TriFine: A Large-Scale Dataset of Vision-Audio-Subtitle for Tri-Modal Machine Translation and Benchmark with Fine-Grained Annotated Tags

Boyu Guan[1,2], Yining Zhang[1,2,3], Yang Zhao[1,2]*, Chengqing Zong[1,2]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Zhongguancun Academy, Beijing, China
{guanboyu2022, zhangyining2024}@ia.ac.cn, {yang.zhao, cqzong}@nlpr.ia.ac.cn

Abstract

Current video-guided machine translation (VMT) approaches primarily use coarse-grained visual information, resulting in information redundancy, high computational overhead, and neglect of audio content. Our research demonstrates the significance of fine-grained visual and audio information in VMT from both data and methodological perspectives. From the data perspective, we have developed a large-scale dataset TriFine, the first vision-audio-subtitle tri-modal VMT dataset with annotated multimodal fine-grained tags. Each entry in this dataset not only includes the triples found in traditional VMT datasets but also encompasses seven fine-grained annotation tags derived from visual and audio modalities. From the methodological perspective, we propose a **F**ine-grained **I**nformation-enhanced **A**pproach for **T**ranslation (FIAT). Experimental results have shown that, in comparison to traditional coarse-grained methods and text-only models, our fine-grained approach achieves superior performance with lower computational overhead. These findings underscore the pivotal role of fine-grained annotated information in advancing the field of VMT.

- **Source Subtitle**: It's actually **overwhelming**.
- **Audio Sentiment Tag: positive**
- **Target Subtitle**: 这真是太令人高兴了。
  (It is actually so ***delightful****.)*
- **Text-only MT**: 这确实令人难以忍受。
  (It's really ***unbearable****.)*
- **MT + Frames:** 这实际上是压倒性的。
  (This is actually quite ***compelling****.)*
- **MT + Audio Sentiment Tag**: 这真是太令人开心了。
  (It is actually so ***delightful****.)*

Figure 1: An example in TriFine. Only source text or with video frames cause false translation (in **red**); while the audio sentiment as complementary disambiguation cue (in **blue**) generate the correct translation (in **green**).

1 Introduction

Multimodal machine translation (MMT) enhances the quality of translations by integrating contextual information derived from complementary modalities in addition to textual input. Early MMT research mainly centered on image-guided machine translation tasks (Zhang et al., 2020; Fang and Feng, 2022; Futeral et al., 2023). Recently, video-guided machine translation (VMT) has garnered significant attention and emerged as a prominent approach (Wang et al., 2019; Kang et al., 2023; Li et al., 2023a; Li et al., 2023b). Compared to image-guided machine translation that utilizes image to translate source text, VMT utilizes corresponding video clip to translate video subtitle. Current state-of-the-art VMT approaches employ visual models to extract coarse-grained features from video frames, which are then integrated with text for both training and inference.

Our experiments (Section 6.1) reveal that previous VMT methods exhibit two key limitations: 1) **Information redundancy and high computational overhead.** The existing approaches require selecting multiple frames from video to extract coarse-grained visual features. This not only decelerates the processing speed but also introduces information redundancy that is irrelevant to the translation task. 2) **The overlooked audio information in VMT studies.** Prior work on VMT has focused solely on visual information from videos, neglecting to analyze the impact of inherent audio information on the VMT task.

Figure 1 illustrates that traditional VMT methods fail to correctly translate *"overwhelming"* when confronted with a large number of irrelevant and noisy video frames. However, incorporating the speaker's positive audio sentiment, a fine-grained

*Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8215–8231
January 19–24, 2025.

yet crucial detail, enhances the probability of assigning a positive connotation to *"overwhelming"*.

Based on the observations above, we propose mitigating these limitations by introducing fine-grained information from visual and audio modalities. On one hand, this can provide the translation system with more detailed information from both modalities, sufficient for completing the translation tasks (e.g. audio sentiment in Figure 1, and the location and action in Figure 2, with more examples in the appendix A). On the other hand, introducing fine-grained information can also alleviate the problems of information redundancy and high computational overhead caused by excessive irrelevant images.

Therefore, we introduce **TriFine**, a large-scale vision-audio-subtitle **Tri**-modal VMT dataset with multimodal **Fine**-grained tags. The dataset contains 1.2 million English-to-Chinese (En→Zh) entries and 1.18 million Chinese-to-English (Zh→En) entries. Each entry features the standard bilingual subtitles-video clip triplet common in VMT datasets, enriched with seven fine-grained tags from audio and visual modalities: `audio sentiment`, `audio stress`, `expression`, `action`, `location`, `entities`, and `video caption`, along with seven test sets for systematic analysis. Meanwhile, to validate the efficacy of fine-grained information in the VMT task, we propose a novel VMT method called FIAT. FIAT is a model-agnostic input-enhanced approach that utilizes multimodal tags and soft attention mask for VMT task without altering the existing neural machine translation model. Experiments demonstrate that FIAT significantly outperforms both the text-only approach and traditional VMT methods using coarse-grained visual feature, such as the Transformer Video Encoder (Shurtz et al., 2024), across all three metrics.

