When Are Images Needed in Multimodal Machine Translation? Text-Only Routing for On-Demand Visual Context

Xinyu Wang[1], Yingfeng Luo[1], Yuxin Gao[1], Anxiang Ma[1,2]*, Jingbo Zhu[1,2], Tong Xiao[1,2]
[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2] NiuTrans Research, Shenyang, China
2601866@stu.neu.edu.cn
{maanxiang,zhujingbo,xiaotong}@mail.neu.edu.cn

Abstract

Image-guided multimodal machine translation (MMT) leverages visual context to resolve translation-relevant uncertainty. However, prevailing systems either process images unconditionally or apply gating only after visual encoding, leaving redundant visual computation unavoidable. To mitigate this computational overhead, routing decisions must rely on a pre-visual source signal. Crucially, surface-level ambiguity does not equate to visual dependence, as intra-sentence textual context often suffices for disambiguation. We therefore formulate visual invocation as source-text sufficiency prediction: determining exclusively from the source text whether translation-relevant uncertainty remains unresolved. To support this formulation, we construct a visual-dependency dataset with three fine-grained labels—`no_ambiguity`, `text_resolved`, and `vision_needed`—and verify discriminative visual evidence for `vision_needed` instances. We further propose PreVIR, a lightweight, backend-agnostic text-only router that selects between text-only and multimodal branches before computing visual features. Experiments on English-to-Chinese MMT demonstrate that PreVIR achieves a favorable COMET–budget trade-off relative to strong routing baselines. At its default operating point, it substantially reduces visual calls while maintaining translation quality comparable to always-on multimodal inference. Label-wise and case analyses show that this trade-off stems from distinguishing genuine visual dependency from contextually resolved ambiguity.

1 Introduction

Image-guided multimodal machine translation (MMT) incorporates aligned visual context to resolve translation-relevant uncertainty in the source sentence. Such uncertainty is especially common in short or underspecified descriptions, where multiple interpretations may lead to different target-language translations (Futeral et al., 2023; Ma et al., 2024). Motivated by this, many prior studies have explored incorporating visual information into machine translation systems (Calixto et al., 2017; Libovický and Helcl, 2017).

Despite evidence that visual contributions in MMT are highly sample-dependent—primarily beneficial when textual cues are insufficient (Elliott, 2018; Caglayan et al., 2019)—prevailing systems still incorporate images unconditionally (Futeral et al., 2023, 2025). While recent adaptive approaches (Liu et al., 2025; Guan et al., 2025, 2026) introduce visual-input selection, their gating mechanisms are typically *post-hoc* (i.e., triggered after visual encoding). Consequently, they fail to eliminate redundant computational overhead and potential visual interference.

Motivated by these limitations, we propose an efficient, pre-visual routing mechanism. This design is feasible because the usefulness of visual context is largely determined by whether translation-relevant uncertainty remains unresolved in the source sentence after accounting for its textual context (Caglayan et al., 2019). The core challenge, then, lies in defining an accurate routing signal from the source. While a naive strategy might trigger visual invocation upon detecting any ambiguous word, lexical ambiguity does not equate to visual dependence. As shown in Figure 1, the local ambiguity of "*bank*" is fully resolved by the intra-sentence context ("of the river"), whereas "*bat*" remains underdetermined without external visual evidence. This suggests that effective routing requires a sentence-level signal that distinguishes unresolved translation uncertainty from ambiguity already resolved by textual context.

In this work, we formulate visual invocation as a source-text sufficiency prediction problem: the source-only router determines whether the source

*Corresponding author.

(a) `no_ambiguity`



Source: The speed limit is 30.
Ambiguity: None.
Resolving evidence: The source is self-contained.
Visual role: Not required.

(b) `text_resolved`



Source: He sat on the bank of the river.
Ambiguity: *bank* may denote a financial institution or a riverbank.
Resolving evidence: "of the river" selects the riverside sense.
Visual role: Not required.

(c) `vision_needed`



Source: He is holding a bat .
Ambiguity: *bat* may denote a baseball bat or the animal.
Resolving evidence: The image shows that the man is holding a baseball bat.
Visual role: Required.

Figure 1: Representative examples of the three visual-dependency labels. Highlighted spans indicate potentially ambiguous expressions. In `text_resolved`, the textual context already determines the intended meaning, whereas `vision_needed` requires discriminative evidence from the paired image.

sentence provides sufficient evidence for the relevant translation choice before any visual features are computed. If the text is insufficient, the paired image is invoked as external evidence; during data construction, we further verify that the image provides discriminative evidence for the unresolved interpretations.

Based on this formulation, we construct a visual-dependency dataset with three labels: `no_ambiguity`, `text_resolved`, and `vision_needed`. We then train a lightweight XLM-RoBERTa-based (Conneau et al., 2020) router that maps the first two categories to a text-only branch and `vision_needed` to a multimodal branch. Since routing relies only on the source sentence and is independent of the translation backend, text-sufficient samples can bypass visual encoding and multimodal fusion entirely.

We evaluate the proposed router against rule-based, LLM-based, and learned baselines, and examine end-to-end translation under varying visual-call budgets. Our approach substantially reduces visual invocation while retaining most of the gains of always-on multimodal inference. Label-wise and case analyses further show that translation gains are concentrated in `vision_needed` samples, highlighting the importance of distinguishing unresolved from text-resolved ambiguity.

Our main contributions are threefold:

- We formulate selective visual use in MMT as a source-only, pre-visual routing problem based on source-text sufficiency, thereby distinguishing visual necessity from the mere presence of ambiguity.
- We construct a visual-dependency dataset with three fine-grained labels—`no_ambiguity`, `text_resolved`, and `vision_needed`—and verify that the paired images of visually dependent samples contain discriminative evidence.
- We develop a lightweight, translation-backend-independent routing pipeline that substantially reduces visual processing while retaining most of the translation benefit provided by always-on multimodal inference.

2 Related Work

2.1 Multimodal Machine Translation and the Role of Vision

Image-guided MMT uses aligned images to resolve linguistic uncertainty and enrich source representations. Multi30K established the standard image-caption translation setting (Elliott et al., 2016). Early systems incorporated global or spatial features through visual and multi-source attention (Calixto et al., 2017; Libovický and Helcl, 2017), while recent work adapts pretrained text-only translators or vision–language models (Futeral et al., 2023, 2025). Nevertheless, most systems still condition every sample on its image.

A parallel line of work has questioned whether MMT systems actually use this visual input. Ad-

versarial image replacement or shuffling often has little effect on translation quality, suggesting weak visual grounding (Elliott, 2018). Source-side masking experiments show that images become substantially more useful when textual evidence is removed, whereas intact captions are frequently sufficient by themselves (Caglayan et al., 2019). Subsequent analyses attribute some apparent multimodal gains to regularisation rather than grounded visual use and show that conclusions about visual utility depend strongly on benchmark composition and image–text alignment (Wu et al., 2021; Li et al., 2021; Long et al., 2024). Li et al. (2022) further show that stronger visual encoders can improve performance on text-masking probes, while gains on standard MMT benchmarks may not reliably reflect effective use of visual information. These studies establish that the contribution of vision is heterogeneous, but they primarily diagnose visual use at the model or dataset level. Our work instead turns this observation into a sample-level decision made before any visual features are computed.

