Letting Uncertainty Guide Your Multimodal Machine Translation

Wuyi Liu[*1] **Yue Gao**[*1] **Yige Mao**[2] **Jing Zhao**[†1]

[1]Computer Science Dept., East China Normal University, Shanghai, China
[2]Beihang University, Beijing, China,

Abstract

Multimodal Machine Translation (MMT) leverages additional modalities, such as visual data, to enhance translation accuracy and resolve linguistic ambiguities inherent in text-only approaches. Recent advancements predominantly focus on integrating image information via attention mechanisms or feature fusion techniques. However, current approaches lack explicit mechanisms to quantify and manage the uncertainty during translation process, resulting in the utilization of image information being a black box. This makes it difficult to effectively address the issues of incomplete utilization of visual information and even potential degradation of translation quality when using visual information.To address these challenges, we introduce a novel Uncertainty-Guided Multimodal Machine Translation (UG-MMT) framework that redefines how translation systems handle ambiguity through systematic uncertainty reduction. Designed with plug-and-play flexibility, our framework enables seamless integration into existing MMT systems, requiring minimal modification while delivering significant performance gains.

1 INTRODUCTION

In traditional machine translation models, encountering sentences such as "a man is walking on the bank" can often lead to easily deducing that "bank" means "riverbank" due to the presence of "on." However, text alone does not always provide sufficient information to resolve the ambiguity of certain words. This is where the concept of multimodal translation emerges. Yao and Wan [2020] defined multimodal machine translation (MMT) as a novel machine translation task that aims to design better translation systems using context from the additional image modality.

In recent years, the realm of MMT has made significant strides to enhance translation accuracy by incorporating visual data alongside textual inputs. Various models have been developed to harness these additional modalities, each offering unique approaches to improving translation performance. For example, the *Multimodal Transformer* [Yao and Wan, 2020] employs cross-modal attention to dynamically align relevant image regions with the corresponding parts of the text being translated, thereby enhancing contextual interpretation. Innovative techniques like *Inversion Knowledge Distillation* [Peng et al., 2023] improves MMT outputs by distilling image information, thus minimizing the need for direct visual input during inference. Additionally, Valhalla [Li et al., 2022] leverages visual hallucination strategies, generating more robust translations by simulating visual contexts even in the absence of actual visual data.

Although methodological advancements have improved translation performance, empirical analyses reveal certain limitations in current approaches. Our evaluation shows that instances of elevated uncertainty account for approximately 45% of cases in the Gated Fusion model [Wu et al., 2021] and 37% of cases in the Revisit MMT model [Wu et al., 2021]. This phenomenon is deeply concerning, as it conflicts directly with the foundational principle of multimodal machine translation (MMT). The core objective of MMT is to leverage visual modality to reduce ambiguity in text , assisting in resolving linguistic uncertainties that purely textual models struggle to address.

The key issue is that existing models lack a clear metric for expressing and measuring uncertainty, making it impossible to quantify whether visual information actually helps in reducing ambiguity. This stands in contrast to other domains, such as multi-class classification tasks [Sensoy et al., 2018], where uncertainty modeling is well-established. Nevertheless, recent works such as MAP [Ji et al., 2023] and UNO [Tian et al., 2020] have demonstrated that uncertainty model-

[*] Equal contribution.
[†] Corresponding author: jzhao@cs.ecnu.edu.cn

ing can also be effectively leveraged in multimodal learning tasks. However, the application of uncertainty modeling to machine translation remains largely unexplored. This is primarily due to the intrinsic complexities of the translation process. First, machine translation involves generating sequences across an effectively infinite output space, where conventional uncertainty modeling metrics such as probability distributions over finite class sets become inapplicable. Second, ambiguities can propagate across sequence tokens, creating cascading uncertainties that traditional approaches fail to address. Third, the cross-modal interactions in MMT involve richer contextual dependencies, often introducing unpredictable noise rather than resolving ambiguities. For instance, visual context may mislead the model when images contain irrelevant or conflicting information.

To address these gaps, we propose the Uncertainty-Guided Multimodal Machine Translation (UG-MMT) framework. By explicitly modeling uncertainty at the token and sequence levels, UG-MMT not only quantifies the contribution of visual information but also guides the cross-modal fusion process toward consistent ambiguity reduction, addressing the fundamental challenges outlined above.

Our approach introduces three key innovations:

- A novel uncertainty modeling framework specifically designed for sequence generation, which captures ambiguity at both token and sentence levels while handling the infinite output space of translation.
- An uncertainty-guided cross-modal fusion mechanism that explicitly optimizes for uncertainty reduction, ensuring visual information serves its intended purpose of disambiguation.
- Comprehensive validation demonstrating that UG-MMT significantly outperforms existing approaches across multiple datasets and metrics, achieving state-of-the-art performance on several standard benchmarks.

Through extensive experiments on established MMT frameworks like Gated Fusion and Revisit-MMT [Wu et al., 2021], we demonstrate that our uncertainty-guided approach consistently improves performance across all evaluation metrics, validating the effectiveness of explicit uncertainty modeling in multimodal translation.

2 RELATED WORK

2.1 MULTIMODAL MACHINE TRANSLATION

Multimodal Machine Translation extends traditional neural machine translation by introducing additional modalities, such as images, to reduce linguistic ambiguity and improve translation robustness. Recent studies have explored various strategies for integrating visual information into translation pipelines, which can be broadly categorized into three main approaches.

