreover, the dominance of text-sufficient samples during training biases the model toward neglecting visual inputs, impairing generalization to unseen datasets (e.g., VATEX) and causing severe degradation on the Ambiguity subset where multimodal reasoning is essential. By contrast, DART's SFT stage reduces visual neglect and consistently improves BLEU, COMET, and BLEURT across general, Ambiguity, and VATEX benchmarks, indicating more effective use of video information. The subsequent RL stage further enhances reasoning and translation quality, yielding additional performance gains.

The results in Table 3 confirm the distinct functions of the reward components in DART's RL stage. Removing $r_{\text{qual}}$ leaves logical correctness as the sole learning signal, leading to a marked decline in translation performance. Removing $r_{\text{logic}}$ reduces optimization to COMET alone, encouraging evaluator-aligned outputs that yield higher COMET scores but degrade overall translation quality as reflected by other metrics. Eliminating $r_{\text{len}}$ makes training vulnerable to repetitive, overlong outputs, severely slowing optimization and preventing convergence.

### 5.3 Analysis of TVRF Data Construction

Table 4 presents a fine-grained analysis within the TVRF pipeline, examining when and how different multimodal cues contribute to VMT. Across 87K samples, 97.89% can be reliably extracted, yet

| Cue type | Availability | | Total items | Avg. items | Prec. (%) | Helpful (%) |
|---|---|---|---|---|---|---|
| | # Non-empty | % | | | | |
| JSON extraction success: 97.89% on 87K samples | | | | | | |
| *people* | 69K | 80.89% | 134K | 1.57 | 90.5% | 9.92% |
| *objects* | 83K | 96.95% | 498K | 5.81 | 88.5% | 13.97% |
| *actions* | 71K | 83.53% | 201K | 2.34 | 86.0% | 11.97% |
| *OCR* | 70K | 82.00% | 563K | 6.58 | 96.0% | 15.41% |
| *spatial relations* | 84K | 97.95% | 517K | 6.04 | 74.5% | 14.52% |
| *pointing gaze* | 44K | 51.47% | 67K | 0.78 | 83.0% | 7.18% |
| Text-only sufficient (none of the six cues helpful) | | | – | | | 68.88% |

Table 4: Statistics of VMT data analyzed with the TVRF pipeline. "# Non-empty" indicates the number of samples where a cue is present; "Total items" and "Avg. items" denote the total and per-sample average counts, respectively. "Prec." is the manual precision evaluated on 200 samples per cue, and "Helpful" denotes the proportion of samples where the cue benefits translation.

| **Setting** | **General (zh→en)** | **General (en→zh)** | **Ambiguity (en→zh)** | **VATEX (en→zh)** |
|---|---|---|---|---|
| | BLEU ↑ / COMET ↑ / BLEURT ↑ | | | |
| $w_{qual} = 0, w_{logic} = 1.0$ | 21.32/73.25/54.70 | 30.01/74.97/57.06 | 31.15/79.93/59.31 | 25.12/76.10/58.10 |
| $w_{qual} = 0.5, w_{logic} = 0.5$ | 23.84/75.41/58.87 | 32.83/79.02/60.59 | 32.92/81.57/61.95 | 28.61/77.47/59.01 |
| $w_{qual} = 0.7, w_{logic} = 0.3$ | 24.85/76.92/61.02 | 33.79/79.84/61.06 | 34.02/82.31/63.91 | 30.84/78.83/59.02 |
| $w_{qual} = 0.9, w_{logic} = 0.1$ **(Our setting)** | **25.21**/77.04/**61.51** | **34.22**/80.21/**61.73** | **34.60**/82.91/**64.27** | **31.21**/79.91/**59.65** |
| $w_{qual} = 0.95, w_{logic} = 0.05$ | 25.04/77.19/61.25 | 33.98/80.24/61.28 | 33.82/83.04/63.86 | 30.56/79.93/58.20 |
| $w_{qual} = 1.0, w_{logic} = 0$ | 24.79/**77.62**/60.04 | 32.70/**80.44**/60.03 | 31.95/**83.08**/63.21 | 30.45/**80.16**/56.85 |
| $w_{qual} = 1.0, w_{logic} = 0$, text input only | 24.41/77.31/59.78 | 32.65/80.26/59.35 | 31.49/82.93/60.80 | 29.86/80.24/56.23 |

Table 5: Validation of the weighting scheme for translation ($w_{qual}$) and logic ($w_{logic}$) rewards.

68.88% are text-sufficient, indicating that video information is often unnecessary and that unfiltered multimodal training may be inefficient. Among the cues, OCR, objects, and spatial relations are the most helpful, each benefiting translation in about 14-15% of samples. OCR stands out due to its high precision and dense occurrences, supplying explicit visual text that is difficult to infer from subtitles alone. Objects also show broad coverage and strong precision, supporting entity grounding, while spatial relations contribute complementary structural information despite noisier extraction. In contrast, people and actions offer moderate gains, suggesting partial redundancy with textual context, and pointing gaze is the least frequently helpful, reflecting its sparse and highly context-dependent nature. Human annotators were employed to verify the accuracy of the extracted multimodal information; annotator recruitment and compensation are detailed in Appendix F. Overall, TVRF highlights that multimodal cues contribute unevenly, motivating selective, cue-aware utilization rather than uniform video conditioning.

### 5.4 Impact of Reward Weighting Schemes

An ablation study was conducted to determine the optimal balance between translation performance and reasoning integrity, with the results summarized in Table 5. Notably, when $w_{qual} = 1.0$ and $w_{logic} = 0$, the model achieves its highest COMET score. However, because the translation reward ($r_{qual}$) is explicitly based on COMET, this peak represents reward hacking targeted at the COMET metric rather than a genuine improvement in translation quality—a conclusion corroborated by the concurrent decline in both BLEU and BLEURT scores.To avoid this degradation while maintaining balance, we selected the configuration $w_{qual} = 0.9$ and $w_{logic} = 0.1$. This distribution is specifically designed for magnitude alignment rather than a categorical preference for translation quality over reasoning logic. Because the model has undergone a SFT phase prior to reinforcement learning, the translation reward ($r_{qual}$) already operates within a highly optimized, narrow variance (approximately 0.08). In contrast, the logic reward ($r_{logic} \in [0, 1]$) operates on a significantly broader scale. Under this weighting scheme, the effective contribution of the logic reward ($0.1 \times 1.0 = 0.1$) still exceeds that of the translation reward ($0.9 \times 0.08 = 0.072$).