2.2 Ambiguity-Centric Evaluation and Visual-Dependency Data

Several datasets target cases where images can affect translation. Multimodal Lexical Translation annotates ambiguous words and their target-language lexicalisations in visual and sentential context (Lala and Specia, 2018). CoMMuTE pairs an ambiguous source sentence with contrastive images and translations corresponding to different senses, enabling targeted evaluation of visual disambiguation (Futeral et al., 2023). 3AM provides 26K English–Chinese image–sentence pairs selected using word-sense disambiguation signals (Ma et al., 2024), while VIDA curates 2,500 visually dependent instances and introduces span-level, disambiguation-centric evaluation (Pan et al., 2026).

These resources provide increasingly challenging tests of visual grounding, but they do not directly supply the mixed supervision required by a visual-invocation router. Multi30K mixes visually useful and text-sufficient samples without labeling them; 3AM identifies potential ambiguity but not whether sentential context resolves it; and CoMMuTE and VIDA emphasize visually dependent cases. Because ambiguity does not necessarily imply a need for vision, we distinguish `no_ambiguity`, `text_resolved`, and `vision_needed`, verifying both textual insufficiency and discriminative evidence in the image. This taxonomy supplies positive and negative supervision for pre-visual routing.

2.3 Selective Visual Use and Translation Routing

DeDiT is the closest prior work: it detects contextually unresolved expressions and performs visual disambiguation only when needed (Liu et al., 2025). However, because the image is already included in the vision–language model input, it does not avoid visual encoding. Similarly, SHIFT selects between text-only and key-frame inputs only after encoding candidate video frames (Guan et al., 2025), while DART learns direct or cue-grounded translation within a unified video-conditioned model (Guan et al., 2026). These methods adapt visual use after or within visual processing, whereas our text-only router makes the decision before the visual branch is invoked.

Related routing methods pursue different objectives. Using different model scales from the NiuTrans.LMT family as its small- and large-model translation backends (Luo et al., 2026b), RouteLMT predicts the marginal gain of invoking the larger text-only translator and routes requests between model sizes under a fixed call budget (Luo et al., 2026a). VisNec, in contrast, estimates visual contribution through with- and without-image loss differences for multimodal data selection (Dong et al., 2026). Our work differs from both settings: rather than allocating requests between model capacities or selecting training data, we determine whether an additional visual modality should be invoked based on source-text sufficiency. This pre-visual, translation-backend-independent design allows text-sufficient inputs to bypass visual encoding and multimodal fusion entirely.

3 Methodology

To enable routing prior to visual encoding, we formally define the routing task and design a complete on-demand translation pipeline guided by a text-only router. As illustrated in Figure 2, the three fine-grained visual-dependency labels are mapped to binary supervision for training the router. At inference time, the router predicts visual necessity from the source sentence alone, allowing text-sufficient samples to bypass visual encoding and multimodal fusion entirely.

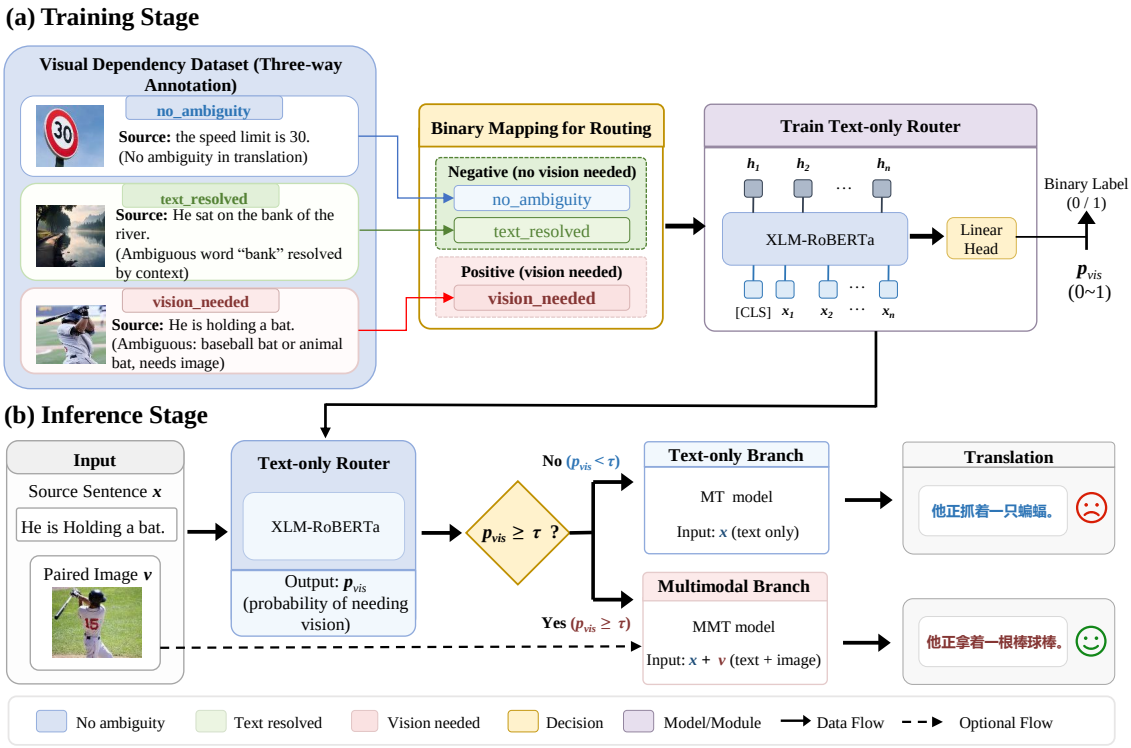


Figure 2: Overview of the proposed pre-visual routing framework. (a) A three-way visual dependency dataset is first constructed with the labels `no_ambiguity`, `text_resolved`, and `vision_needed`. These annotations are then mapped to binary routing labels to train a text-only router based on XLM-RoBERTa. (b) At inference time, the router predicts $p_{\text{vis}} = P(\text{vision_needed} \mid x)$ from the source sentence alone. The input is routed to the text-only branch when $p_{\text{vis}} < \tau$, and to the multimodal branch otherwise. Both branches use the same translation backend and differ only in whether the paired image is provided.