The first category involves feature concatenation methods, where visual and textual features are simply combined during encoding. For example, Yao and Wan [2020] utilized a multi-modal Transformer framework, concatenating the visual features from pre-trained image embeddings with textual inputs. Similarly, Takushima et al. [2019] incorporated global image embeddings to construct multimodal feature representations for improving translation performance.

The second key direction leverages cross-modal interactive attention mechanisms. Nishihara et al. [2020] enhanced translation by allowing the model to attend to both token-level textual information and region-level visual features. Zhao et al. [2022] further extended this line of research by proposing a cross-modal interaction module, integrating visual and textual features through region-level and word-level attention mechanisms.

The third prominent approach is the gated fusion mechanism, which dynamically controls the contributions of different modalities based on their contextual importance. For instance, Wu et al. [2021] proposed a multimodal fusion method using independently encoded text and image representations, integrating them through a gating mechanism. Building upon this, Lin et al. [2020] designed a model that incorporated dynamic context-guided capsule networks for robust visual feature extraction, followed by a gating mechanism to align and fuse modalities.

Despite these developments, prior work has largely focused on architectural improvements to enhance multimodal embeddings without explicitly addressing the uncertainty inherent in cross-modal alignment. As noted in Table 1, visual information can sometimes increase, rather than decrease, translation uncertainty, emphasizing the need for uncertainty-aware multimodal translation frameworks.

2.2 MULTIMODAL UNCERTAINTY LEARNING

Uncertainty modeling has become an essential tool for quantifying model confidence and managing ambiguity, particularly in high-stakes AI applications. Traditional uncertainty estimation techniques are well-established in tasks like multi-class classification [Sensoy et al., 2018], where methods such as Dirichlet-based evidential learning enable models to represent and quantify classification uncertainty effectively.

Recently, the field has expanded to multimodal uncertainty learning, focusing on integrating uncertainty estimation into multimodal tasks. For instance, Jung et al. [2024] proposed a Bayesian framework for generalizing uncertainty estimation to multimodal settings, achieving state-of-the-art results in uncertainty-aware learning. Ji et al. [2023] introduced

MAP, a multimodal uncertainty-aware vision-language pre-training model, modeling sequence-level interactions between visual and textual data to align probabilistic representations. Gao et al. [2024] further emphasized the importance of aleatoric uncertainty in multimodal fusion, demonstrating its impact on improving prediction robustness across different modalities.

Beyond vision-language models, specific multimodal applications such as emotion recognition [Chen et al., 2022] and intention detection [Trick et al., 2019] have introduced task-specific uncertainty modeling frameworks. Chen et al. [2022] designed a hierarchical uncertainty module that captures both context-level and modality-level uncertainties, enabling more accurate predictions in conversational scenarios. Trick et al. [2019] proposed an uncertainty-reduction pipeline for intention recognition, demonstrating that explicit cross-modal uncertainty management significantly improves system robustness.

Furthermore, Ott et al. [2018] and Wang et al. [2020] primarily focus on uncertainty in neural machine translation, addressing uncertainty calibration from the perspectives of data distribution and the inference stage, respectively. Ott et al. [2018] leverages uncertainty estimation tools to measure distributional discrepancies in the data and tackles the issue in NMT models where low-frequency words are assigned low probabilities in the predictive distribution, resulting in a lack of diversity in the translation outputs. Wang et al. [2020] proposes a stepwise label smoothing method to quantify confidence calibration bias in NMT during the inference stage.

Despite these advancements, the application of multimodal uncertainty learning to complex generation tasks, such as machine translation, remains underexplored. Unlike classification tasks with finite output classes, translation involves generating sequences over an effectively infinite output space, where ambiguities propagate across tokens. This fundamental challenge necessitates new strategies for token-level uncertainty modeling and sequence-level uncertainty fusion, as proposed in this paper.

3 METHODS

3.1 UNCERTAINTY LEARNING FOR TRANSLATION MODELS

To achieve our goal of reducing uncertainty through new modalities in MMT, it is crucial to first enable the model to quantify uncertainty. In multi-class classification tasks, modeling uncertainty using Dirichlet distributions has been demonstrated to be effective and well-established [Sensoy et al., 2018]. However, machine translation presents unique challenges for uncertainty modeling. Unlike classification, which operates on fixed and bounded label spaces, translation involves generating sequences from immense, almost infinite vocabularies. Each token-level prediction depends not only on the source input but also on all prior tokens in the output sequence, creating cascading dependencies. This sequential nature amplifies uncertainty, as small ambiguities in earlier tokens can propagate and impact subsequent predictions. Furthermore, the integration of visual modalities introduces richer yet noisier feature spaces, increasing the complexity of precise uncertainty estimation. To address uncertainty in this complex setting, we build on the principles of evidential learning but adapt them for sequence generation tasks. The natural starting point is the transformation of logits—predicted by each time step during translation—into probabilistic models that quantify evidence for class predictions.

In transformer-based translation models, the final outputs are typically a series of logits, which are the unnormalized scores for each word in the vocabulary. These logits represent the raw predictions of the model before any normalization or activation is applied. Upon passing these logits through a softmax layer, they are converted into probabilities, indicating the likelihood of each word being the correct one at a given position.