In summary, our main contributions are three-fold: 1) We have constructed TriFine, which is the first large-scale tri-modal VMT dataset that comprises triplets, seven types of fine-grained tags from audio and visual modalities, and seven test sets. 2) We propose a novel model-agnostic input-enhanced VMT method FIAT, which makes use of multimodal fine-grained tags and soft attention mask. FIAT is the first to validate the effectiveness of fine-grained information from audio and visual modalities for the VMT task. 3) Extensive experiments on the TriFine dataset using FIAT have been conducted to set a benchmark for the research on this topic. The FIAT dataset and the code of our FIAT framework can be accessed at `https://github.com/BoyuGuan/TriFine`.

2 Related Work

Multimodal Machine Translation (MMT) aims to enhance translation by utilizing information from other modalities, typically visual modality, in addition to textual information (Shen et al., 2024). Unlike translation tasks that involve other modalities, such as speech translation (Liu et al., 2020; Zhang et al., 2023a; Yu et al., 2024b), document image translation (Ma et al., 2023; Zhang et al., 2023b; Zhu et al., 2023a; Tian et al., 2023; Liang et al., 2024;) or simultaneous machine translation (Zhang and Feng 2023; Guo et al., 2024; Yu et al., 2024a), in multimodal machine tasks, the information from other modalities typically plays a supportive, augmentative role for the textual modality. The dominant source of information remains the text rather than the information derived from other modalities. Currently, multimodal machine translation can be primarily categorized into two types: image-guided machine translation and video-guided machine translation.

Image-guided Machine Translation. Image-guided machine translation (IMT) refers to the task of translating a text while simultaneously providing one or more images associated with the text to supplement the information and improve the translation quality. The introduction of the Multi30K dataset (Elliott et al., 2016) accelerated the development of MMT, leading to various methods (Ive et al., 2019; Lin et al., 2020; Caglayan et al., 2021; Li et al., 2022a; Huang et al., 2023; Zhu et al., 2023b; Liu et al., 2020; Fei et al., 2023; Yang et al., 2024).

Video-guided Machine Translation. Similar to IMT, video-guided multimodal machine translation (VMT) uses videos alongside text to enhance translation. With the introduction of VMT datasets such as How2 (Sanabria et al., 2018), VATEX (Wang et al., 2019), VISA (Li et al., 2022b), EVA (Li et al., 2023b), BIGVIDEO (Kang et al., 2023), and MAD-VMT (Shurtz et al., 2024), a series of VMT methods have emerged (Hirasawa et al., 2020; Gu et al., 2021). However, the existing VMT datasets exhibit limitations in one or more aspects such as scale, diversity, and audio-subtitle alignment. The VMT task still faces a data scarcity (Shen et al., 2024), especially the lack of datasets that

Class	Num	Visual					Audio			Others
		Caption	Location	Action	Entity	Expression	Sentiment	Pattern	Stress	
En→Zh	250	221	142	92	189	88	110	24	74	3
Zh→En	250	212	133	112	178	71	67	32	57	4
Sum	500	433	275	204	367	159	177	56	131	7
Percentage(%)		86.6	55.0	40.8	73.4	31.8	35.4	11.2	26.2	1.4
In TriFine		✓	✓	✓	✓	✓	✓	✗	✓	✗

Table 1: Human evaluation on 500 samples to assess the effectiveness of various multimodal fine-grained information in aiding translation. Fine-grained information annotated with ✓ are annotated and analyzed in this paper, whereas those marked with ✗ are excluded.

- **10-s Video clip** (with audio):

- **Source Subtitle**: A lot of bugs.
- **Target Subtitle**: 很多虫子。

Multimodal Fine-grained Tags	
Action: run away	**Audio Sentiment:** negative
Expression: none	**Entities**: people, beach, net
Location: beach	**Audio Stress:** 0.97, 1.02, 0.91, 1.08
Caption: Two people on a beach by the sea, one of them runs away quickly after touching a fishing net, while the other one has been standing on the right side.	

Figure 2: A sample of English-to-Chinese translation from the TriFine dataset.

can support a comprehensive analysis of the impact of fine-grained information from visual and audio modalities on the VMT task.

Several studies (Caglayan et al., 2019; Wang et al., 2019; Long et al., 2024) suggest that visual information offers limited assistance in multimodal machine translation, emphasizing the need for effective use of multimodal data. OVC (Wang and Xiong, 2021), EMMT (Huang et al., 2021) and MSCTD (Liang et al., 2022), which utilize multimodal fine-grained information on entities and sentiments respectively, present a promising direction. Building upon those, we introduce the TriFine dataset, which explores a novel approach by including seven types of fine-grained tags from visual and speech modalities.

3 Dataset

To maintain consistency with the previous VMT datasets (Wang et al., 2019; Kang et al., 2023; Li et al., 2023b; Shurtz et al., 2024), our TriFine dataset also focuses on the En-Zh VMT task.

3.1 Motivation

As we mentioned before, the current VMT methods suffer from information redundancy, high computational overhead, and overlooking audio information. A feasible mitigation approach is to incorporate fine-grained multimodal information in place of the original coarse-grained visual features. In this section, we answer the following two key questions: i) *what fine-grained information have we chosen to incorporate?* And ii) *what are the underlying reasons for our selection*?