| | | |
|---|---|---|
| Video Clip |  |  |
| Source Sentence | It's a nice day. | Keep it tight, keep it tight. |
| Reference Sentence | 今天天气很好。<br>*(The weather is nice today.)* | 拉紧，拉紧。<br>*(pull it taut, pull it taut.)* |
| LMRM Reasoning | Okay, let me try to figure out how to approach this. The user provided a video with several frames showing ... Let me confirm once more. Yes, that's the standard translation. (737 tokens) | Okay, let's see. The user wants me to translate the sentence "Keep it tight, keep it tight." from English to Chinese ... I think that's correct. So the answer should be "抓紧，抓紧。" (388 tokens) |
| LMRM Output | 今天天气不错。<br>*(The weather is nice today.)* | 抓紧，抓紧。<br>*(Grip tightly, grip tightly.)* |
| DART (Ours) Reasoning | This text can be translated without video information. The translation is: (16 tokens) | To translate this text, video information action is required. It is : pull. So the translation is : (21 tokens) |
| DART (Ours) Output | 今天天气真好。<br>*(The weather is nice today.)* | 拉紧，拉紧。<br>*(pull it taut, pull it taut.)* |

Figure 4: Case study of DART and LMRM on VMT at two ambiguity levels; **blue** marks video-dependent tokens, **green**/**red** indicate correct/incorrect translations, and key reasoning tokens in **purple**.

Consequently, this configuration ensures that the optimization process remains logic-centric while effectively regularizing the model to prevent the aforementioned reward hacking, thereby preserving true translation fluency.

### 5.5 Case Study

Figure 4 compares DART and LMRM on VMT cases with varying disambiguation requirements. In a simple, unambiguous example ("It's a nice day"), LMRM incurs substantial overhead by generating 737 reasoning tokens, whereas DART requires only 16 tokens to conclude that video cues are unnecessary and translates directly from text. In an ambiguous case ("keep it tight, keep it tight"), where the video shows a man pulling a rope-connected trap, LMRM generates 388 reasoning tokens yet misinterprets *keep* as *grab*. In contrast, DART uses 21 tokens to identify the need for action cues, correctly infers *pull*, and produces the correct translation. The reduced token generation in ambiguous cases indicates that LMRM fails to capture disambiguation needs, reflecting limited adaptive reasoning.

## 6 Conclusions

This paper introduces DART, a disambiguation-adaptive translation framework that dynamically selects translation paradigms based on input complexity. To our knowledge, this is the first approach to explicitly incorporate reasoning into VMT task. DART first estimates the ambiguity of each instance: unambiguous cases are translated using text-only information, whereas ambiguous instances trigger explicit reasoning over relevant video-based multimodal cues, which are then integrated into translation. In addition, we propose TVRF to identify VMT samples where video content benefits translation, thereby alleviating long-standing data bias. We demonstrate the effectiveness of the proposed method across multiple metrics and datasets.

## 7 Limitations

Despite the language-agnostic by design of our method, the current scarcity of diverse language pairs in VMT datasets has restricted us from conducting evaluations on a wider scale. Although our 4B-scale DART model already surpasses many widely adopted 8B-scale models in both translation quality and processing efficiency. The substantial computational overhead introduced by reinforcement learning training and video processing makes it difficult, under our constrained computational resources, to scale experiments to larger model sizes.

## Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62476271 and 62336008). We would like to express our sincere gratitude to the anonymous reviewers for their valuable feedback and constructive suggestions.

## References

Sohyun An, Ruochen Wang, Tianyi Zhou, and Cho-Jui Hsieh. 2025. Don't Think Longer, Think Wisely: Optimizing Thinking Dynamics for Large Reasoning Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 46 others. 2025. Qwen3-vl technical report.

Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025a. Evaluating o1-like llms: Unlocking reasoning for translation through comprehensive analysis. *ArXiv*, abs/2502.11544.

Jianghao Chen, Junhong Wu, Yangyifan Xu, and Jiajun Zhang. 2025b. LADM: Long-context training data selection with attention-based dependency measurement for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3076–3090,

Vienna, Austria. Association for Computational Linguistics.

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Soul-mix: Enhancing multimodal machine translation with manifold mixup. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11283–11294, Bangkok, Thailand. Association for Computational Linguistics.

Dipankar Das, Naveen Mellempudi, Dheevatsa Mudigere, Dhiraj Kalamkar, Sasikanth Avancha, Kunal Banerjee, Srinivas Sridharan, Karthik Vaidyanathan, Bharat Kaul, Evangelos Georganas, Alexander Heinecke, Pradeep Dubey, Jesus Corbal, Nikita Shustrov, Roma Dubtsov, Evarist Fomenko, and Vadim Pirogov. 2018. Mixed precision training of convolutional neural networks using integer operations. In *International Conference on Learning Representations*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: LLM Learns When to Think. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland. Association for Computational Linguistics.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, Toronto, Canada. Association for Computational Linguistics.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025a. Video-r1: Reinforcing video reasoning in MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yi Feng, Chuanyi Li, Jiatong He, Zhenyu Hou, and Vincent Ng. 2025b. Multimodal neural machine translation: A survey of the state of the art. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22141–22158, Suzhou, China. Association for Computational Linguistics.

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

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

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*, pages 9274–9287, Vienna, Austria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Weiqi Gu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2021. Video-guided machine translation with spatial hierarchical attention network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 87–92, Online. Association for Computational Linguistics.

Boyu Guan, Chuang Han, Yining Zhang, Yupu Liang, Zhiyang Zhang, Yang Zhao, and Chengqing Zong. 2025a. SHIFT: Selected helpful informative frame for video-guided machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3249–3267, Suzhou, China. Association for Computational Linguistics.

Boyu Guan, Yining Zhang, Yang Zhao, and Chengqing Zong. 2025b. TriFine: A large-scale dataset of vision-audio-subtitle for tri-modal machine translation and benchmark with fine-grained annotated

tags. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8215–8231, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiawei Guo, Feifei Zhai, Pu Jian, Qianrun Wei, and Yu Zhou. 2025. CROP: Contextual region-oriented visual token pruning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9756–9772, Suzhou, China. Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. 2025. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3619–3638, Vienna, Austria. Association for Computational Linguistics.