Essentially, this transformation turns the task into a multi-class classification problem, where each position in the sequence corresponds to a different class. Given this setup, the logits, denoted as z, can be considered as the model's predictions before normalization. We utilize these logits by transforming them into evidence values using a ReLU activation function. Specifically, logit z_w for each token is processed with the ReLU function to obtain the evidence $e_w = \mathrm{ReLU}(z_w)$. Following this transformation, the evidence e_w is augmented by adding one, resulting in the Dirichlet parameters $\alpha_w = e_w + 1$. This augmentation step is necessary to satisfy the properties of the Dirichlet distribution, where the parameters α_w must each be greater than zero. These Dirichlet parameters are then utilized in further computations, allowing us to model the uncertainty and variability inherent in the translation task.

Given these parameters, the uncertainty u and belief masses b_w for each word w in the vocabulary are formulated as follows:

$$b_w = \frac{\alpha_w - 1}{S} \quad \text{and} \quad u = \frac{V}{S}, \tag{1}$$

where $S = \sum_{w=1}^{V} \alpha_w$ represents the Dirichlet strength and V denotes the size of the target vocabulary. The uncertainty is inversely related to the total evidence, encapsulating the "I do not know" stance when evidence is low.

While these formulations offer a sound theoretical basis for capturing uncertainty, their direct application to machine translation tasks proved challenging due to the nature of sequence generation. A naive implementation of uncertainty modeling would employ the Kullback-Leibler (KL) diver-

gence term in the loss function to align the predicted Dirichlet distribution with the target probabilities. However, this method often over-penalizes high uncertainty predictions and discourages exploration in ambiguous contexts. This issue becomes particularly noticeable in translation tasks, where the immense vocabulary size amplifies the impact of such penalties. Tokens associated with synonyms, polysemy, or cultural nuances inherently exhibit higher uncertainty, and over-penalizing these cases can hinder the model's ability to flexibly adapt to the diversity and complexity of language.

Also, relying solely on an uncertainty loss based on Dirichlet distribution, without the standard cross-entropy loss, can lead to significant performance degradation during training. Cross-entropy explicitly measures the probability assigned by the model to the correct ground-truth token at each time step. This ensures token-level precision by driving the logits distribution $p(y_i|\text{context})$ toward the correct output y_i. In translation tasks, where maintaining token-by-token alignment is critical, such direct supervision is indispensable. Uncertainty losses like $\mathcal{L}_{\text{err}}$ 2, on the other hand, focus on minimizing the error between predicted distributions and confidence scores, while $\mathcal{L}_{\text{var}}$ 3 regularizes the Dirichlet variance to prevent overconfident predictions. While these components improve calibration of uncertainty, they do not explicitly enforce the correct token being predicted, thus failing to address token-level precision challenges in sequence generation tasks.

To address these challenges, we adopted a hybrid approach. Instead of treating uncertainty as the primary optimization goal, it was incorporated as an auxiliary regularization term into the standard cross-entropy loss. This adjustment preserved the strengths of cross-entropy for token-level accuracy while leveraging uncertainty regularization to calibrate the predictions for ambiguous tokens. Specifically, our final loss function comprises:

- $\mathcal{L}_{\text{err}}$: This measures the squared error between the true token distribution and the model's predicted token probabilities across all tokens in a sentence:

$$\mathcal{L}_{\text{err}} = \sum_{i=1}^{N} \sum_{w \in V} (y_{iw} - \hat{p}_{iw})^2, \tag{2}$$

  where N represents the number of tokens in the given sentence, y_{iw} is the one-hot encoded true distribution for the i-th token w, and $\hat{p}_{iw}$ is the predicted probability for token w.

- $\mathcal{L}_{\text{var}}$: This captures the uncertainty in predictions by incorporating the variance from the Dirichlet distribution across all tokens in the sentence:

$$\mathcal{L}_{\text{var}} = \sum_{i=1}^{N} \sum_{w \in V} \frac{\hat{p}_{iw}(1 - \hat{p}_{iw})}{S_i + 1}, \tag{3}$$

  where $S_i = \sum_{w \in V} \alpha_{iw}$ represents the evidence (Dirichlet strength) for the i-th token in the sentence.

The application of Equation 3 is pivotal in making the model mathematically more confident in leveraging image data effectively to manage uncertainty in multimodal contexts.However, simply adding this term could easily lead to overconfidence on prediction, thus limiting the performance improvement. As a resolution, we adopted label-smoothed cross-entropy, also used by our baseline models, which prevents the model from becoming overly confident by distributing small probabilities to incorrect options. This allowed us to manage uncertainty effectively without imposing rigid constraints on prediction distributions, thus maintaining the quality of the translations.

The total loss function is thus a combination of these two components:

$$L(\Theta) = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{err}} + \lambda_2 \mathcal{L}_{\text{var}}. \tag{4}$$

This formulation enables the model to simultaneously minimize prediction errors and account for uncertainty, thereby calibrating the confidence in its predictions in a more comprehensive manner. By effectively utilizing the multimodal translation logits and the optimized loss function, our approach significantly improves the robustness of classification decisions under uncertain conditions.