We intuitively selected 8 types of fine-grained information from the visual modality (video caption, location, action, entities, expression) and audio modality (sentiment, pattern, stress). We randomly chose 500 samples challenging for text-based translation. Three annotators independently evaluated these, choosing one or more types of fine-grained information that could aid in translation for each sample. If the required multimodal information was not among the eight types specified, it was categorized as "Others." The detailed evaluation guidelines used in this paper can be found in the Appendix B.1. The evaluation results are shown in Table 1, it indicates that seven out of eight types of fine-grained information could significantly aid translation; captions (86.6%) are most effective, while audio patterns (11.2%) are minimally beneficial. **Therefore, we selected the following seven types of fine-grained information: video caption, location, action, entities, expression, audio sentiment, and audio stress.**

3.2 Dataset Overview

As show in Table 2, the TriFine dataset consists of 1.2 million En→Zh entries and 1.18 million Zh→En entries in train set. Figure 2 presents an entry from the TriFine dataset. Each data entry in the TriFine dataset contains a <Chinese subtitle, video clip, English subtitle> triplet similar to traditional VMT datasets, but also includes seven fine-grained multimodal tags: `audio sentiment`, `audio stress`, `video caption`, `location`, `action`, `entities`, and `expression`. Following previous work, we created a general test set and an ambiguity test set for the TriFine dataset. Ad-

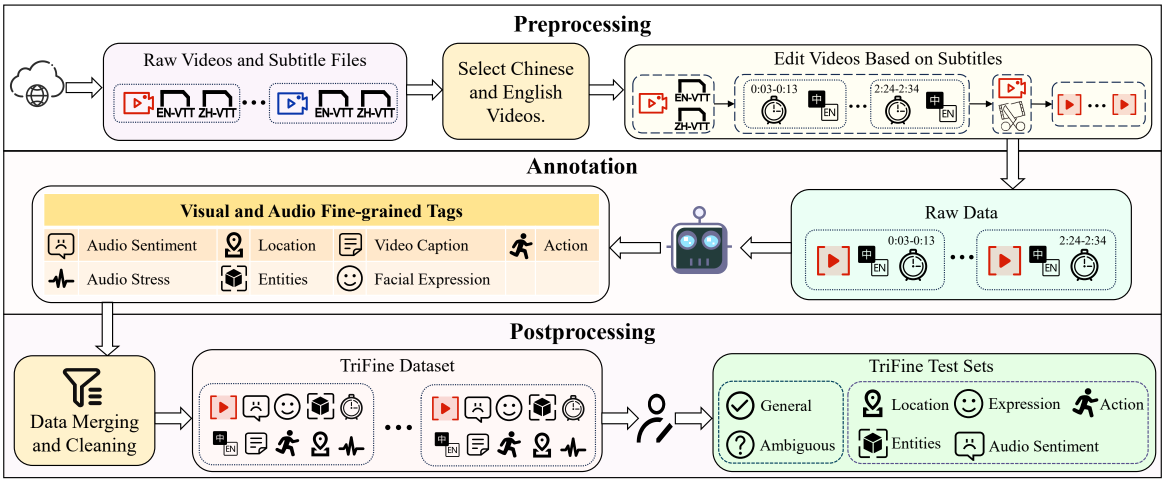


Figure 3: The whole process of TriFine dataset construction. Each entry in the TriFine dataset includes Chinese-English subtitle pair, video clip, subtitle timestamps, and seven fine-grained tags from audio and visual modalities. We constructed general and ambiguity test sets like previous VMT datasets. Furthermore, we developed five information-specific test sets, each focusing on a particular type of fine-grained multimodal information: location, expressions, audio sentiments, actions, and entities.

Class	# Videos	# Clips	AM	# FG
		Train		
En→Zh	18K	1.20M	Auto	7
Zh→En	12K	1.18M		
		Test		
General (En→Zh)	5463	7,000	Auto	7
General (Zh→En)	5892	7,000	Auto	
Ambiguous	35	1,001	Manual	
Location	31	1,000	Manual	
Entities	32	1,000	Manual	
Action	30	1,000	Manual	
Audio Sentiment	29	500	Manual	
Expression	29	500	Manual	

Table 2: Detailed statistics of the TriFine dataset. AM is short for the annotation method, and # FG indicates the types of fine-grained tags.

ditionally, we developed five information-specific test sets rich in audio sentiment, entities, action, and facial expression, location respectively.

3.3 Dataset Construction

The full process for constructing the TriFine dataset is shown in Figure 3.

3.3.1 Preprocessing

Data source. We collected videos along with their corresponding Chinese and English subtitles from YouTube[1]. The video IDs were obtained through a combination of self-collection and selection from the BIGVIDEO (Kang et al., 2023) dataset. All selected videos feature manually uploaded Chinese and English subtitles, ensuring higher quality compared to automatically translated or automatically recognized subtitles.

Language	# Video	# Clip	Precision
ALL	133K	10.39M	-
None	50K	2.34M	91.50%
English	21K	2.48M	98.50%
Chinese	14K	1.98M	99.00%
Others	48K	3.59M	93.75%

Table 3: The videos collected in the initial stage, which included both Chinese and English subtitles, were subjected to video language statistics and manual accuracy evaluation (randomly selecting 400 samples).