3.1 Problem Formulation

Let M be a translation backend supporting both text-only and image-text inputs. Each translation request consists of a source sentence x and its paired image v. The system can utilize two branches of the backend: a computationally efficient text-only branch M_t and a resource-intensive multimodal branch M_m. For a request (x, v) these two branches generate translations as follows:

$$\hat{y}_{\text{t}} = M_t(x), \qquad \hat{y}_{\text{m}} = M_m(x, v) \tag{1}$$

The router selects which branch to use for serving the request, yielding the final output $\hat{y}$.

To determine whether the source sentence x requires additional visual information before accessing the image, we design a pre-translation text-only router. Let $z(x) \in \{0, 1\}$ denote the routing decision. Here $z = 1$ indicates routing to M_m, meaning that relying solely on the source text is insufficient to support reliable translation, and the paired image may provide discriminative evidence for semantic choices. Conversely, $z = 0$ indicates routing to M_t, meaning that the text information in the source sentence is already sufficient to support translation generation in the current direction.

We model the router's decision-making as a probabilistic prediction task. Given the input source sentence x, the text router predicts the probability of requiring visual context:

$$p_{\text{vis}} = P_\theta(z = 1 \mid x) \tag{2}$$

where θ denotes the parameters of the router, and p_{vis} represents the probability that the source sentence has a visual demand.

Given a routing threshold τ, the visual necessity probability is converted into a discrete routing decision:

$$\hat{z} = I\left[p_{\text{vis}} \geq \tau\right] \tag{3}$$

where $I[\cdot]$ is the indicator function. When $p_{\text{vis}} \geq \tau$, the system considers the current source sentence to have a high demand for vision and routes it to the multimodal translation branch; otherwise, the system deems the textual information sufficient and utilizes the text-only translation branch.

Accordingly, the entire on-demand multimodal translation process can be formalized as:

$$\hat{y} = \begin{cases} M_t(x) & \hat{z} = 0, \\ M_m(x, v) & \hat{z} = 1, \end{cases} \tag{4}$$

3.2 Router

Architecture. To predict the visual necessity p_{vis} defined in Eq. 2, we employ XLM-RoBERTa-base as a lightweight text-only router. Given a source sentence x, we extract the contextualized representation of its start token (e.g., <s>) to serve as the sentence-level embedding:

$$h_x = \text{XLMR}_{\theta_e}(x) \tag{5}$$

where θ_e denotes the parameters of the text encoder. Because h_x is jointly encoded by the complete intra-sentence context, this representation can simultaneously capture potential ambiguous expressions and their contextual disambiguation cues. We then apply a linear classification head to map h_x to the routing probability:

$$\boldsymbol{s} = W h_x + b, \qquad p_{\text{vis}} = \text{softmax}(\boldsymbol{s})_1 \tag{6}$$

where $W \in R^{2\times d_h}$ and $b \in R^2$ are learnable parameters, and the subscript 1 indexes the `vision_needed` class. This continuous probability is subsequently thresholded by τ to yield the discrete routing decision $\hat{z}$ (Eq. 3).

Training Objective. In practice, the distribution of visual dependency is inherently skewed, as genuine multimodal requirements are relatively sparse compared to text-sufficient samples. To prevent majority-class bias during training, we optimize the router using a class-weighted cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{1} w_c y_{i,c} \log P_\theta(z = c \mid x_i) \tag{7}$$

where $y_{i,c}$ is the ground-truth routing label and $w_c = N/(2N_c)$ is the inverse frequency weight for class c (N_c being the sample count of class c).

We conduct full-parameter fine-tuning to jointly optimize the text encoder and the classification head. Guided by this balanced objective, the router learns to effectively weigh ambiguous expressions against their textual constraints to determine source-text sufficiency.

Table 1: Taxonomy of visual-dependency labels.

Label	Ambiguity	Text sufficient	Image required
`no_ambiguity`	No	Yes	No
`text_resolved`	Yes	Yes	No
`vision_needed`	Yes	No	Yes

4 Dataset

To train and evaluate the proposed pre-visual router, we construct a visual-dependency dataset by re-annotating samples from 3AM, VIDA, and CoMMuTE (Ma et al., 2024; Pan et al., 2026; Futeral et al., 2023). The construction process combines translation-relevant ambiguity annotation, source-text sufficiency assessment, visual evidence verification, and human quality control. The following sections describe the label taxonomy, data construction procedure, and resulting dataset statistics.

4.1 Label Taxonomy

As summarized in Table 1, we determine ground-truth visual dependency through three progressive criteria: potential translation ambiguity, textual sufficiency, and visual validity. This sequential verification yields three distinct categories. **`no_ambiguity`** denotes sentences lacking any translation-relevant ambiguity. **`text_resolved`** indicates that while local ambiguities exist, intra-sentence contextual cues fully determine the intended meaning. For both of these classes, image access is unnecessary. Conversely, **`vision_needed`** signifies that the source text contains unresolved ambiguities, and the paired image explicitly provides the discriminative evidence required for accurate translation. Instances where both the text and image fail to resolve the ambiguity are discarded during data construction.

4.2 Data Construction

Figure 3 illustrates our three-stage annotation pipeline, followed by human quality control. Before annotation, we apply rule-based filtering and source-sentence deduplication, then retain image–text pairs only when the paired source sentence ranks within the top 0.5% of all texts for its image under SigLIP similarity. Detailed filtering rules, adjudication criteria, and quality-control statistics are provided in Appendix A.

Ambiguity candidate mining. DeepSeek-V4-Flash extracts potentially ambiguous spans, am-

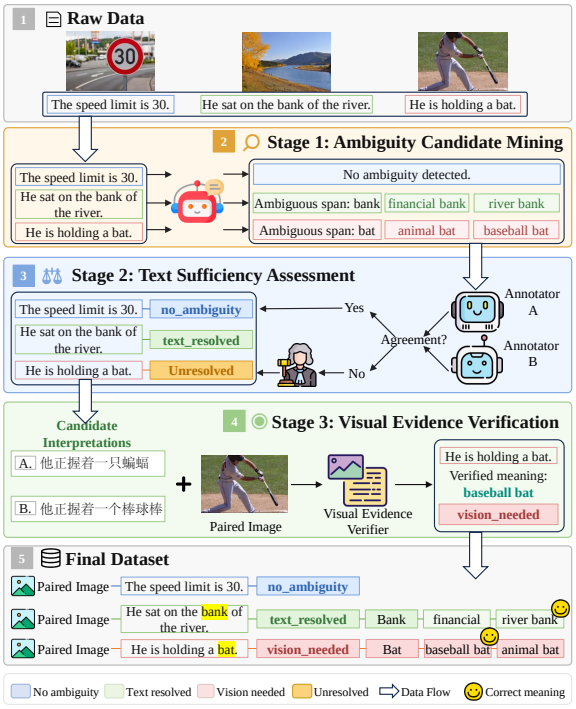


Figure 3: Overview of the visual-dependency data construction pipeline. Three examples illustrate how ambiguity candidate mining, text sufficiency assessment, and visual evidence verification produce the final labels: `no_ambiguity`, `text_resolved`, and `vision_needed`. Only textually unresolved samples undergo visual verification.

biguity types, and candidate interpretations from each source sentence. We retain candidates whose plausible interpretations yield different target-language translations, and use spaCy part-of-speech analysis to filter out punctuation, function words, and unlikely ambiguity candidates. Figure 3 illustrates this step with *bank* and *bat*.