3.2 UNCERTAINTY-GUIDED MULTIMODAL MACHINE TRANSLATION

The overall architecture of our proposed framework is illustrated in Figure 1. This framework integrates both textual and visual modalities through a Gated Fusion mechanism. Text sequences are processed via word and positional embeddings, while images are transformed into visual embeddings. These features are dynamically fused and passed through a Transformer decoder for sequence-level language generation.

Incorporating additional modalities like images into translation tasks is intended to help disambiguate and reduce uncertainty, thereby improving translation accuracy. However, after successfully integrating uncertainty modeling into the multimodal translation task, we observed that the inclusion of images did not consistently result in reduced uncertainty across various scenarios. This observation indicated that the model was not effectively leveraging images to resolve textual ambiguities, contradicting the fundamental goal of multimodal translation. From the data presented in Table 1, we can infer that the current translation models show some potential for using the visual modality to reduce uncertainty. This suggests that while the models have the capacity to improve translation accuracy through multimodal integration, the strategies for leveraging visual information are not yet fully optimized.

To address this challenge, we sought a metric to assess the images' impact on reducing uncertainty. We decided on the

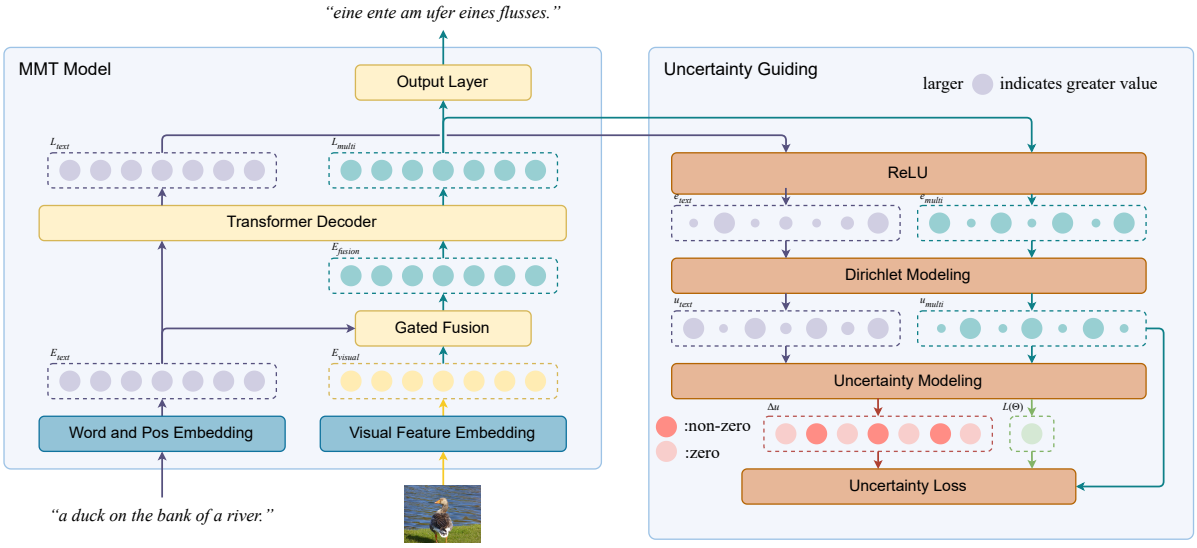


Figure 1: Architecture of the proposed Uncertainty-Guided Multimodal Machine Translation (UG-MMT) framework. The left side of the figure illustrates the multimodal translation pipeline, incorporating both textual and visual features. Text sequences are processed via word and positional embeddings, while images are transformed into visual embeddings. These features are fused using a Gated Fusion mechanism before being passed through the Transformer decoder for sequence generation. The right panel highlights the uncertainty modeling process. Text-only and multimodal logits are transformed into evidence values via the ReLU activation function. A higher evidence value indicates stronger confidence, resulting in lower uncertainty. The figure also shows the computation of the relative uncertainty difference (Δu), where the color depth reflects the magnitude of Δu. Specifically, when multimodal uncertainty (u_{multi}) exceeds text-only uncertainty (u_{text}), $\Delta u > 0$, shown as deeper-colored nodes. In contrast, when $u_{\text{multi}} \leq u_{\text{text}}$, $\Delta u = 0$ due to the ReLU activation, effectively ignoring such cases. The uncertainty loss incorporates both absolute multimodal uncertainty (u_{multi}) and relative uncertainty difference (Δu), guiding the model to leverage visual features effectively for ambiguity reduction.

uncertainty difference between translations with and without images. Our objective was that the inclusion of images should consistently lead to lower uncertainty. Initially, a straightforward approach was to incorporate this difference as a regularization term in the loss function. Yet, merely maximizing this difference risked the model overemphasizing the role of images. Therefore, to prevent such an imbalance, we applied the ReLU function to this difference, ensuring the regularization effect only activates when the multimodal uncertainty surpasses the text-only uncertainty:

$$\Delta u = \text{ReLU}\left(\frac{u_{\text{multi}}}{u_{\text{text}} + \epsilon} - 1.0\right) \tag{5}$$

where u_{multi} represents the uncertainty in multimodal translation, u_{text} represents the uncertainty in text-only translation, and ϵ is a small constant added to avoid division by zero. The ratio reflects the relative change in uncertainty between multimodal and text-only settings, ensuring that only when multimodal uncertainty surpasses text-only uncertainty does the regularization term activate. Compared to directly subtracting these uncertainties ($\Delta u = u_{\text{multi}} - u_{\text{text}}$), this ratio-based approach provides smoother and more balanced adjustments. By normalizing the uncertainties, it ensures that their relative contributions are independent of their magnitude scales, mitigating sensitivity to large or small absolute values. Additionally, it avoids abrupt gradient contributions common with simple subtraction, enhancing training stability and preventing the model from over-relying on images. Finally, the use of ReLU further restricts optimization to cases where multimodal uncertainty truly exceeds text-only uncertainty, ensuring the regularization targets meaningful scenarios aligned with reducing overall ambiguity.