Select Chinese and English videos. To assess the impact of audio on VMT, we utilized the SpeechBrain toolkit (Ravanelli et al., 2021; Ravanelli et al., 2024) to recognize the language from three random segments of each video. If two or three segments were identified as the same language with an average confidence score above 0.65, the video was categorized under that language. We applied the above strategy to classify the spoken language in videos with both Chinese and English subtitles collected during the initial phase. The classification results, along with the precision of the manual evaluation based on a random sample of 400 videos, are shown in Table 3. Videos labeled "None" lack human voices or have excessively noisy audio. Only Chinese and English videos with corresponding subtitles were retained. Videos were segmented into 10-second clips centered on subtitle midpoints, preserving original timings for replication and analysis.

[1] https://www.youtube.com

Dataset	Language	Domain	# Clip	Duration	# FG	Audio	Amb	Info-spec	A-S Align
How2 (2018)	En-Pt	instruction	189K	5.8s	0	✓	✗	✗	✓
VATEX (2019)	EN-Zh	caption	41K	10s	0	✓	✗	✗	✗
VISA (2022b)	En-Ja	subtitle	40K	10s	0	✗	✓	✗	✗
MSCTD (2022)	En-Zh/De	subtitle	172K	-	1	✗	✗	✗	✗
EVA (2023b)	En-Zh/Ja	subtitle	1.4M	10s	0	✗	✓	✗	✗
BigVideo (2023)	En-Zh	subtitle	3.3M*	8s	0	✓	✓	✗	✗
MAD-VMT (2024)	En-Zh	caption	193K	-	0	✗	✗	✗	✗
TriFine (Ours)	En-Zh	subtitle	2.4M	10s	7	✓	✓	✓	✓

Table 4: Dataset statistics for TriFine and comparable VMT datasets. "# FG" denotes the count of fine-grained tag types. "Amb" and "Info-spec" indicate ambiguity and information-specific test sets. "A-S Align" signifies audio-subtitle alignment. *Note: BigVideo initially reported 4.5 million clips, but only 3.3 million are publicly accessible due to privacy constraints.

modality	Category	Accuracy	# Samples
Visual	Location	89.50%	400
	Entity	88.00%	
	Expression	86.50%	
	Action	93.25%	
	Caption	93.75%	
Audio	Audio Sentiment	79.50%	400

Table 5: Manual accuracy assessment of 400 randomly chosen samples from TriFine's automatically annotated data.

3.3.2 Annotation

We employed MiniCPM-Llama3-V-2.5 (Yao et al., 2024), a multimodal large language model, to annotate five fine-grained visual tags (location, expression, entities, action, caption) using distinct strategies (e.g. staged reasoning) tailored to each tag's characteristics. The multimodal large model annotated tags based on the video frames corresponding to the start, middle, and end timestamps of the source subtitles. Specific prompts and strategies used are detailed in Appendix B.2. Audio segments corresponding to subtitle timestamps were processed using emotion2vec (Ma et al., 2024) to annotate audio emotions. Following Liang et al. (2022), we classified the eight types of audio emotion outputs into polarity labels: positive (happy), neutral, and negative (angry, disgusted, fearful, sad) and "unknown". The "Other" and "surprised" labels were disregarded due to their combined insignificant proportion (sum less than 1.5%). Each data entry was annotated in the same language as the audio of the corresponding video clip. Three annotators independently evaluated 400 random samples, results are shown in Table 5. The evaluation rules are as specified in Appendix B.3. Audio stress is calculated by the formulas presented in Section 4.2.

3.3.3 Postprocessing

Data cleaning. We conducted data cleaning on the dataset using four methods: Chinese-English sentence length ratio, COMET score (Rei et al., 2020), fast_align (Dyer et al., 2013), and the quality of multimodal fine-grained tags. The detailed process is described in Appendix B.4. 40% of the substandard data was filtered out.

Test sets. We randomly selected 7,000 entries each from En→Zh and Zh→En TriFine subsets to form a general test set. To evaluate the role of multimodal fine-grained tags in translation disambiguation, we constructed a 1,001-sample high-quality ambiguity test set from 14,000 entries (criteria are detailed in Appendix B.5). We developed five information-specific test sets enriched with facial expressions, actions, locations, audio sentiments, and entities to evaluate fine-grained tags' impact on translation quality. The ambiguity and information-specific test sets were selected by three annotators. Information-specific test sets are compiled as follows: 1,000 each for entities, location and action, and 500 each for audio sentiment and facial expression.

3.4 Comparison with Existing VMT Datasets

Table 4 presents a comparison between the TriFine dataset and all existing VMT datasets. In terms of scale, the TriFine dataset is second only to BigVideo (Kang et al., 2023) far surpassing other datasets. Compared to other datasets, TriFine includes seven types of fine-grained information from visual and audio modalities. The video clips in the TriFine dataset all contain audio that corresponds to the subtitle.