Text sufficiency assessment. Given the source sentence, candidate interpretations, and target language, Qwen-Max and DeepSeek-V4-Pro independently assign `no_ambiguity`, `text_resolved`, or `unresolved`, without access to the paired image. Their exact label agreement is 79.04%. Agreements are retained directly, while disagreements are adjudicated by GPT-5.5 based on label validity, ambiguity interpretation, textual evidence, and reasoning consistency. The top annotation is accepted if its score is at least 11 out of 16 and leads by at least 2 points. Cases meeting the score threshold but not the margin threshold undergo human review; lower-scoring cases are discarded.

Visual evidence verification. For each `unresolved` sample, we convert the candidate interpretations into complete descriptions associated with distinct Chinese translations and use SigLIP to score their compatibility with the paired image. A sample is automatically labeled `vision_needed` when the top score is at least 0.60 on a $[0, 1]$ scale and its margin over the runner-up is at least 0.08. Cases meeting the score threshold but not the margin threshold undergo human review; those below the score threshold are discarded. The two text-sufficient classes bypass this stage.

Table 2: Statistics of the visual-dependency dataset. `no_vision` combines `no_ambiguity` and `text_resolved` as the negative class, while `vision_needed` is treated as the positive class.

Split	No amb.	Text res.	No vis.	Vis. req.	Total
Train	832	616	1,448	973	2,421
Dev	104	77	181	107	288
Test	104	78	182	162	344
Total	1,040	771	1,811	1,242	3,053

Human quality control. Two computer science master's students, both native Chinese speakers proficient in English, independently review borderline cases and screen all 3,300 candidate instances for label correctness, interpretation plausibility, evidence sufficiency, and translation relevance. Their raw pass/fail agreement before discussion is 77%, with disagreements resolved through discussion. Supported label errors are corrected, while 247 invalid instances (7.48%) are removed, yielding the final 3,053-instance dataset.

4.3 Dataset Statistics

The dataset is constructed primarily from 3AM and VIDA, with deduplicated and re-annotated CoMMuTE instances included to evaluate generalisation to an independent data source. Exact and near-duplicate removal is performed before data splitting to prevent overlap across partitions. After annotation and quality control, the final dataset contains 3,053 instances. The label distribution across the training, development, and test sets is reported in Table 2.

5 Experiments

5.1 Data and Model

Data To adapt the dataset for the binary routing task, we group `no_ambiguity` and `text_resolved` into a single negative class (*vision-*

unnecessary), retaining `vision_needed` as the positive class (*vision-required*). As detailed in Table 2, the dataset exhibits a mild class imbalance (roughly a 2:3 positive-to-negative ratio), which we address by applying inverse-frequency class weights derived from the training set. The test split is then used for a dual evaluation: assessing the standalone classification accuracy of the router and measuring the end-to-end performance of the on-demand translation pipeline.

Model To prevent different model architectures from introducing confounding variables into the analysis of visual contributions, our adaptive framework employs a unified Vision-Language Model (VLM) backend. The text-only and multimodal branches strictly share the same parameters, instructions, and decoding configurations; the only difference is the presence or absence of visual input.

Under a unified zero-shot evaluation on Multi30K, Qwen3.5-9B achieves the best macro-average performance among the candidate VLMs (26.11 BLEU and 81.10 COMET; Figure 4). We therefore use it as the shared backend for both translation branches.

5.2 Baselines

To comprehensively evaluate the effectiveness of our method, we establish two sets of evaluation settings: routing baselines to assess the standalone classification accuracy of our text-only router, and end-to-end translation strategies to demonstrate the efficacy of the entire on-demand pipeline.

Routing Baselines To evaluate the classification performance of our fully fine-tuned text-sufficiency router (based on XLM-RoBERTa), we compare it against the following baselines:

- **Lexicon-based Heuristic**: A rule-based baseline that triggers the multimodal branch if the source sentence contains words from a predefined ambiguous lexicon; otherwise, it relies solely on text.
- **LLM Zero-Shot**: A prompting-based baseline where a state-of-the-art LLM (DeepSeek-V4-Pro) is instructed to determine zero-shot whether the current sentence requires visual supplementation based on textual semantics.
- **Frozen Encoder + Linear**: An ablation baseline that freezes the XLM-RoBERTa backbone and only trains the linear classification head, evaluating the necessity of full-parameter fine-tuning.

End-to-End Translation Baseline We compare PreVIR with endpoint policies, budget-matched routing baselines, and oracle references. **Text-only** and **Always-Multimodal** route all samples to the text-only and multimodal branches, respectively. For a fair comparison, **Random**, **Ambiguity-based**, **Length-based**, **Sentinel-src-25**, and **PreVIR** operate under the same 46.22% visual budget. Random routing selects samples uniformly; Ambiguity-based routing prioritizes sentences containing more lexical ambiguity cues; Length-based routing prioritizes shorter sentences with less textual context; and Sentinel-src-25 (Proietti et al., 2025) prioritizes sentences with lower predicted translation quality. PreVIR ranks samples by their predicted visual-necessity probabilities. Finally, the **Gold-label Router** uses the annotated visual-dependency labels, while the **Gain-Oracle** selects the multimodal output only when it yields a positive reference-aware COMET gain.

5.3 Implementation Details

Router Training. We fully fine-tune XLM-RoBERTa-base for 20 epochs on a single NVIDIA A800 80GB GPU. To strictly prevent information leakage, the router accesses only the source text during both training and inference. We optimize the model using AdamW with a learning rate of 2×10^{-5} and a batch size of 8. To mitigate class imbalance in our training data, we employ a weighted cross-entropy loss with inverse-frequency weights (0.836 for the negative class and 1.244 for the positive `vision_needed` class). At inference, the threshold τ is set to 0.5.

Translation Backend. We deploy Qwen3.5-9B using vLLM (Kwon et al., 2023) in bfloat16 precision. For both branches, we apply deterministic greedy decoding (temperature = 0, top-p=1.0). All generated outputs undergo a post-processing step to strip non-translation artifacts such as code blocks, thinking traces, and prefixes.