Lingjie Jiang, Xun Wu, Shaohan Huang, Qingxiu Dong, Zewen Chi, Li Dong, Xingxing Zhang, Tengchao Lv, Lei Cui, and Furu Wei. 2025. Think Only When You Need with Large Hybrid-Reasoning Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Liyan Kang, Luyang Huang, Ningxin Peng, Peihao Zhu, Zewei Sun, Shanbo Cheng, Mingxuan Wang, Degen Huang, and Jinsong Su. 2023. BigVideo: A large-scale video subtitle translation dataset for multimodal machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8456–8473, Toronto, Canada. Association for Computational Linguistics.

Shaharukh Khan, Ayush Tarun, Ali Faraz, Palash Kamble, Vivek Dahiya, Praveen Pokala, Ashish Kulkarni, Chandra Khatri, Abhinav Ravi, and Shubham Agarwal. 2024. Chitranuvad: Adapting multi-lingual LLMs for multimodal translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 839–851, Miami, Florida, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Mingjie Li, Po-Yao Huang, Xiaojun Chang, Junjie Hu, Yi Yang, and Alex Hauptmann. 2023. Video pivoting unsupervised multi-modal machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3918–3932.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, and 3 others. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *ArXiv*, abs/2505.04921.

Guosheng Liang, Longguang Zhong, Ziyi Yang, and Xiaojun Quan. 2025a. ThinkSwitcher: When to Think Hard, When to Think Fast. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5185–5201, Suzhou, China. Association for Computational Linguistics.

Yupu Liang, Yaping Zhang, Zhiyang Zhang, Zhiyuan Chen, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2025b. Improving MLLM's document image machine translation via synchronously self-reviewing its OCR proficiency. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23659–23678, Vienna, Austria. Association for Computational Linguistics.

Yupu Liang, Yaping Zhang, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2025c. Single-to-mix modality alignment with multimodal large language model for document image machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12391–12408, Vienna, Austria. Association for Computational Linguistics.

Chenyu Lin, Cheng Chi, Jinlin Wu, Sharon Li, and Kaiyang Zhou. 2025a. Learning to Think Fast and Slow for Visual Language Models. *Preprint*, arXiv:2511.16670.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1320–1329, New York, NY, USA. Association for Computing Machinery.

Zhengkai Lin, Zhihang Fu, Ze Chen, Chao Chen, Liang Xie, Wenxiao Wang, Deng Cai, Zheng Wang, and Jieping Ye. 2025b. Controlling Thinking Speed in Reasoning Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jinghui Lu, Haiyang Yu, Siliang Xu, Shiwei Ran, Guozhi Tang, Siqi Wang, Bin Shan, Teng Fu, Hao Feng, Jingqun Tang, Han Wang, and Can Huang. 2025. Prolonged Reasoning Is Not All You Need: Certainty-Based Adaptive Routing

for Efficient LLM/MLLM Reasoning. *Preprint*, arXiv:2505.15154.

Jinze Lv, Jian Chen, Zi Long, Xianghua Fu, and Yin Chen. 2025. Topicvd: A topic-based dataset of video-guided multimodal machine translation for documentaries.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning Models Can Be Effective Without Thinking. *Preprint*, arXiv:2504.09858.

OpenAI. 2024a. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2024b. Openai o1 system card. https://openai.com/index/openai-o1-system-card/.

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 (ACL 2002)*, pages 311–318. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300.

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyun Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges. *ArXiv*, abs/2405.12669.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297, New York, NY, USA. Association for Computing Machinery.

Ammon Shurtz, Lawry Sorenson, and Stephen D. Richardson. 2024. The effects of pretraining in video-guided machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15888–15898, Torino, Italia. ELRA and ICCL.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. 2024. Visual Agents as Fast and Slow Thinkers. In *The Thirteenth International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2720–2728. AAAI Press.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyun Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kaiming Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Internvideo2.5: Empowering video mllms with long and rich context modeling. *ArXiv*, abs/2501.12386.

Yusong Wang, Dongyuan Li, Jialun Shen, Yicheng Xu, Mingkun Xu, Kotaro Funakoshi, and Manabu Okumura. 2024. LAMBDA: Large language model-based data augmentation for multi-modal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15240–15253, Miami, Florida, USA. Association for Computational Linguistics.

Chao Wu, Baoheng Li, Mingchen Gao, and Zhenyi Wang. 2025a. From Efficiency to Adaptivity: A Deeper Look at Adaptive Reasoning in Large Language Models. *Preprint*, arXiv:2511.10788.

Di Wu, Seth Aycock, and Christof Monz. 2025b. Please translate again: Two simple experiments on whether human-like reasoning helps translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20435–20451, Suzhou, China. Association for Computational Linguistics.

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

Wenyi Xiao and Leilei Gan. 2025. Fast-Slow Thinking GRPO for Large Vision-Language Model Reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jiafeng Xiong and Yuting Zhao. 2025. GIIFT: Graph-guided inductive image-free multimodal machine translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 98–112, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *ArXiv*, abs/2505.09388.

Jian Yang, Hongcheng Guo, Yuwei Yin, Jiaqi Bai, Bing Wang, Jiaheng Liu, Xinnian Liang, LinZheng Chai, Liqun Yang, and Zhoujun Li. 2024. m3P: Towards multimodal multilingual translation with multimodal prompt. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10858–10871, Torino, Italia. ELRA and ICCL.

Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, Houwen Peng, and Jie Jiang. 2025b. R-4B: Incentivizing General-Purpose Auto-Thinking Capability in MLLMs via Bi-Mode Annealing and Reinforce Learning. *Preprint*, arXiv:2508.21113.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025c. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2376–2385.

Zhishen Yang, Tosho Hirasawa, Mamoru Komachi, and Naoaki Okazaki. 2022. Why videos do not guide translations in video-guided machine translation? an empirical evaluation of video-guided machine translation dataset. *Journal of Information Processing*, 30:388–396.