4 Method

We previously emphasized the significance of fine-grained information for VMT task in manual anal-

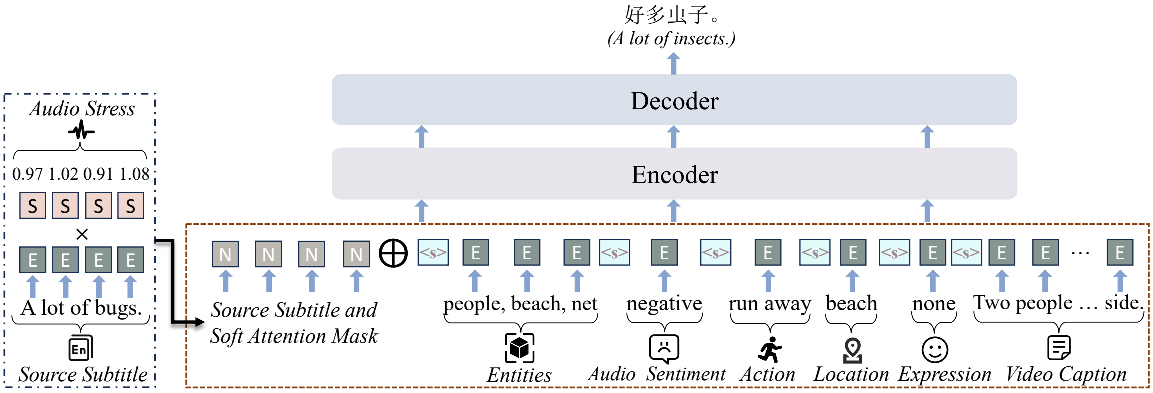


Figure 4: Diagram of FIAT. It employs seven fine-grained types of information from visual and audio modalities to enhance the input. E, S, and N are abbreviations for embedding, soft attention mask, and new embedding, respectively. <s> stands for separate token.

ysis. However, it is still unclear whether these fine-grained information proves to be effective for neural machine translation models. To validate its applicability in models, we developed FIAT, the first audio-and-visual-aware VMT framework. FIAT utilizes fine-grained tags and soft attention masks to enhance input, enabling the completion of VMT task without modifying the neural machine translation model architecture.

4.1 FIAT

The architecture of FIAT is illustrated in Figure 4. As previously mentioned, each entry in the TriFine dataset contains not only bilingual subtitle pair but also seven types of multimodal fine-grained information: entities, audio sentiment, action, location, expression, video caption, and audio stress. We obtain embeddings for the source language subtitle L and all fine-grained tags T except for audio stress.

$$L_e[le_1, \dots, le_N] = \text{Embedding}(L) \tag{1}$$

$$T_e^i[te_1, \dots, te_{N_i}] = \text{Embedding}(T^i) \tag{2}$$

The T^i in Equation 2 represents different multimodal fine-grained tags. Then, we compute the soft attention mask sa_i for each embedding le_i of sentence L using audio stress values (the detailed calculation process is described in Section 4.2). For each pair of le_i and sa_i, we compute their dot product to obtain the new embedding ne_i.

$$ne_i = sa_i \cdot le_i \tag{3}$$

Construct the multimodal fine-grained information-enhanced input E by concatenating sentence $L_{ne} = \{ne_i\}_{i=1}^{M}$ with six types of multimodal fine-grained tags T^i, using the $\langle\text{sep}\rangle$ token as a separator between each element.

$$S = \oplus_{i=1}^{6}(\langle\text{sep}\rangle \oplus T_i) \tag{4}$$

$$E = L \oplus S \tag{5}$$

After inputting E into the encoder and decoder of the existing neural machine translation model, the target subtitle L_T is obtained.

$$L_T = \text{Decoder}(\text{Encoder}(E)) \tag{6}$$

It can be observed that our FIAT method is highly modular, facilitating the combination of various fine-grained multimodal information.

4.2 Soft Attention Mask

In speech, word emphasis is often achieved through stress modulation. In FIAT, to incorporate stress information from speech, we propose the soft attention mask. In contrast to traditional attention masks using binary values, the soft attention mask calculates the root mean square of the audio associated with the source subtitle. It then scales the values to fit a normal distribution $N = (\mu, \sigma^2)$ and truncates it within the range $[Min, Max]$. The process for this calculation is detailed below.

For the embeddings $L_e = \{le_i\}_{i=1}^{N}$ of a subtitle L, $A = [a_1, \dots, a_M]$ is the corresponding audio of subtitle. We calculate the root mean square of le_i using the audio A.

$$am'_i = \sqrt{\frac{1}{\lfloor \frac{M}{N} \rfloor} \sum_{k=start_i}^{\lfloor \frac{M}{N} \rfloor \cdot i} a_k^2} \tag{7}$$

In Equation (7), $start_i = \lfloor \frac{M}{N} \rfloor \cdot (i-1) + 1$. Next, we map $\{am'_i\}_{i=1}^{N}$ onto a normal distribution.

$$\tilde{am}_i = \frac{am'_i - \bar{am'}}{\bar{am'}} \cdot \sigma + \mu \tag{8}$$

In Equation 8, $\bar{am}' = \frac{\sum_{i=1}^{N} am'_i}{N}$. Truncating $\tilde{am_i}$ to the interval $[Min, Max]$.

$$am_i = \min(\max(\tilde{am_i}, Min), Max) \tag{9}$$

Finally, we obtain the soft attention mask am_i for each embedding le_i of source subtitle.