6 Results and Analysis

6.1 Routing Results

Table 3 reports routing performance at the default threshold $\tau = 0.5$. A useful pre-visual router

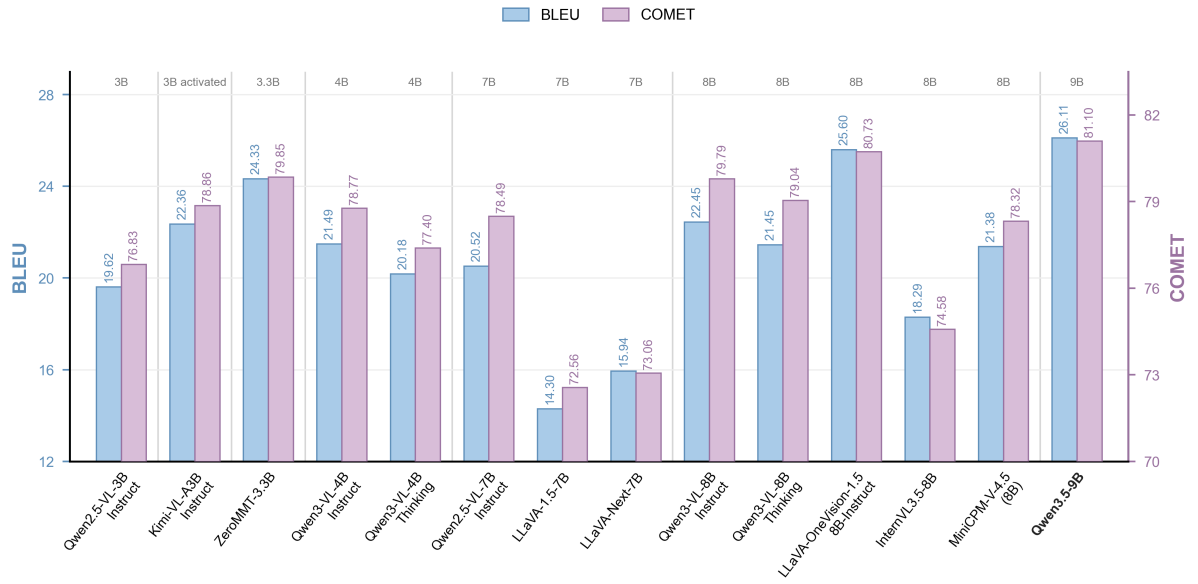


Figure 4: Translation performance of candidate vision-language models on Multi30K, macro-averaged over the En→De, En→Fr, and En→Cs directions. Models are ordered by their advertised parameter scales.

Table 3: Visual-dependency routing results on the test set. ZS denotes zero-shot inference. P, R, and F1 denote precision, recall, and F1-score, respectively, for the positive `vision_needed` class. Call denotes the percentage of samples routed to the multimodal translation path.

Method	Acc.↑	P↑	R↑	F1↑	Call (%)
Always Neg.	0.53	0.00	0.00	0.00	0.00
Lexicon Rule	0.44	0.42	0.48	0.45	53.78
DeepSeek-ZS	0.84	**0.91**	0.73	0.81	38.08
Frozen XLM-R	0.47	0.47	**1.00**	0.64	100.00
PreVIR (Ours)	**0.86**	0.86	0.85	**0.85**	46.22

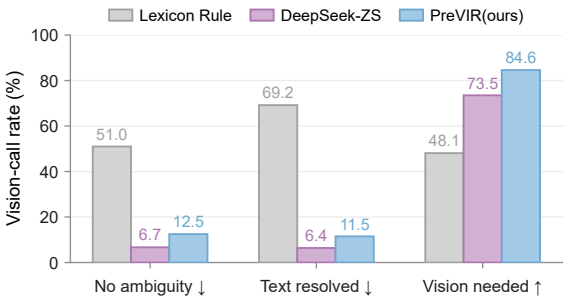


Figure 5: Vision-call rates across the three fine-grained visual-dependency categories. Lower rates are preferred for `no_ambiguity` and `text_resolved`, while a higher rate is preferred for `vision_needed`.

must balance two competing objectives: covering `vision_needed` instances so that potentially useful visual evidence is not missed, while avoiding unnecessary visual calls for text-sufficient inputs. The two extreme baselines illustrate why no single metric is sufficient for evaluating this trade-off. Always Neg. benefits from the negative-class majority but fails to identify any visually dependent sample, whereas Frozen XLM-R attains perfect recall only by routing every input to the multimodal branch. Among the non-degenerate strategies, PreVIR achieves the best overall accuracy and F1 score. DeepSeek-ZS maintains the highest precision but follows a more conservative routing policy, missing more than one quarter of the `vision_needed` instances. PreVIR improves their coverage while retaining selective visual invocation, yielding a stronger overall balance between missed visual evidence and unnecessary calls.

The category-wise results in Figure 5 explain where this advantage comes from. The Lexicon Rule invokes vision more frequently for `text_resolved` samples than for `vision_needed` samples (69.2% versus 48.1%), showing that the presence of a potentially ambiguous expression is poorly aligned with visual dependency when the surrounding text has already resolved its meaning. In contrast, PreVIR invokes vision for only 11.5% of the `text_resolved` samples while covering 84.6% of the `vision_needed` samples. This separation is consistent with the router using sentence-level textual context rather than treating ambiguity

itself as a visual trigger.

Overall, the results support source-text sufficiency as an effective signal for pre-visual routing: task-specific supervision does not simply increase the number of visual calls, but directs them toward samples whose translation-relevant uncertainty remains unresolved by text. The remaining routing errors also indicate room for improving the identification of difficult visual-dependency cases, whose downstream translation effects are examined in the following experiments.

6.2 End-to-End Translation Results

Table 4 compares endpoint policies, budget-matched routing strategies, and oracle diagnostics in terms of translation quality and visual invocation rate using Qwen3.5-9B as the translation backend. Results for additional backends are reported in Appendix B.

Budget-Matched Comparison. Under the same 46.22% visual budget, PreVIR achieves the highest overall COMET score. Although the Ambiguity-based strategy obtains the best overall BLEU, PreVIR performs best on the `vision_needed` subset in both metrics, outperforming it by 1.70 BLEU and 1.03 COMET. Meanwhile, COMET differences on the two text-sufficient subsets remain within 0.25 points. These results indicate that PreVIR allocates more of the fixed visual budget to genuinely vision-dependent inputs.

Quality–Efficiency Trade-off. Compared with Text-only translation, PreVIR improves BLEU by 0.97 and COMET by 1.72. It remains within 0.40 BLEU and 0.32 COMET of Always-Multimodal while reducing visual invocation by 53.78%. Paired significance tests show that the COMET improvement over Text-only is statistically significant, whereas the BLEU improvement is not. Neither metric differs significantly from Always-Multimodal. Detailed confidence intervals and significance tests are reported in Appendix C. Therefore, selective routing recovers most of the benefit of visual context with fewer than half of the visual calls. At matched request batch sizes, PreVIR also improves throughput by 18.7–46.1% over Always-Multimodal, including router overhead. Detailed settings and timing results are provided in Appendix D.