Another critical component of the loss function is u_{multi}, which explicitly penalizes high multimodal uncertainty during training. This term is crucial for ensuring that the additional modalities, particularly the visual inputs, actively contribute to reducing ambiguity within the translation process. Without directly enforcing an uncertainty penalty, the model might ignore the uncertainty from multimodal inputs or fail to optimize it effectively.

While Δu encourages relative uncertainty reduction to optimize the visual modality's contribution, u_{multi} focuses on the absolute multimodal uncertainty, playing a complementary role in the loss function. Incorporating u_{multi} ensures that the multimodal system minimizes overall uncertainty in every

scenario, independent of the relative differences between modalities. This term serves several important purposes.

First, minimizing u_{multi} directly penalizes high levels of multimodal uncertainty, driving the model toward producing sharper and more confident probability distributions. These sharper distributions improve token-level precision during sequence generation, aligning with the goal of increasing prediction accuracy. By enforcing this absolute certainty, the model learns to construct more robust feature representations from both the textual and visual inputs. Second, u_{multi} prevents potential exploitation of the Δu term. When only a relative uncertainty difference is regularized, the model might retain an overall high uncertainty in multimodal predictions while artificially lowering Δu. This could undermine the true goal of reducing ambiguities. The inclusion of u_{multi} ensures that uncertainty optimization is not just relative but also absolute, pushing the system toward reliably low uncertainty in multimodal contexts.

Overall, the inclusion of u_{multi} complements Δu by addressing both the absolute uncertainty minimization and the relative uncertainty difference, ensuring a more balanced and principled approach to optimizing multimodal predictions.

To integrate this into our training process, we define the loss function as follows:

$$\mathcal{L} = u_{\text{multi}} + \beta \cdot \Delta u + \lambda * L(\Theta) \tag{6}$$

where β is a scaling factor dependent on the training epoch, and $L(\Theta)$ represents the regularization term defined in the previous Section 4.

Algorithm 1 Uncertainty-Guided Multimodal Machine Translation

Require: Text logits z_{text}, Multimodal logits z_{multi}
Ensure: Effectively use the new modality to reduce uncertainty

$e_{\text{text}} \leftarrow \text{ReLU}(z_{\text{text}}) + 1$
$e_{\text{multi}} \leftarrow \text{ReLU}(z_{\text{multi}}) + 1$
$u_{\text{text}} \leftarrow \frac{V}{\sum e_{\text{text}}}$
$u_{\text{multi}} \leftarrow \frac{V}{\sum e_{\text{multi}}}$
$\Delta u \leftarrow \text{ReLU}\left(\frac{u_{\text{multi}}}{u_{\text{text}}+\epsilon} - 1.0\right)$
$\mathcal{L} \leftarrow u_{\text{multi}} + \beta \cdot \Delta u + L(\Theta)$

4 EXPERIMENTS

4.1 DATASET

In this section, we evaluate our framework with the widely used Multi30K benchmark [Elliott et al., 2016]. The training and validation sets consisted of 29, 000 and 1,014 instances. We evaluate on TEST2016, TEST2017 (1,000 instances), and MSCOCO [Elliott et al., 2017] (461 challenging out-of-domain samples). As the process in the project [Wu et al., 2021],We merge the source and target sentences in the officially preprocessed version of Multi30k to build a joint vocabulary. We then apply the byte pair encoding (BPE) algorithm [Sennrich, 2015] with 10,000 merging operations to segment words into subwords, which generates a vocabulary of 9,712(9,544) tokens for En-De (En-Fr).

4.2 SETUP

Our experimental setup closely follow the methodologies described in the papers of Gated Fusion and Revist MMT, ensuring consistent variable control to effectively highlight the impact of our introduced component. For optimization, we used the Adam optimizer with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate initially increased linearly from 10^{-7} to 0.005 during the warm-up phase and then decayed in proportion to the number of updates.

Each training batch was composed of up to 16,384 source/target tokens. We applied label smoothing with a weight of 0.1 and a dropout rate of 0.3 to prevent overfitting. Training was scheduled to halt early if the validation loss did not improve over 10 consecutive epochs [Zhang et al., 2020]. During inference, we averaged the results of the last 10 checkpoints and performed beam search with a beam size of 5 to select the best translation candidates. Evaluation metrics included 4-gram BLEU and METEOR scores across all test sets. All models are trained and evaluated on one single machine with one RTX 4090 GPU (5-10 minutes for the entire training process).