Jingyang Yi, Justin Wang, and Sida Li. 2025. Shorter-Better: Guiding Reasoning Models to Find Optimal Inference Length for Efficient Reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Ying Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jing Tang, Hongyuan Liu, Qining Guo, and 15 others. 2025a. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *ArXiv*, abs/2509.18154.

Zhuang Yu, Shiliang Sun, Jing Zhao, Tengfei Song, and Hao Yang. 2025b. Imagination and contemplation: A balanced framework for semantic-augmented multimodal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10913–10928, Suzhou, China. Association for Computational Linguistics.

Armel Randy Zebaze, Rachel Bawden, and Benoît Sagot. 2025. Llm reasoning for machine translation: Synthetic data generation over thinking tokens. *ArXiv*, abs/2510.11919.

Jiajie Zhang, Nianyi Lin, Lei Hou, Ling Feng, and Juanzi Li. 2025a. AdaptThink: Reasoning Models Can Learn When to Think. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3716–3730, Suzhou, China. Association for Computational Linguistics.

Xiaoyun Zhang, Jingqing Ruan, Xing Ma, Yawen Zhu, Haodong Zhao, Hao Li, Jiansong Chen, Ke Zeng, and Xunliang Cai. 2025b. When to Continue Thinking: Adaptive Thinking Mode Switching for Efficient Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5808–5828, Suzhou, China. Association for Computational Linguistics.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

Zhihui Zhang, Shiliang Sun, Jing Zhao, Tengfei Song, and Hao Yang. 2025c. VQA-augmented machine translation with cross-modal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10113–10124, Suzhou, China. Association for Computational Linguistics.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Zhiyuan Chen, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025d. A query-response framework for whole-page complex-layout document image translation with relevant regional concentration. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7138–7149, Vienna, Austria. Association for Computational Linguistics.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Cong Ma, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025e. Understand layout and translate text: Unified feature-conductive end-to-end document image translation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3358–3376.

Jiawei Zheng, Feiyan Liu, and Xiaoli Wang. 2025. Seeing through ambiguity: Effective video-guided machine translation via chaotic fusion and causally aligned spatio-temporal attention. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 4837–4845, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. 2024. Sglang: efficient execution of structured language model programs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

# A Implementation Details of TVRF

We initially performed random sampling of 50,000 English-to-Chinese and 50,000 Chinese-to-English instances from the TriFine training split. This initial pool was subsequently refined using the COMET metric to exclude substandard entries, yielding a final dataset of approximately 87,000 high-quality instances.

To balance translation quality and computational efficiency, we leverage the Qwen3-VL-32B-Instruct model deployed via the vLLM (Kwon et al., 2023) framework. This setup extracts detailed multimodal cues from videos and outputs them in JSON format. These cues subsequently facilitate the inference stage. The specific prompt utilized for this extraction process is detailed in Figure 5.

We use COMET as the translation quality evaluation function $m(\cdot)$, and set $\delta = 0.04$ in Equation 4. For each instance, the extracted multimodal information is transformed into natural language descriptions to generate multimodal cues. These cues are subsequently integrated with the source sentence to construct a cue-augmented translation prompt. The architecture of the prompt construction template is depicted in Figure 6. Following the prompt illustrated in Figure 7, the LLM is tasked with generating translations derived exclusively from text-only source sentences, thereby excluding any auxiliary multimodal information.

We completed the entire TVRF pipeline using four NVIDIA A100 80GB GPUs, with a total runtime of approximately 29 hours.

# B Training Settings for DART

## B.1 SFT-stage training of DART.

All experiments are conducted on four NVIDIA A100 80GB GPUs, leveraging PyTorch DDP (Paszke et al., 2019) and DeepSpeed ZeRO (Rasley et al., 2020) Stage-2 for memory efficiency. The model was fine-tuned for one epoch using 18K TVRF-constructed $zh \rightarrow en$ samples and 18K $en \rightarrow zh$ samples. Per-device batch size of 4 and gradient accumulation over 4 steps. Optimization is performed with AdamW (Loshchilov and Hutter, 2019) using a learning rate of $1 \times 10^{-5}$, zero weight decay, cosine scheduling with a 3% warmup ratio, gradient clipping at 1.0, and bfloat16 precision (Das et al., 2018). During training, the vision encoder is frozen, while the multimodal projection and language model parameters are updated. The maximum sequence length is set to 16,384 tokens.

The SFT loss is formulated as in Equation 14.

$$\mathcal{L}_{\text{SFT}}(\theta) = -\,\mathbb{E}_{(x,v,z)\sim\mathcal{D}_{\text{SFT}}} \left[ \sum_{t=1}^{|z|} \log \pi_\theta(z_t \mid z_{<t}, x, v) \right] \tag{14}$$

## B.2 RL-stage training of DART

Building on the model obtained through DART-based SFT, GRPO-based reinforcement learning is further applied to jointly improve reasoning capability and translation quality. Since COMET exhibits relatively low variance and reflects translation quality that is crucial to the task, we set $w_{\text{qual}}$ and $w_{\text{logic}}$ in Equation 9 to 0.9 and 0.1, respectively. The textual similarity function $\text{sim}(\cdot)$ in Equation 11 is computed using a normalized string alignment metric that quantifies the degree of structural overlap between two texts, producing

**TVRF Prompt for Multimodal Cue Extraction**

```
You are a multimodal information extraction engine for video.
Your job: extract factual, observable cues from the given video/frames. Do NOT narrate.
Rules:
1) Only report what is visible/legible/audible in the video.
2) Never hallucinate names, brands, or text. OCR must match exactly what is readable.
3) Use stable IDs across the clip: persons P1,P2,... objects O1,O2,... regions R1,R2,...
4) Output MUST be valid JSON and NOTHING ELSE.

Given the input video, extract the following 6 cue types:
(1) Who-is-who: person/entity identity registry with stable IDs
(2) Object category + object attributes
(3) Action category + action-object binding
(4) OCR on-screen text
(5) Spatial relations (in/on/under + direction)
(6) Pointing/gaze target grounding (this/that target)
Return JSON with this schema:
{
  "people": [
    {
      "person_id":"P1",
      "role_guess":{"role":"<e.g., cashier/teacher/driver/unknown>"},
    }
  ],

  "objects": [
    {"object_id":"O1","category":"<e.g.,
    cup/knife/phone/unknown>","attributes":{"color":"...","state":"..."}}
  ],

  "actions": [
    {
      "action_id":"A1",
      "predicate":"<verb label, e.g., pour/cut/open/hand_over>",
      "agent_id":"P1|O1|unknown",
      "patient_id":"O2|P2|unknown",
      "instrument_id":"O3|unknown"
    }
  ],

  "ocr": [
    {
      "text":"<exact string>"
    }
  ],

  "spatial_relations": [
    {
      "subject_id":"O1|P1",
      "relation":"in|on|under|left_of|right_of|in_front_of|behind|near|towards|away_from|upwar
      d|downward",
      "object_id":"O2|P2|R1",
    }
  ],

  "pointing_gaze": [
    {
      "source_id":"P1",
      "type":"pointing|gaze|head_turn",
      "target_id":"O2|P2|R1|unknown",
      "target_description":"<if target_id unknown, describe the region/object>",
    }
  ]
}

Additional constraints:
- Prefer fewer, high-precision items over many low-confidence items.
- Use "unknown" rather than guessing.
- If no cue of a type exists, return an empty list for that field.
```