Oracle Diagnostics. The Gold-label Router exceeds PreVIR by only 0.43 BLEU and 0.19 COMET, showing that relatively little performance is lost due to routing prediction errors. The Gain-Oracle achieves substantially higher performance by selecting outputs according to their realized reference-aware gains. This gap suggests that annotated visual necessity and actual backend gain are related but not identical, as the latter also depends on image informativeness and the backend's ability to exploit visual evidence.

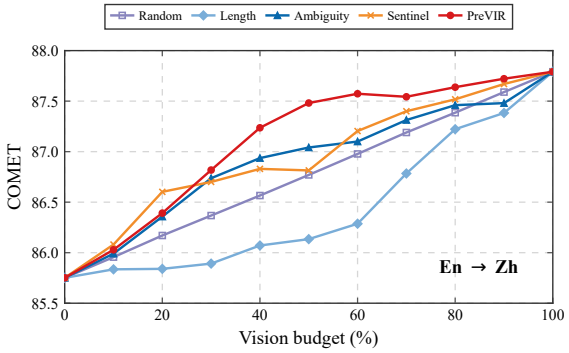


Figure 6: Quality—budget trade-offs of hybrid translation routing.

6.3 Quality–Budget Trade-off

We further explored the performance of various methods at different visual invocation rates. As shown in Figure 6, Sentinel-src-25 performs better under extremely restricted budgets of 10% and 20%. PreVIR becomes the strongest method at 30% and achieves the highest COMET at every evaluated budget from 30% to 90%, with its clearest advantage appearing in the practical 40–60% range. This trend shows that the improvement observed at the default 46.22% operating point extends across a broad range of vision budgets.

The relative performance of the heuristic methods further illustrates the importance of the routing signal. Length-based routing remains below random routing at every intermediate budget, indicating that sentence length provides an ineffective prioritization signal in this setting. Ambiguity-based routing performs better than random at most budgets, suggesting that lexical ambiguity contains useful coarse information. Nevertheless, it remains below PreVIR throughout the 10–90% range, particularly at intermediate budgets. This consistent gap supports the distinction between detecting potentially ambiguous expressions and determining whether the corresponding uncertainty remains unresolved by the textual context.

Table 4: Overall and label-wise translation results. No amb., Text res., and Vision req. correspond to the `no_ambiguity`, `text_resolved`, and `vision_needed` gold labels, respectively. Random, Ambiguity-based, Length-based, Sentinel-src-25, and PreVIR are evaluated under the same 46.22% vision budget. Text-only and Always-Multimodal are endpoint policies, while the oracle methods are diagnostic references evaluated at their native call rates. Bold and underlined values denote the best and second-best results, respectively, among the budget-matched routing methods.

	BLEU ↑				COMET ↑				
Strategy	Overall	No amb.	Text res.	Vision req.	Overall	No amb.	Text res.	Vision req.	Vision Call
Gain-Oracle	44.83	49.66	45.57	38.05	88.93	90.70	88.56	87.96	41.57%
Gold-label Router	41.96	–	–	–	87.66	–	–	–	47.09%
Text-only	40.56	46.24	42.88	31.19	85.75	89.19	87.31	82.79	0.00%
Always-Multimodal	41.93	45.65	43.27	35.73	87.79	89.51	87.45	86.85	100.00%
Random	41.19	45.98	<u>43.03</u>	33.29	86.69	89.35	87.37	84.66	46.22%
Ambiguity-based	**42.03**	**47.51**	**43.08**	<u>34.12</u>	<u>87.04</u>	**89.56**	87.45	<u>85.23</u>	46.22%
Length-based	41.41	47.11	42.98	32.77	86.11	89.44	87.33	83.39	46.22%
Sentinel-src-25	<u>41.63</u>	<u>47.27</u>	42.80	33.16	86.83	<u>89.49</u>	**87.58**	84.76	46.22%
PreVIR (Ours)	41.53	46.36	41.36	**35.82**	**87.47**	89.31	<u>87.55</u>	**86.26**	46.22%

The curve also reveals a limitation of necessity-based routing. Although PreVIR is optimized to identify `vision_needed` samples, Sentinel-src-25 performs better when only the top 10–20% of inputs can invoke vision. This suggests that identifying whether vision is needed is not identical to ranking the magnitude of the translation gain realized by a particular multimodal backend. By the 50% budget point, PreVIR has already selected 143 of the 162 gold `vision_needed` samples and retains approximately 84.8% of the full COMET improvement delivered by Always-Multimodal inference. Consequently, the curve largely plateaus beyond this point, with minor non-monotonic fluctuations. These results characterize PreVIR as an effective visual-necessity router while leaving backend-specific gain prediction as a complementary direction for future work.

6.4 Case Studies

As shown in Table 5, the successful routing cases (TP and TN) demonstrate PreVIR's ability to distinguish unresolved visual dependency from surface-level lexical ambiguity. Our router correctly invokes the multimodal branch when the source text lacks cues to disambiguate terms like *dummies*. Conversely, it accurately bypasses visual processing when intra-sentence context (e.g., *baseball*) already resolves local ambiguities like *pitcher* and *pitch*, thereby eliminating redundant computation.

Furthermore, the error cases (FN and FP) highlight a critical asymmetric risk in multimodal routing. Under-invocation (FN) on genuinely vision-dependent instances (e.g., *showers*) deprives the model of necessary evidence, leading to severe adequacy drops. In contrast, over-invocation (FP) on text-sufficient instances primarily wastes the computational budget and introduces minor quality fluctuations without reliable gains. This asymmetry underscores that maintaining high recall on truly vision-required samples is paramount for robust on-demand MMT.

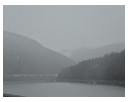
7 Conclusion

In this paper, we formulate visual invocation in MMT as source-text sufficiency prediction and propose PreVIR, a lightweight, translation-backend-independent router. By distinguishing genuine visual necessity from contextually text-resolved ambiguity, we construct a fine-grained dataset to supervise this pre-visual routing. Experiments demonstrate that PreVIR establishes a highly favorable performance-efficiency trade-off, preserving the vast majority of multimodal translation gains while substantially reducing visual calls. Label-wise, budgeted, and case analyses confirm that PreVIR succeeds by strictly allocating visual computation to genuinely vision-dependent samples. Ultimately, our findings establish source-only pre-visual routing as a highly efficient and principled alternative to indiscriminate multimodal processing.

Limitations

Our study presents three main limitations. First, our 3,053-instance dataset focuses on English-to-Chinese translation. Because visual necessity inherently depends on target-language lexicalization,

Table 5: Case studies of pre-visual routing. Highlighted spans indicate translation-relevant ambiguity or its corresponding realization in the translation. COMET scores are reported for the two translation branches.