5 Experiments

Baselines. We set the Transformer (Vaswani et al., 2017) model as the text-only baseline model, configured with 6 layers in both the encoder and decoder, 8 attention heads, a hidden size of 512, and an FFN size of 2048. We chose the Transformer Video Encoder (TVE) and Conformer Video Encoder (CVE), proposed by Shurtz et al. (2024), as baseline methods for traditional VMT approaches. These two methods employ visual models to extract coarse-grained visual features from multi-frame images, which are then interacted with text. To ensure fair comparison, the encoder and decoder in FIAT share the same architecture as the transformer model in the text-only baseline.

Setup. In the computation of the soft attention mask, we utilized a truncated normal distribution with $\mu = 1$ and $\sigma = 0.05$ within the range $[0.9, 1.1]$. We employ three widely used metrics to evaluate the results: SacreBLEU[2] (Post, 2018), METEOR (Rei et al., 2020), and COMET (Rei et al., 2020). During the inference, the beam size and the length penalty are set to 4 and 1.0. The results are averaged over three different random seeds. More details can be found in Appendix C.

6 Results and Analysis

6.1 Main Results

Table 6 presents the main experimental results of all methods on the general test set.

Multimodal vs. text-only. The experimental results presented in rows 1 and 13 indicate that our FIAT method, which integrates multimodal fine-grained information, significantly outperforms the text-only baseline. Specifically, FIAT achieved an improvement of 1.93/1.84 BLEU score and 2.53/1.94 METEOR score, and 1.73/1.31 COMET score on Zh→En and En→Zh, respectively.

Fine-grained vs. coarse-grained. Based on rows 2, 3, and 13, we can conclude that our FIAT method, which utilizes fine-grained multimodal information, not only greatly outperforms traditional VMT methods such as TVE and CVE that rely on coarse-grained visual information, but also substantially reduces computational overhead. TVE and CVE exhibit similar performances in our experiments, aligning with Shurtz et al.'s (2024) findings. Specifically, compared to TVE, FIAT achieved an improvement of 1.66/1.51 BLEU score, 2.11/1.59 METEOR score, and 1.01/0.84 COMET score on Zh→En and En→Zh, respectively.

Audio information in VMT. The results of rows 1, 4, and 5 indicate that audio information is beneficial for the VMT task. Row 5 shows that using audio sentiment information alone yields significantly better results for the En→Zh task compared to Zh→En, likely due to more semantically rich emotional content in English audio. This highlights the necessity of including audio aligned with source language subtitles in the VMT dataset.

Selection of fine-grained information. The table shows that caption is the most effective single multimodal fine-frained information type to incorporate, aligns with the manual evaluation results in Section 3.1. Notably, fusing all fine-grained information except caption (row 12) achieves comparable results to integrating all information types, with significantly reduced computational overhead. We explore further relationships between captions and other fine-grained multimodal information in Appendix D.

Regularization. Some studies (Caglayan et al., 2019; Wu et al., 2021) suggest that in multimodal machine translation, visual features may primarily serve as regularization. However, comparing rows 1, 2, 6 and 9, we can observe that although using the same input format, FIAT+Expression (row 6) performs notably worse than the baselines, while FIAT+Entities (row 9) far exceeds them. The observed discrepancy highlights the critical role of multimodal tag content, rather than simple regularization, in determining the efficacy of translation models.

6.2 Results On Ambiguity Test Set

Similar to previous VMT methods, a key application of FIAT is in handling translation disambiguation tasks. Table 7 presents the experimental results of each method on the ambiguity test set. Our FIAT method shows substantial improvements over the text-only, TVE, and CVE methods on the ambiguity test set in all three metrics. This indicates that our FIAT method maintains superior performance in handling translation disambiguation tasks.

[2] https://github.com/mjpost/sacrebleu

	Method	Zh→En			En→Zh			GPU Hours↓
		BLEU↑	METEOR↑	COMET↑	BLEU↑	METEOR↑	COMET↑	
1	Text-only	23.58	47.86	71.86	36.22	45.16	75.17	**8.7**
2	TVE	23.85	48.28	72.58	36.55	45.51	75.64	182.1
3	CVE	23.97	48.30	72.60	36.43	45.42	75.58	193.6
	FIAT (Ours)							
4	+ Stress	23.72	48.25	72.75	36.58	45.64	75.64	11.6
5	+ Sentiment	23.78	48.25	72.78	37.17	45.96	75.96	8.8
6	+ Expression	22.33	46.26	71.25	33.54	43.11	74.14	8.8
7	+ Action	24.05	48.34	72.65	36.65	45.67	75.70	8.9
8	+ Location	23.82	48.15	72.20	36.70	45.69	75.67	8.9
9	+ Entities	24.56	49.10	72.88	37.14	46.24	75.89	9.0
10	+ Caption	24.71	49.48	73.14	37.76	47.06	76.33	27.4
11	+ Stress + Sentiment + Caption	24.88	49.62	73.26	38.00	**47.11**	76.41	28.3
12	+ ALL (except Caption)	25.45	50.38	73.55	37.75	46.52	76.23	12.4
13	+ ALL	**25.51**	**50.39**	**73.59**	**38.06**	**47.11**	**76.48**	28.8

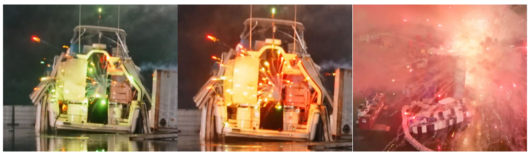
Table 6: The main experimental results of various methods on the general test set of the TriFine dataset. All results are mean values of three different random seeds. The GPU Hours reported in the table represent the average training time for both En→Zh and Zh→En. The Best result in each column is in **bold**.