Figure 5: Prompt for multimodal cue extraction in the TVRF framework using MLLMs

**TVRF Prompt for Cue-Conditioned Translation**

```
You are a professional translator.
Translate the source sentence into the target language.
Use the provided video cue ONLY as auxiliary context when it helps disambiguate meaning.
Do NOT invent any details beyond the cue.
Output ONLY the final translation, with no explanation.

Task: Translate from [SOURCE LANGUAGE] to [TARGET LANGUAGE].
Source sentence: [SOURCE SENTENCE]
Video cue:[CUE TEXT]
```

Figure 6: TVRF Prompt for Cue-Conditioned Translation

**TVRF Prompt for Text-only Translation**

```
You are a professional translator.
Translate the source sentence into the target language.
Use the provided video cue ONLY as auxiliary context when it helps disambiguate meaning.
Do NOT invent any details beyond the cue.
Output ONLY the final translation, with no explanation.

Task: Translate from [SOURCE LANGUAGE] to [TARGET LANGUAGE].
Source sentence: [SOURCE SENTENCE]
Video cue: No video cue is provided. Translate using text only.
```

Figure 7: TVRF Prompt for Text-only Translation

**TVRF Response Template for Video-Dependent Cases**

```
To translate this text, video information [CUE TYPE] is **required**.
It is:
[CUE CONTENT]
So the translation is:
[REFERENCE SENTENCE]
```

Figure 8: Response template used in TVRF when the textual input requires multimodal cue information from the video to complete the translation.

**TVRF Response Template for Text-Sufficient Cases**

```
This text can be translated **without** video information.
The translation is:
[REFERENCE SENTENCE]
```

Figure 9: Response template used in TVRF when the textual input is sufficiently clear and video information is unnecessary.

a bounded score that reflects their relative correspondence. The parameter $\lambda$ in Equation 11 is set to 0.7, while $\alpha$ in Equation 13 is set to 0.001, with $L_{\text{limit}} = 500$. Balancing stability and efficiency, we adopt SGLang (Zheng et al., 2024) as the rollout engine. The model is initialized from a DART-SFT–finetuned checkpoint and optimized with a learning rate of $5 \times 10^{-7}$ for a total of 5 epochs. Each training batch contains 128 samples, with a maximum prompt length of 8192 tokens and a maximum response length of 1024 tokens. For policy optimization, GRPO is employed with 4 rollouts per prompt, a PPO mini-batch size of 64, and a micro-batch size of 2 per GPU. Training is performed on 4 NVIDIA A100 80 GB GPUs using FSDP with bfloat16 precision, gradient checkpointing enabled, and no parameter or optimizer offloading. KL regularization is applied via a low-variance KL loss with a coefficient of 0.01, while entropy regularization is set to 0.001; the KL term is not included directly in the reward. It requires approximately 40 hours of training.

The GRPO loss is formulated as in Equation 15.

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}_{(x,v)\sim\mathcal{D}_{\text{RL}},\hat{z}\sim\pi_\theta} \left[ A(\hat{z}) \log \frac{\pi_\theta(\hat{z} \mid x, v)}{\pi_{\theta_{\text{old}}}(\hat{z} \mid x, v)} \right] \tag{15}$$

We deployed the COMET model as a low-latency inference service using FastAPI and Python's multiprocessing (spawn mode). To maximize throughput and minimize overhead for RL training, the system utilizes 8 worker processes distributed across two GPUs (4 per device). Each worker is restricted to a single CPU thread to prevent resource contention and employs TensorFloat-32 (TF32) precision along with a warmup mechanism. By bypassing high-level framework wrappers and directly invoking the model's core inference steps, the architecture ensures high-frequency, stable reward feedback.

## C Deployment Details

To ensure fair comparison with prior work (Guan et al., 2025a), we adopt identical prompts for the same input modalities.

**VMT Prompt for Text Inputs**

```
Please translate the following input
sentence from [SOURCE LANGUAGE] to [TARGET
LANGUAGE]. ONLY output the translated
sentence.
Input sentence:
[SOURCE SENTENCE]
Translated sentence:
```

Figure 10: VMT prompt for text inputs.

For text-only inputs (rows 5–9 in Table 1), the prompt shown in Figure 10 is used.

**VMT Prompt for Video–Text Inputs**

```
Please translate the following input
sentence from [SOURCE LANGUAGE] to [TARGET
LANGUAGE] according to the video. ONLY
output the translated sentence.
Input sentence:
[SOURCE SENTENCE]
Translated sentence:
```

Figure 11: VMT prompt for video–text inputs.

For video–text inputs (rows 10–14, 16, and 19 in Table 1), we employ the prompt illustrated in Figure 11.

To verify that MLLMs lack the inherent capability to directly select translation-relevant multimodal cues and perform translation accordingly, we apply the prompt shown in Figure 12, which yields the results reported in rows 15 and 17 of Table 1.

**Multimodal Cue–Based Self-Reasoning for VMT**

```
I will give you an input sentence, which
is a subtitle of a video clip, and I will
also input the corresponding video clip.
I need to translate this input sentence
from [SOURCE LANGUAGE] to [TARGET
LANGUAGE]. Please refer to thevisual cues
in the video, such as people, objects,
actions, OCR, spatial relations, and
pointing/gaze cues when producing the
translation of this sentence.
Input sentence:
[SOURCE SENTENCE]
Translated sentence:
```

Figure 12: Multimodal cue–based self-reasoning for VMT.