Case	Image	Item	Text	COMET
Multimodal Branch Vision req. Δ=+40.308		Source	The dummies need to be cleaned.	–
		Reference	奶嘴 需要清洗。	–
		Text-only	假人 需要被清理。	57.642
		Multimodal	奶嘴 需要清洗。	97.950
Text-only Branch Text res. Δ=0.000		Source	A **baseball** pitcher is throwing a pitch .	–
		Reference	一个棒球 投手 正在投出一个 球 。	–
		Text-only	一名棒球 投手 正在投 球 。	92.267
		Multimodal	一名棒球 投手 正在投 球 。	92.267
Text-only Branch Vision req. Δ=+30.650		Source	I know – good job there are showers .	–
		Reference	我知道，幸运的是下了 阵雨 。	–
		Text-only	我知道——太好了，这里有 淋浴间 。	55.420
		Multimodal	我知道——幸好有 阵雨 。	86.070
Multimodal Branch Text res. Δ=-3.289		Source	A baseball player slides into **home plate** and is called safe by the umpire.	–
		Reference	一名棒球运动员滑进本垒，被裁判 判为安全 。	–
		Text-only	一名棒球运动员滑垒至本垒，裁判 判其安全 。	87.958
		Multimodal	一名棒球选手滑垒至本垒，裁判 判定其安全上垒 。	84.669

this scope may not fully capture diverse cross-lingual ambiguities. Second, strictly pre-visual routing prevents the model from identifying uninformative or misaligned images at test time, which poses risks when deploying on uncurated, noisy corpora. Third, the router optimizes for binary visual necessity rather than realized translation gain. While effective at general classification, it struggles to optimally rank high-gain instances under extremely tight visual budgets. Future work will explore multilingual expansion and gain-aware ranking objectives.

References

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mingkang Dong, Hongyi Cai, Jie Li, Sifan Zhou, Bin Ren, Kunyu Peng, and Yuqian Fu. 2026. VisNec: Measuring and Leveraging Visual Necessity for Multimodal Instruction Tuning. *arXiv preprint arXiv:2603.01195*.

Desmond Elliott. 2018. Adversarial Evaluation of Multimodal Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

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.

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 Asso-*

ciation for Computational Linguistics: NAACL 2025, pages 761–778, Albuquerque, New Mexico. Association for Computational Linguistics.

Boyu Guan, Chuang Han, Yining Zhang, Yupu Liang, Zhiyang Zhang, Yang Zhao, and Chengqing Zong. 2025. 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, Chuang Han, Yang Zhao, and Chengqing Zong. 2026. DART: Disambiguation-Aware Reasoning for Video-guided Machine Translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7772, San Diego, California, United States. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Chiraag Lala and Lucia Specia. 2018. Multimodal Lexical Translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision Matters When It Should: Sanity Checking Multimodal Machine Translation Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Danyang Liu, Fanjie Kong, Xiaohang Sun, Dhruva Patil, Avijit Vajpayee, Zhu Liu, Vimal Bhat, and Najmeh Sadoughi. 2025. Detect, Disambiguate, and Translate: On-Demand Visual Reasoning for Multimodal Machine Translation with Large Vision-Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1559–1570, Albuquerque, New Mexico. Association for Computational Linguistics.

Zi Long, ZhenHao Tang, Xianghua Fu, Jian Chen, Shilong Hou, and Jinze Lyu. 2024. Exploring the Necessity of Visual Modality in Multimodal Machine Translation using Authentic Datasets. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 36–50, Torino, Italia. ELRA and ICCL.

Yingfeng Luo, Hongyu Liu, DingYang Lin, Kaiyan Chang, Chenglong Wang, Bei Li, Quan Du, Tong Xiao, and JingBo Zhu. 2026a. RouteLMT: Learned Sample Routing for Hybrid LLM Translation Deployment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1886–1897, San Diego, California, USA. Association for Computational Linguistics.

Yingfeng Luo, Ziqiang Xu, Yuxuan Ouyang, MuRun Yang, DingYang Lin, Kaiyan Chang, Tong Zheng, Bei Li, Peinan Feng, Quan Du, Tong Xiao, and JingBo Zhu. 2026b. NiuTrans.LMT: Toward inclusive and scalable multilingual machine translation with LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25151–25179, San Diego, California, United States. Association for Computational Linguistics.

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

Jingheng Pan, Xintong Wang, Longyue Wang, Liang Ding, Weihua Luo, and Chris Biemann. 2026. VIDA: A Dataset for Visually Dependent Ambiguity in Multimodal Machine Translation. *arXiv preprint arXiv:2605.02035*.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, 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.

A Dataset Construction Details

Preprocessing. We retain source sentences containing 3–40 words and remove exact duplicates and near duplicates with a source-sentence Jaccard similarity of at least 0.7. Deduplication is performed before dataset splitting. To filter mismatched image–text pairs, the paired source sentence must rank within the top 0.5% of candidate texts for its image under SigLIP similarity.

Automatic annotation and adjudication. Qwen-Max and DeepSeek-V4-Pro independently assess textual sufficiency without access to the paired image. Their exact three-way label agreement is 79.04%. GPT-5.5 adjudicates disagreements using the scoring rubric in Table 6. All thresholds are set heuristically and kept fixed throughout data construction.

The higher-scoring annotation is automatically accepted when its total score is at least 11 and its lead is at least 2 points. If the top annotation reaches 11 but its lead is below 2, the instance is referred for human review. Instances whose top annotation scores below 11 are discarded.

Visual evidence verification. For each textually unresolved instance, SigLIP assigns image–description compatibility scores in $[0, 1]$. An instance is automatically accepted as `vision_needed` when the highest score is at least 0.60 and its margin over the runner-up is at least 0.08. Cases meeting the score threshold but not the margin threshold undergo human review, while those below the score threshold are discarded.

Human quality control. Two computer science master's students, both native Chinese speakers proficient in English, independently review the borderline cases and all 3,300 instances entering final screening. For each instance, they inspect the source sentence, paired image, proposed ambiguity, candidate interpretations and translations, and supporting textual or visual evidence. Each reviewer independently records a pass/fail decision before discussion.

The raw pass/fail agreement is 77%. Disagreements are resolved through discussion. Labels are corrected when the available evidence supports a clear revision; otherwise, the instance is removed. Final screening removes 247 instances (7.48%) and retains 3,053.

Table 6: Scoring rubric for GPT-based adjudication.

Scoring dimension	Range
Label validity	1–5
Ambiguity interpretation	1–5
Textual evidence sufficiency	1–3
Reasoning consistency	1–3
Total	4–16

B Cross-Backend Evaluation

We further evaluate PreVIR with InternVL3.5-8B, LLaVA-1.5-7B, and LLaVA-OneVision-1.5-8B on the same 344-instance test set. The trained XLM-R router is reused without backend-specific retraining, yielding identical routing decisions and a visual-call rate of 46.22%. Table 7 reports overall and label-wise translation results.