Table 1: Proportion of Instances Where Visual Modality Increased Uncertainty

Models	Before UG-MMT Integration	After UG-MMT Integration
Gated Fusion	45.0% of cases	0.0% of cases
Revisit MMT	37.0% of cases	0.0% of cases

4.3 RELULTS

To position our work within the broader context of multimodal translation research, we compared our approach with current state-of-the-art MMT models. Table 2 shows the comparison results on the Multi30k dataset. Notably, our UG-MMT enhanced models achieved SOTA performance on the Test2016 dataset, with a BLEU score of 42.82 for En→De translation. This result not only validates the effectiveness of our uncertainty-guided approach but also demonstrates its potential to advance the field of multimodal translation.

In our preliminary analysis of existing multimodal translation models, we observed a concerning phenomenon where

Table 2: Comparison with existing MMT systems on Multi30K dataset (B: BLEU, M: METEOR)

System	En→De						En→Fr					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	B	M	B	M	B	M	B	M	B	M	B	M
Existing Traditional MMT Systems												
Multimodal Self-attn [Yao and Wan, 2020]	41.02	-	33.36	-	29.88	-	61.8	-	53.46	-	44.52	-
Gated Fusion$^{\diamond}$ [Wu et al., 2021]	41.56	68.17	32.74	60.99	29.04	56.00	61.05	80.1	54.09	75.47	44.25	69.12
Revisit MMT$^{\diamond}$ [Wu et al., 2021]	40.8	68.01	32.94	61.33	28.83	56.02	62.05	81.12	53.79	76.28	44.87	69.33
IKD-MMT [Peng et al., 2023]	41.28	58.93	**33.83**	53.21	**30.17**	48.93	-	-	-	-	-	-
MGNMT(TF PCL-O) [Yin et al., 2023]	40.4	58.4	32.5	52.0	29.0	48.5	61.3	75.8	54.4	70.7	-	-
RG-MMT-EDC [Tayir et al., 2024]	42.00	60.20	33.40	53.70	30.00	49.60	**62.90**	77.20	**55.80**	72.00	45.10	64.90
UG-MMT+Gated Fusion (Ours)	**42.82**	**69.11**	33.78	**61.49**	28.93	56.03	62.01	81.41	54.43	76.47	**45.31**	**69.93**
UG-MMT+RMMT (Ours)	42.01	68.59	33.2	61.44	30.01	**56.6**	62.28	**81.56**	54.47	**76.67**	45.16	69.76

Note: $^{\diamond}$ means to reproduce previous MMT methods based on the settings mentioned on experiment section. Best results are shown in **bold**, second best results are underlined. '-' indicates unavailable results.

visual information frequently led to increased uncertainty in translation decisions. After integrating our UG-MMT framework, we observe a dramatic shift in this pattern, which is shown in Table 1. This transformation in uncertainty management precedes and directly contributes to improved translation performance, suggesting that uncertainty reduction serves as a driving force for enhanced translation quality rather than merely being a byproduct of better translations.This causal relationship between uncertainty reduction and performance improvement is further validated by our experimental results presented in Table 3. The integration of UG-MMT yields substantial improvements across multiple evaluation metrics. Particularly, for the En→De translation task, we observe improvements of up to 1.26 BLEU points on Test2016 with Gated Fusion, while RMMT shows similar positive trends with a 1.21 BLEU point increase. These improvements are notably consistent across different test sets and language pairs, demonstrating the robustness of our uncertainty-guided approach.

To further validate whether uncertainty-guided translation truly leads to more accurate and contextually appropriate translations, we conducted a detailed qualitative analysis. The examples in Table 4 provide concrete evidence of improved translation quality. In the first example, UG-MMT demonstrates superior verb disambiguation, correctly translating "scanning" where the baseline model incorrectly used "winning". The second example showcases improved handling of complex scene understanding, with more precise role identification and better context integration.

5 ANALYSIS

5.1 ABLATION STUDY

To systematically evaluate the contribution of each component in the proposed UG-MMT framework, we performed ablation experiments using Gated Fusion as the baseline model. By introducing different components of UG-MMT (u_{multi}, Δu, and $L(\Theta)$) individually and in combination, we analyzed their effects on translation performance, measured by BLEU scores on the Test2016 dataset. The results of our experiments are summarized in Table 5.

Introducing u_{multi} alone led to a BLEU score improvement from 41.57 to 42.07 ($+0.50$). This improvement can be attributed to u_{multi} encouraging the model to minimize token-level uncertainty during sequence generation. By decreasing the overall uncertainty, the model is incentivized to prioritize predictions with stronger evidence e_k. This inherent preference for confident predictions forces the model to output sharper probability distributions, favoring correct predictions while suppressing incorrect ones.The prioritization of low-uncertainty predictions amplifies the training signal during errors: when the model makes an incorrect prediction, the sharpness of the prediction results in a higher cross-entropy loss compared to normal settings. This reinforcement effect leads to better gradient signals, encouraging the model to improve its generation consistency over time. As a result, u_{multi} directly contributes to improving model convergence and robustness by aligning predictions with token-level confidence and evidence.