**VMT Prompt for Image–Text Inputs**

```
Please translate the following input
sentence from [SOURCE LANGUAGE] to [TARGET
LANGUAGE] according to the image. ONLY
output the translated sentence.
Input sentence:
[SOURCE SENTENCE]
Translated sentence:
```

Figure 13: VMT prompt for image–text inputs.

Finally, since the SHIFT method (row 18 in Table 1) ultimately feeds a single image–text pair into the MLLM, we use the prompt depicted in Figure 13.

## D Robustness to Multimodal Cue Misclassification

Although multiple strategies are employed in both data construction and model training, the model may still make occasional selection errors, albeit at a relatively low frequency. Importantly, such misclassifications do not necessarily translate into noticeable degradation in translation quality. This is because these cases typically exhibit weak or ambiguous multimodal dependency, where the inclusion or omission of multimodal cues has limited impact—particularly for conservative false positives. For instance, a sample that human annotators judge as not requiring multimodal cues may be classified by the model as requiring them, reflecting a more cautious decision-making tendency rather than a critical error; the additional multimodal in-

formation in such cases generally has negligible influence on the final output. In contrast, samples for which accurate multimodal cue selection is crucial to translation quality are, in most cases, correctly identified by the model.

## E More Results

Additional experimental results on the video subtitle benchmark TriFine (including the Chinese–English general test set and the ambiguity test set) and the video description benchmark VATEX are reported in Table 6. The table additionally reports results from directly using the models employed in our data annotation pipeline, namely Qwen3-VL-32B (Row 24) and Qwen3-30B-A3B-Instruct (Row 8). Despite their substantially larger parameter scales, these models exhibit either low efficiency or limited multimodal capability, resulting in inferior overall performance compared to DART. We also include two-step (Row 17) and three-step (Row 18) decision-making processes that mirror the DART logic: the former first determines whether and which multimodal cues are needed before translation, while the latter further decomposes this into deciding cue necessity, selecting cue types, and then generating the translation. The results show that the improvements of DART on VMT cannot be achieved by prompting alone, thereby confirming the effectiveness and necessity of the proposed reasoning paradigm.

## F Human Evaluation

Our evaluators are volunteers recruited from forums; they are current PhD students in computer science with strong bilingual proficiency in both Chinese and English. To ensure evaluation quality, we conduct random checks throughout the process. Given the high level of expertise required, their hourly compensation is approximately twice the local average wage.

A multimodal cue is considered correctly extracted only if it (1) actually appears in the video, and (2) no other translation-relevant visual information has been overlooked. Each sample is independently assessed by at least three annotators, and only judgments supported by a majority of evaluators are adopted.

The detailed guidelines are illustrated in Figure 14.

### F.1 SFT Loss Curve Comparison

As illustrated in Figure 15, under identical experimental settings, the SFT stage of our DART method demonstrates superior convergence behavior compared to conventional SFT. Benefiting from a reasoning process highly aligned with VMT, DART achieves a sharper and more rapid decline in training loss. Notably, the loss curves depicted are calculated based exclusively on generated translation tokens.

## G Baselines

### G.1 Traditional VMT Methods

**Transformer model** (Vaswani et al., 2017). Following the architectural settings of prior VMT frameworks, we incorporate a 6-layer Transformer encoder-decoder as our text-only baseline. The model utilizes a hidden size of 512 and an inner-layer dimension of 2048 for the feed-forward sub-layers.

**TVE and CVE** (Shurtz et al., 2024). Building upon the doubly-attentive Transformer architecture, TVE and CVE process video sequences via uniform sampling at 5 FPS and leverage off-the-shelf CLIP features. The Transformer Video Encoder (TVE) relies on self-attention to aggregate frame-level information, while the Conformer Video Encoder (CVE) enhances this process by interleaving convolutions to exploit local spatial-temporal features. In both frameworks, the visual and textual encoders operate independently. Their respective representations are subsequently fused within the decoder through separate multi-head attention modules, enabling the model to generate the final translation by effectively integrating cues from both the source text and the corresponding video context.

**FIAT** (Guan et al., 2025b). FIAT is an input-augmentation strategy that enhances VMT by substituting redundant video features with discrete, fine-grained multimodal tags extracted from both visual and acoustic modalities. By fusing source sentences with these granular tags, the framework enriches the textual context while preserving the structural integrity of the underlying Transformer backbone. FIAT delivers higher-quality translations while maintaining a substantially lower computational footprint than coarse-grained VMT baselines.