Across these additional backends, PreVIR improves COMET over Text-only by 0.43–1.96 points. Compared with Always-Multimodal, COMET decreases by 0.43 points on InternVL3.5 and 0.39 points on LLaVA-OneVision, while increasing by 0.37 points on LLaVA-1.5. On the vision_needed subset, COMET gains over Text-only range from 1.49 to 3.88 points. These results support the reuse of the learned routing policy across the evaluated backends, although the resulting translation gains remain backend-dependent.

C Statistical Reliability

Table 8 reports paired score differences, 95% confidence intervals, and two-sided p-values with Holm correction across the four comparisons. PreVIR significantly improves COMET over Text-only by 1.72 points (95% CI: [1.01, 2.48]; adjusted $p < 0.001$), whereas the BLEU improvement is not statistically significant. Compared with Always-Multimodal, neither metric shows a statistically significant difference, while PreVIR uses vision for only 46.22% of inputs. However, these results do not establish equivalence or non-inferiority. The intervals quantify test-sample uncertainty for the evaluated system outputs.

D Practical Efficiency

Experimental Setup. We benchmark PreVIR and Always-Multimodal on the 344-instance test set, using the PreVIR routing decisions from Table 4 (46.22% visual calls). Both methods use Qwen3.5-9B in BF16 with vLLM 0.18.0 on a single

Table 7: Overall and label-wise translation results with additional backends. Column definitions follow Table 4. The same trained XLM-R router is reused across all backends without retraining. PreVIR uses a visual-call rate of 46.22%, while Gain-Oracle is a diagnostic upper bound evaluated at its backend-specific native call rate.

	BLEU↑				COMET↑				
Strategy	Overall	No amb.	Text res.	Vision req.	Overall	No amb.	Text res.	Vision req.	Vision Call
InternVL3.5-8B									
Gain-Oracle	48.80	54.66	48.59	41.72	88.31	90.00	88.73	87.02	45.06%
Gold-label Router	46.78	–	–	–	87.09	–	–	–	47.09%
Text-only	44.92	51.16	47.65	34.28	84.96	88.27	87.19	81.75	0.00%
Always-Multimodal	45.96	50.37	45.62	40.54	87.34	88.92	87.42	86.29	100.00%
PreVIR (Ours)	46.41	50.99	48.19	39.01	86.91	88.59	87.33	85.63	46.22%
LLaVA-1.5-7B									
Gain-Oracle	29.60	32.00	30.00	25.06	81.80	83.19	80.70	81.44	43.60%
Gold-label Router	27.36	–	–	–	79.24	–	–	–	47.09%
Text-only	27.20	29.31	28.81	22.54	78.32	80.88	77.87	76.89	0.00%
Always-Multimodal	26.96	28.16	27.98	22.89	78.38	77.95	77.98	78.85	100.00%
PreVIR (Ours)	27.14	29.01	28.39	22.91	78.75	80.21	77.56	78.38	46.22%
LLaVA-OneVision-1.5-8B									
Gain-Oracle	47.15	51.32	49.13	39.88	88.48	89.55	88.82	87.62	38.95%
Gold-label Router	45.21	–	–	–	87.25	–	–	–	47.09%
Text-only	43.79	49.13	46.56	33.75	85.43	88.18	87.28	82.78	0.00%
Always-Multimodal	45.21	48.72	47.45	38.47	87.57	88.81	87.83	86.65	100.00%
PreVIR (Ours)	44.87	49.26	46.48	37.26	87.18	88.59	87.56	86.09	46.22%

Table 8: Paired comparisons on the 344-instance test set. Δ denotes PreVIR minus the baseline. We report pointwise 95% confidence intervals for the differences and two-sided p-values adjusted using Holm's procedure across all four comparisons.

Metric	Δ	95% CI	p_{Holm}
PreVIR vs. Text-only			
BLEU	+0.97	[−0.13, 2.09]	0.2916
COMET	+1.72	[1.01, 2.48]	**0.0004**
PreVIR vs. Always-MM			
BLEU	−0.40	[−2.05, 1.25]	0.6333
COMET	−0.32	[−0.75, 0.08]	0.2990

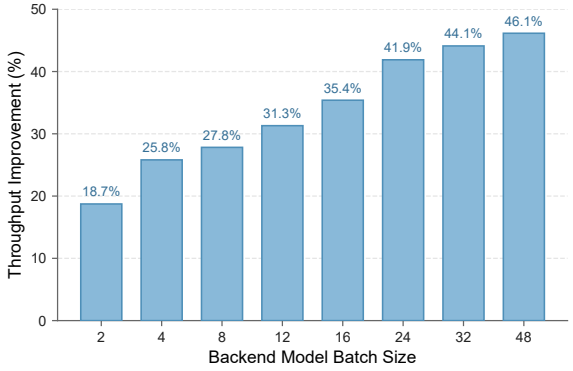


Figure 7: Throughput improvement of PreVIR over Always-Multimodal at matched request batch sizes, computed from mean total runtimes including router overhead. Runtime variability is reported in Table 9.

NVIDIA A800 80GB GPU. The router batch size is fixed at 32. PreVIR first routes all inputs, then translates the text-only and multimodal subsets sequentially. We vary the translation request batch size from 2 to 48, using the same value for both methods. Images are resized offline to at most 307,200 pixels, and decoding is greedy with a maximum of 128 output tokens. Timing includes routing (for PreVIR), online image processing, and translation generation. We use throughput to quantify the efficiency gains. The relative improvement over Always-Multimodal is $(\overline{T}_{\text{Always-MM}}/\overline{T}_{\text{PreVIR}} - 1) \times 100\%$.

Results. Figure 7 shows throughput improvements of 18.7–46.1% across the evaluated request batch sizes. Table 9 reports the corresponding runtimes and their variability. Routing takes only 0.238–0.258 seconds per test-set pass, accounting for 0.6–2.6% of PreVIR's total runtime. At a request batch size of 48, PreVIR reduces total runtime from 13.592 to 9.301 seconds, a 31.6% reduction. Although the additional router incurs its own cost, the overall system still delivers practical efficiency gains that far exceed this cost.

Table 9: Processing time for the complete 344-instance test set, reported as mean $\pm$ standard deviation over five repeated runs. Always-MM and PreVIR are reported in seconds, whereas Router is reported in milliseconds. PreVIR includes both routing and translation. Batch denotes the translation request batch size.

Batch	Always-MM (s)	PreVIR (s)	Router (ms)
2	48.45 ± 0.16	40.81 ± 0.17	255 ± 5
4	32.88 ± 0.06	26.14 ± 0.02	247 ± 7
8	23.15 ± 0.18	18.14 ± 0.15	253 ± 9
12	19.76 ± 0.04	15.05 ± 0.05	252 ± 6
16	17.99 ± 0.05	13.30 ± 0.07	238 ± 10
24	15.815 ± 0.132	11.147 ± 0.081	241 ± 16
32	14.731 ± 0.034	10.223 ± 0.087	258 ± 9
48	13.592 ± 0.068	9.301 ± 0.042	245 ± 17