On the other hand, incorporating $L(\Theta)$ alone improved the BLEU score to 41.79 ($+0.22$). The relatively modest improvement indicates that $L(\Theta)$ acts primarily as a stabilizing regularization term rather than directly optimizing for accuracy. By leveraging Dirichlet-based evidential learning Sensoy et al. [2018], $L(\Theta)$ helps to balance uncertainty distributions across predictions, particularly in challenging translation tasks. This regularization ensures that the uncertainty model remains well-calibrated, reducing overfitting while preparing the system to utilize uncertainty effectively in conjunction with other components.

When Δu was introduced as a standalone component, the

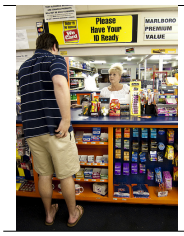
Table 3: Effect of integrating UG-MMT into Gated Fusion and RMMT models on BLEU scores for En→De and En→Fr tasks.

#	Model	En→De			En→Fr		
		Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO
	Baseline Models						
1	Gated Fusion	41.56	32.74	29.04	61.05	54.09	44.25
2	RMMT	40.8	32.94	28.83	62.05	53.79	44.52
	Baseline Models With UG-MMT						
3	Gated + UG	42.82 ↑1.26	33.78 ↑1.04	28.93 ↓0.11	62.01 ↑0.96	54.43 ↑0.34	45.31 ↑1.06
4	RMMT + UG	42.01 ↑1.21	33.2↑0.26	30.01 ↑1.18	62.33 ↑0.28	54.47 ↑0.68	45.16 ↑0.64

Note: All baseline models were re-implemented and evaluated in our experimental environment using identical hyperparameters as specified in Section 4.1. Green arrows (↑) indicate improvements over our re-implemented baselines, while red arrows (↓) indicate decreased performance.

Table 4: Example of translation improvement using UG-MMT

SRC: the gentleman is scanning the image that the woman in the blue shirt is providing him.
MMT: der herr ~~gewinnt~~ das bild der frau im blauen hemd, ~~die ihn anhält~~.
(The gentleman wins the image of the woman in the blue shirt who stops him.)
UG-MMT: der herr **scannt** das bild von der frau im blauen hemd.
(The gentleman scans the image from the woman in the blue shirt.)
REF: der herr scannt das bild, das ihm die frau im blauen hemd zeigt.
(The gentleman scans the image that the woman in the blue shirt shows him.)



SRC: a clerk in a convenience store asks a customer buying alcohol for his age and identification.
MMT: ein ~~kunde~~ in einem ~~nachbarschaftsladen schreibt~~ einen kunden ~~für seine alkohol und identisch~~.
(A customer in a neighborhood store writes a customer for his alcohol and identical.)
UG-MMT: ein **verkäufer** in einem **laden fragt einen kunden nach seinem ausweis** beim alkoholkauf.
(A clerk in a store asks a customer for his identification when buying alcohol.)
REF: ein mitarbeiter in einem laden fragt einen kunden, der alkohol kauft, nach seinem alter und einem ausweis.
(An employee in a store asks a customer who is buying alcohol for his age and identification.)

Table 5: Ablation Study Results on Test2016

u_{multi}	$L(\Theta)$	Δu	BLEU	Δ
			41.57	-
✓			42.07	+0.50
	✓		41.79	+0.22
		✓	40.80	-0.77
✓	✓	✓	**42.82**	+1.25

BLEU score decreased to 40.80 (-0.77). This regression highlights the challenges of relying on cross-modal uncertainty differences without proper regularization. Specifically, Δu, by definition, quantifies the difference in uncertainty levels between the textual and visual modalities. However, without any regularization term to ensure the correctness of the uncertainty estimates, Δu lacks reliability; it fails to capture the "true" uncertainty gap between modalities, and thus, cannot meaningfully guide the optimization process. In essence, Δu requires reliable and calibrated uncertainty estimates from both modalities to meaningfully quantify their disparity. Without such calibration, the model cannot accurately assess the comparative "value" of each modality for ambiguity resolution, leading to inconsistent predictions and compromised performance.

When all three components were integrated, the BLEU score increased significantly to 42.82 ($+1.25$), demonstrating the synergistic interaction among u_{multi}, $L(\Theta)$, and Δu. Each component provides unique benefits:

- u_{multi} encourages confident and low-uncertainty predictions at the token level, improving translation consistency.
- $L(\Theta)$ ensures stable and well-calibrated uncertainty estimation, preventing overfitting and misalignment of cross-modal predictors.

- Δu reinforces the contribution of visual inputs by dynamically prioritizing uncertainty reduction across modal inputs, ensuring that image features are utilized effectively to resolve textual ambiguities.

The comprehensive framework not only improves prediction accuracy but also ensures that visual modality consistently contributes to reducing translation uncertainty, as demonstrated by the elimination of uncertainty increases observed in our error analysis.

5.2 UNDERSTANDING UNCERTAINTY

In this section, we aim to demonstrate through experiments that our Uncertainty-Guided Multimodal Machine Translation (UG-MMT) framework possesses the ability to comprehend and manage uncertainty effectively. To establish this capability, we need examples that elicit high uncertainty outputs alongside those that convey low uncertainty.