| # | Method | TriFine | | | VATEX | Speed |
|---|---|---|---|---|---|---|
| | | General (zh→en) | General (en→zh) | Ambiguity (en→zh) | Test (en→zh) | |
| | | BLEU ↑ / COMET ↑ / BLEURT ↑ | | | | SPS ↑ |
| ***Traditional VMT Methods*** | | | | | | |
| 1 | Transformer | 23.47/71.89/56.42 | 36.19/75.11/54.42 | 29.74/74.32/52.98 | 29.61/ 73.07/53.82 | **75.38** |
| 2 | TVE* | 23.85/72.58/57.20 | 36.55/75.64/54.98 | 30.37/74.45/55.55 | 30.30 / 73.37 / — | 1.30 |
| 3 | CVE* | 23.97/72.60/57.19 | 36.43/75.58/55.29 | 30.28/74.39/55.55 | 29.40 / 73.44 / — | 1.28 |
| 4 | FIAT* | **25.51**/73.59/57.89 | **38.06**/76.48/56.15 | 31.24/75.93/56.32 | 30.75/73.92/55.43 | 0.71 |
| ***Open-source LLMs (Text-only)*** | | | | | | |
| 5 | Llama-3-8B* | 14.12/72.48/57.08 | 25.00/75.65/55.57 | 22.50/76.65/56.85 | 25.11/75.33/54.94 | 9.25 |
| 6 | Llama-3.1-8B* | 16.68/72.54/55.78 | 25.11/77.66/57.39 | 24.95/77.14/58.91 | 27.81/78.15/57.95 | 9.21 |
| 7 | Qwen3-4B-Instruct | 17.41/74.56/58.79 | 30.42/78.39/59.31 | 30.05/80.15/61.32 | 29.49/77.87/57.56 | 19.82 |
| 8 | Qwen3-30B-A3B-Instruct | 19.29/75.31/60.79 | 31.54/79.05/59.98 | 32.45/80.74/62.50 | 30.51/79.24/58.21 | 4.89 |
| ***Open-source LRMs (Text-only)*** | | | | | | |
| 9 | DeepSeek-R1-Distill-Llama-8B | 13.38/71.34/54.52 | 25.20/74.44/55.53 | 23.53/76.08/57.13 | 24.48/74.81/54.12 | 0.78 |
| 10 | DeepSeek-R1-Distill-Qwen-7B | 11.26/70.66/54.24 | 22.07/73.60/53.25 | 20.54/73.06/53.12 | 22.82/72.64/51.20 | 0.67 |
| 11 | Qwen3-4B-Thinking | 17.37/74.54/58.86 | 30.63/77.93/58.93 | 29.98/79.97/61.07 | 29.97/77.94/57.42 | 0.31 |
| ***Open-source MLLMs (Text & Video)*** | | | | | | |
| 12 | LLaVA-Next-Video-7B* | 12.38/68.65/55.18 | 23.63/73.63/57.26 | 23.66/76.35/58.22 | 25.62/75.45/55.10 | 0.65 |
| 13 | InternVideo2.5-8B* | 19.60/75.55/60.18 | 30.28/77.59/57.85 | 31.49/80.25/61.41 | 30.09/78.25/58.04 | 0.72 |
| 14 | MiniCPM-V-4.5-8B | 22.01/75.98/60.59 | 27.61/78.38/59.07 | 31.92/80.71/61.98 | 28.73/78.44/57.52 | 0.51 |
| 15 | Qwen3-VL-4B-Instruct | 19.86/75.38/60.12 | 31.82/78.99/60.22 | 32.54/81.52/62.92 | 29.76/78.77/58.63 | 3.19 |
| 16 | + Self-reasoning | 18.11/75.15/59.80 | 31.53/78.85/60.19 | 33.13/81.56/63.18 | 29.45/79.04/58.96 | 2.96 |
| 17 | + 2-step prompting | 18.52/75.39/60.21 | 31.80/79.17/60.24 | 33.63/81.89/63.32 | 29.60/79.01/58.84 | 2.93 |
| 18 | + 3-step prompting | 18.49/75.82/60.32 | 31.91/79.10/60.29 | 33.82/81.73/63.44 | 29.62/79.12/58.87 | 1.46 |
| 19 | + SHIFT | 20.96/75.81/60.90 | 33.25/79.31/60.62 | 33.67/81.79/63.05 | 30.32/79.05/59.03 | 1.25 |
| ***Open-source LMRMs (Text & Video)*** | | | | | | |
| 20 | R1-Onevision-7B | 10.22/68.37/54.15 | 26.91/75.67/55.58 | 28.65/78.53/59.11 | 25.54/75.90/55.63 | 0.36 |
| 21 | Video-R1-7B | 19.59/75.36/60.08 | 32.02/79.15/60.02 | 32.89/81.40/62.97 | **31.46**/78.82/58.83 | 0.07 |
| 22 | Qwen3-VL-4B-Thinking | 18.01/74.83/59.17 | 29.50/77.90/58.72 | 31.26/81.18/62.43 | 27.97/78.57/58.14 | 0.06 |
| 23 | + Self-reasoning | 17.72/74.92/59.52 | 28.54/78.48/59.34 | 29.94/81.51/62.50 | 25.50/78.38/57.41 | 0.06 |
| 24 | Qwen3-VL-32B | 21.96/76.85/61.32 | 32.48/79.84/60.51 | 34.05/82.12/**64.31** | 31.10/79.66/59.36 | 0.33 |
| 25 | **DART-4B (Ours)** | 25.21/**77.04/61.51** | 34.22/**80.21/61.73** | **34.60/82.91**/64.27 | 31.21/**79.91/59.65** | 1.93 |

Table 6: Additional results are reported on the TriFine and VATEX benchmark. Each test set is evaluated under three random seeds, with results averaged; statistical tests ($p < 0.01$) verify stability and robustness. SPS (Samples Per Second) denotes average end-to-end inference throughput. For each dataset and metric, the best result is highlighted in **bold**. Rows marked with * are reported from Guan et al. (2025a)

### G.2 Text-only LLMs

**Llama-3-8B-Instruct and Llama-3.1-8B-Instruct** (Grattafiori et al., 2024). Llama-3-8B-Instruct and Llama-3.1-8B-Instruct represent a family of prominent open-source decoder-only Transformers developed by Meta, optimized for complex instruction following. The former is pretrained on a 15-trillion-token corpus and refined via SFT and DPO, while the latter, Llama-3.1-8B-Instruct, expands the context window to 128K tokens and introduces support for eight high-resource languages. Notably, Chinese is excluded from these officially supported languages, and the performance of both models on Chinese-English translation tasks remains outside their guaranteed scope.

**Qwen3-4B** (Yang et al., 2025a). Qwen3-4B is a 4-billion-parameter variant of the Qwen3 series, representing the latest generation of large language models from Alibaba. Pretrained on a vast corpus supporting over 100 languages, Qwen3-4B exhibits robust multilingual capabilities and superior translation performance. A defining characteristic of this model is its dual-mode architecture, which supports both a "thinking mode" for complex logical reasoning and a "non-thinking mode"

**Human evaluation guidelines for assessing the correctness of visual cue extraction**

```
Objective: To evaluate whether the model captures the ground-truth visual context required to
resolve translation ambiguities.

Evaluator Qualifications:
- Evaluators are PhD students in Computer Science with strong bilingual proficiency in Chinese
and English.

Evaluator Materials:
For each evaluation instance, the following materials are provided:
1. Source Sentence: The original text requiring translation (e.g., a video subtitle).
2. Target Sentence: The ground-truth translation in the target language.
3. Video Clip: The 8-10 second video associated with the text.
4. Visual Cue: The model-extracted visual cues (covering People, Objects, Actions, OCR,
Spatial, or Pointing Gaze).

Criteria for Successful Extraction:
A multimodal cue is considered "Correctly Extracted" only if it satisfies both conditions
below:

1. Visibility (Factuality):
- The cue (e.g., specific objects or actions) must actually appear within the 8-10 second video
clip. Any hallucinated information results in immediate failure.

2. Essentiality (No Critical Omission):
- The model must identify ALL visual information mandatory for resolving linguistic ambiguity
in the source text.
- Failure Condition: The extraction is deemed incorrect if the evaluator identifies any
visible visual hint that is helpful for translation but was overlooked by the model.

Quality Control Process:
- Each sample is independently assessed by at least three annotators.
- Final labels are adopted only when supported by a majority vote.
- Random spot checks are conducted to ensure  consistency.
```

Figure 14: Human evaluation guidelines for assessing the correctness of multimodal cue extracted.