Method	BLEU	METEOR	COMET
Text-only	29.85	42.22	74.39
TVE	30.37	42.73	74.45
CVE	30.28	42.66	74.39
FIAT + ALL (Ours)	**31.24**	**44.89**	**75.93**

Table 7: Experimental results on the ambiguity test set.

6.3 Results on Information-specific Test Sets

Table 8 shows the experimental results on information-specific test sets. In the information-specific test sets for audio sentiment, action, location, and entities, results indicate that the FIAT method utilizing solely the corresponding fine-grained information outperforms FIAT methods employing other individual fine-grained information types. Furthermore, these results are comparable to those obtained when utilizing all 7 types of fine-grained information, and significantly surpasses the baselines. This demonstrates that our FIAT method can effectively utilize these four types of fine-grained multimodal information. However, FIAT underperforms on the information-specific test of expression compared to text-only and TVE methods, consistent with general test set results. In future work, we will focus on developing automated methods for selecting optimal fine-grained information based on specific scenarios, potentially improving performance across all information types, including expression.



- **Source Subtitle**: **It's hitting** his yacht.
- **Entities Tag**: **fireworks**, boat, water
- **Target Subtitle**: 烟火正在打击他的游艇。
  *(The **fireworks** are **striking** his yacht.)*
- **Text-only MT**: 它正在撞击他的游艇。
  *(**It's colliding** with his yacht.)*
- **MT** + **Entity Tag**: 烟花正在打击他的游艇。
  *(The **fireworks** are **striking** his yacht.)*

Figure 5: A case study of the TriFine dataset. The word "fireworks" in entities tag can resolve the "It" in the original text and deduce the correct translation of the corresponding verb "hitting".

6.4 Case Study

Figure 5 illustrates a case study from the TriFine dataset. In the video, fireworks are attacking a yacht through explosions. The English subtitle "It's" refers to "fireworks," and identifying this through frames or the entity itself aids in accurate Chinese translation. Given the context of fireworks, "striking" is a more appropriate translation for "hitting" than "colliding." Our FIAT method addresses issues of unclear references and polysemous words in text-only machine translation by using an additional entities tag for accurate translations.

7 Conclusion

In this paper, we introduce a new VMT dataset called TriFine, the first vision-audio-subtitle tri-modal VMT dataset with annotated fine-grained tags. In addition to the common tuple found in the VMT dataset, each entry in the dataset also includes seven types of fine-grained multimodal information. To validate the effectiveness of multi-modal fine-grained information in the VMT task, we introduce the first audio-and-visual-aware VMT framework, FIAT. FIAT utilizes multimodal fine-

Set Method	Sentiment		Expression		Action		Location		Entities	
	B	M	B	M	B	M	B	M	B	M
Text-only	31.30	44.25	28.55	42.36	29.30	41.97	30.53	43.11	26.80	41.56
TVE	31.54	44.51	28.68	42.63	29.63	42.12	30.56	43.24	27.29	41.83
FIAT (Ours)										
+ Sentiment	32.66	45.86	28.63	42.53	29.93	42.42	31.01	43.87	28.35	42.31
+ Expression	29.35	43.18	25.35	39.71	27.26	40.14	27.19	40.32	24.40	39.69
+ Action	31.72	44.81	28.89	42.98	30.24	42.91	30.73	43.40	27.05	41.71
+ Location	31.83	44.82	28.72	42.84	29.99	42.66	31.43	44.15	28.17	42.36
+ Entities	32.45	45.80	**29.04**	43.00	30.08	42.71	31.32	44.14	28.69	42.51
+ ALL	**32.95**	**46.24**	29.01	**43.06**	**30.42**	**43.51**	**31.92**	**44.39**	**29.08**	**43.07**

Table 8: The experimental results of information-specific test sets in terms of BLEU (B) and METEOR (M). The best result in each column is in **bold**, and the underlined value is the best result for methods other than FIAT+ALL in each column.

grained information to enhance the inputs and complete the VMT task without altering the existing neural machine translation model. The experimental results demonstrate the effectiveness of the FIAT method. We hope that our dataset and method can inspire further research in the field.

8 Limitations

Although we only selected videos with manually uploaded Chinese and English subtitles and conducted extensive data cleaning, there are still some low-quality data in the dataset. To manage costs and time with the expansive TriFine dataset, we did not choose larger models for automatic annotation. Utilizing more powerful models may yield higher-quality annotated data. In this paper, the primary purpose of the proposed FIAT method is to verify the effectiveness of fine-grained multimodal information for the VMT task, which still requires the input of given tags. In future work, we will explore methods for utilizing fine-grained information to aid VMT through cascaded or end-to-end architectures.