Given our focus on translation tasks, a common scenario that naturally arises is the out-of-vocabulary (OOV) situation. OOV refers to cases where the translation encounters words not present in the existing vocabulary, analogous to the occurrence of unseen categories in multiclass classification tasks. These situations should theoretically prompt high uncertainty outputs, indicating the model's recognition of unfamiliar terms. Conversely, words well within the vocabulary should yield lower uncertainty, showing confidence in prediction. Hence, we utilize the OOV scenario as a test to verify whether our model can accurately understand and express uncertainty.

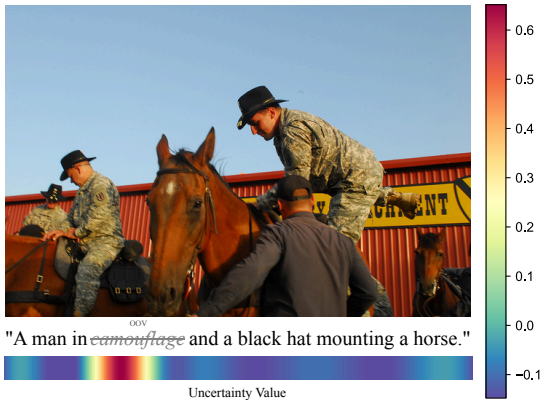


Figure 2: Example of UG-MMT handling uncertainty

Through our experiments, particularly on the Test2017 and MSCOCO datasets, we observed that OOV words consistently resulted in elevated uncertainty scores. This explicit signaling reflects the model's cautious approach when faced with unknowns, dynamically incorporating contextual cues to refine predictions. For instance (see in figure 2), in translating "a man in camouflage and a black hat mounting a horse," the term "camouflage"—absent from the dataset—induced a heightened uncertainty score (0.6), whereas more familiar terms like "man" showed minimal uncertainty (0.005). This distribution underscores the model's ability to distinguish between OOV words and familiar vocabulary, adapting its prediction strategy accordingly.

Quantitative analysis further confirmed that sentences containing high-uncertainty tokens typically achieved lower BLEU and METEOR scores. This correlation highlights the value of uncertainty flags in guiding the model to adjust its predictions amidst linguistic ambiguity. By enabling the system to recognize and act upon uncertainty, UG-MMT enhances both translation accuracy and reliability.

6 CONCLUSION

We proposed UG-MMT, an Uncertainty-Guided Multimodal Machine Translation framework that systematically integrates uncertainty modeling into multimodal translation tasks. By explicitly modeling token-level and sequence-level uncertainties, UG-MMT ensures effective utilization of visual information to disambiguate linguistic ambiguities. UG-MMT eliminates multimodal uncertainty and achieves SOTA performance on Multi30K. These results highlight the importance of combining uncertainty modeling with cross-modal fusion, paving the way for more robust applications of multimodal translation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Project 62476089.

References

Feiyu Chen, Jie Shao, Anjie Zhu, Deqiang Ouyang, Xueliang Liu, and Heng Tao Shen. Modeling hierarchical uncertainty for multimodal emotion recognition in conversation. IEEE Transactions on Cybernetics, 54(1): 187–198, 2022.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In Proceedings of the 5th Workshop on Vision and Language, pages 70–74. Association for Computational Linguistics, 2016.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In Proceedings of the Second Conference on Machine Translation, Volume

2: Shared Task Papers, pages 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4718.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26876–26885, 2024.

Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23262–23271, 2023.

Myong Chol Jung, He Zhao, Joanna Dipnall, and Lan Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation. Advances in Neural Information Processing Systems, 36, 2024.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. Valhalla: Visual hallucination for machine translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5216–5226, 2022.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. Dynamic context-guided capsule network for multimodal machine translation. In Proceedings of the 28th ACM international conference on multimedia, pages 1320–1329, 2020.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. Supervised visual attention for multimodal neural machine translation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4304–4314, 2020.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In International Conference on Machine Learning, pages 3956–3965. PMLR, 2018.

Ru Peng, Yawen Zeng, and Junbo Zhao. Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation, 2023. URL https://arxiv.org/abs/2210.04468.

Rico Sennrich. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems, 31, 2018.

Hiroki Takushima, Akihiro Tamura, Takashi Ninomiya, and Hideki Nakayama. Multimodal neural machine translation using cnn and transformer encoder. Advances in Natural Language Processing, 85, 2019.

Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. Encoder–decoder calibration for multimodal machine translation. IEEE Transactions on Artificial Intelligence, 5(8):3965–3973, 2024.

Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In 2020 IEEE International Conference on Robotics and Automation (ICRA), pages 5716–5723. IEEE, 2020.

Susanne Trick, Dorothea Koert, Jan Peters, and Constantin A Rothkopf. Multimodal uncertainty reduction for intention recognition in human-robot interaction. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 7009–7016. IEEE, 2019.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. arXiv preprint arXiv:2005.00963, 2020.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation, 2021. URL https://arxiv.org/abs/2105.14462.

Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4346–4350, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.400. URL https://aclanthology.org/2020.acl-main.400/.

Yongjing Yin, Jiali Zeng, Jinsong Su, Chulun Zhou, Fandong Meng, Jie Zhou, Degen Huang, and Jiebo Luo. Multi-modal graph contrastive encoding for neural machine translation. Artificial Intelligence, 323:103986, 2023. ISSN 0004-3702.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. In International Conference on Learning Representations, 2020.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. Region-attentive multimodal neural machine translation. Neurocomputing, 476:1–13, 2022.