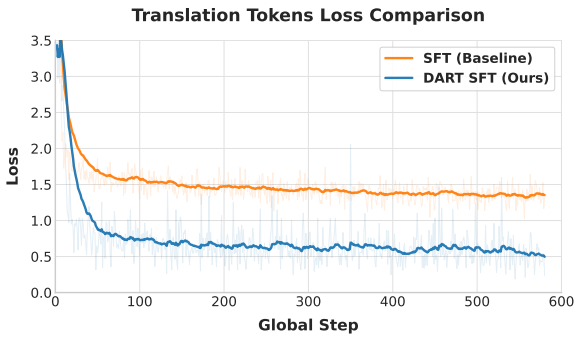


Figure 15: Training loss comparison (translation tokens only) between DART with TVRF and conventional SFT.

for efficient, general-purpose tasks. The transition between these modes is determined by explicit manual selection rather than automated switching. However, a key distinction from our work is that the mode transition in Qwen3-4B is determined by explicit manual selection rather than automated switching. Within the scope of our study, it serves as a sophisticated textual baseline.

**DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B** (DeepSeek-AI et al., 2025). DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B represent a series of dense models developed by distilling the sophisticated reasoning trajectories of DeepSeek-R1 into smaller, more efficient architectures. This distillation process effectively transfers the complex Chain-of-Thought capabilities and logical consistency—initially elicited through large-scale reinforcement learning in the flagship model—to the Qwen2.5 and Llama-3 backbones. Both variants exhibit substantial enhancements in multi-step reasoning, mathematical problem-solving, and coding tasks compared to their non-distilled counterparts. Within the scope of our research, both distilled variants serve as robust, text-only reasoning baselines.

### G.3 MLLMs

**Qwen3-VL-4B-Instruct and Qwen3-VL-4B-Thinking** (Bai et al., 2025). Qwen3-VL-4B-Instruct and Qwen3-VL-4B-Thinking are 4B-parameter multimodal models that natively process

interleaved video–text inputs within a 256K-token context by converting video frames into continuous visual tokens to capture fine-grained spatiotemporal dependencies. The Instruct variant is optimized for multilingual instruction following, whereas the Thinking variant incorporates explicit reasoning to improve logical analysis and temporal grounding in complex videos.

**LLaVA-NeXT-Video-7B** (Zhang et al., 2024). LLaVA-Next-Video-7B leverages Vicuna-7B-v1.5 as its linguistic backbone, optimized through multimodal instruction-tuning across diverse image and video datasets. To enhance temporal reasoning, the model was trained on a comprehensive corpus featuring 100K video-specific instruction pairs from VideoChatGPT-Instruct, supplemented by image-based VQA and GPT-generated synthetic data. This architecture effectively integrates sequential visual frames with textual prompts, enabling robust, context-aware performance on complex video-understanding tasks.

**InternVideo2.5-Chat-8B** (Wang et al., 2025). InternVideo2.5-Chat-8B leverages the InternLM2.5-7B backbone to achieve high-performance bilingual modeling and long-context reasoning. The architecture utilizes an adaptive frame sampling strategy to capture temporal dynamics, followed by a hierarchical spatiotemporal compression pipeline. By integrating spatiotemporal merging and attention-guided pruning, the model efficiently reduces visual redundancy, allowing it to process extended video sequences while retaining the most salient tokens for precise instruction following.

**MiniCPM-V-4.5-8B** (Yu et al., 2025a). MiniCPM-V 4.5 stands as an 8-billion-parameter multimodal model integrated with the Qwen2-7B language backbone, delivering robust bilingual proficiency in English and Chinese alongside sophisticated OCR and reasoning capabilities. To ensure efficient encoding, the model employs a 3D-Resampler to distill features from sampled frames into high-density tokens. This unified architecture effectively captures essential spatiotemporal dynamics while significantly reducing the visual feature dimensionality.

**R1-Onevision-7B** (Yang et al., 2025c). R1-Onevision-7B represents a 7-billion-parameter multimodal reasoning model designed to advance generalized reasoning through a cross-modal formalization framework. This model utilizes the R1 reinforcement learning paradigm to incentivize the generation of structured reasoning trajectories, effectively bridging the gap between visual perception and logical deduction. By formalizing multimodal inputs into intermediate symbolic or structured representations, it demonstrates exceptional proficiency in complex tasks such as mathematical geometry and visual logic.

**Video-R1-7B** (Feng et al., 2025a). Video-R1-7B is a 7-billion-parameter multimodal reasoning model that pioneers the application of the R1 reinforcement learning paradigm to incentivize complex video reasoning. It inherits robust linguistic reasoning and instruction-following capabilities through a cold-start supervised fine-tuning phase using Chain-of-Thought data. The video processing workflow is distinguished by the T-GRPO algorithm, which explicitly rewards the model for utilizing temporal information to solve reasoning tasks. This approach enables the model to perform deep temporal modeling and logical deduction by generating long-form reasoning traces that ground textual answers in dynamic visual evidence. In our study, Video-R1-7B serves as a specialized reasoning baseline.

**SHIFT** (Guan et al., 2025a). SHIFT is a lightweight, plug-and-play framework for video-guided machine translation that adaptively selects the most informative frame from video for MLLMs. The framework utilizes a clustering and selector mechanism to determine the optimal input configuration, feeding either a single key frame combined with the source text or the source text alone into the MLLM to obtain the translation result. This selective input strategy improves efficiency and enhances translation quality.