9 Ethical Considerations

The data was collected in strict adherence to the source's terms of use and copyright policies, utilizing tools commonly employed in previous research. Videos were divided into 10-second segments, with the number of clips per video limited to avoid the potential substitution of the original video. Users of our dataset will be required to sign a usage agreement, stipulating that the dataset is to be used exclusively for academic research purposes. Before the dataset is released, we will perform an additional filtering process to eliminate any privacy and copyright-sensitive content to the best extent possible.

Acknowledgments

We would like to begin by expressing our sincere gratitude to Dr. Tong Rong from the Singapore Institute of Technology for her invaluable guidance on audio data annotation. We also extend our thanks to Zhiyang Zhang, Yupu Liang, Donglei Yu, and Jianghao Chen from the CIP Lab at the Institute of Automation, Chinese Academy of Sciences, for their insightful and constructive feedback. Our deep appreciation goes to the anonymous reviewers for their valuable suggestions. Furthermore, we acknowledge the annotators who contributed to the dataset construction. The research work has been supported by the Natural Science Foundation of China under Grant No. 62336008.

References

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

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.

Dipankar Das, Naveen Mellempudi, Dheevatsa Mudigere, Dhiraj Kalamkar, Sasikanth Avancha, Kunal Banerjee, Srinivas Sridharan, Karthik Vaidyanathan, Bharat Kaul, Evangelos Georganas, et al. 2018. Mixed precision training of convolutional neural networks using integer operations. In *International Conference on Learning Representations*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. 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.

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

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

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.

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

Shoutao Guo, Shaolei Zhang, and Yang Feng. 2024. Decoder-only streaming transformer for simultaneous translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8851–8864, Bangkok, Thailand. Association for Computational Linguistics.

Tosho Hirasawa, Zhishen Yang, Mamoru Komachi, and Naoaki Okazaki. 2020. Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020. *arXiv preprint arXiv:2006.12799*.

Xin Huang, Jiajun Zhang, and Chengqing Zong. 2021. Entity-level cross-modal learning improves multi-modal machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1067–1080, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xin Huang, Jiajun Zhang, and Chengqing Zong. 2023. Contrastive adversarial training for multi-modal machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

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

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

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

Yihang Li, Shuichiro Shimizu, Chenhui Chu, Sadao Kurohashi, and Wei Li. 2023b. Video-helpful multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4281–4299, Singapore. Association for Computational Linguistics.

Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. 2022b. VISA: An ambiguous subtitles dataset for visual scene-aware machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6735–6743, Marseille, France. European Language Resources Association.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. MSCTD: A multimodal sentiment chat translation dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2601–2613, Dublin, Ireland. Association for Computational Linguistics.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou.

2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095, Mexico City, Mexico. Association for Computational Linguistics.

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

Yuchen Liu, Jiajun Zhang, Hao Xiong, Long Zhou, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2020. Synchronous speech recognition and speech-to-text translation with interactive decoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8417–8424.

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. *arXiv preprint arXiv:2404.06107*.

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

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. CCIM: Cross-modal cross-lingual interactive image translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4959–4965, Singapore. Association for Computational Linguistics.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15747–15760, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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

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

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaelle Laperriere, Mickael Rouvier, Renato De Mori, and Yannick Esteve. 2024. Open-source conversational ai with speechbrain 1.0. *Preprint*, arXiv:2407.00463.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

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

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

Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo, and Bin Wang. 2023. In-image neural machine translation with segmented pixel sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15046–15057, Singapore. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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.

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

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint 2408.01800*.

Donglei Yu, Xiaomian Kang, Yuchen Liu, Yu Zhou, and Chengqing Zong. 2024a. Self-modifying state modeling for simultaneous machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9781–9795, Bangkok, Thailand. Association for Computational Linguistics.

Tengfei Yu, Xuebo Liu, Liang Ding, Kehai Chen, Dacheng Tao, and Min Zhang. 2024b. Speech sense disambiguation: Tackling homophone ambiguity in end-to-end speech translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8020–8035, Bangkok, Thailand. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2023. Hidden markov transformer for simultaneous machine translation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuhao Zhang, Chen Xu, Bei Li, Hao Chen, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023a. Rethinking and improving multi-task learning for end-to-end speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10753–10765, Singapore. Association for Computational Linguistics.

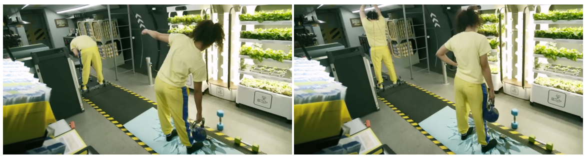
Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023b. LayoutDIT: Layout-aware end-to-end document image translation with multi-step conductive decoder. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10043–10053, Singapore. Association for Computational Linguistics.

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

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023a. PEIT: Bridging the modality gap with pre-trained models for end-to-end image translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13433–13447, Toronto, Canada. Association for Computational Linguistics.

Yaoming Zhu, Zewei Sun, Shanbo Cheng, Luyang Huang, Liwei Wu, and Mingxuan Wang. 2023b. Beyond triplet: Leveraging the most data for multimodal machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2679–2697, Toronto, Canada. Association for Computational Linguistics.

A Examples



- **SRC Subtitle**: We're going to come out so **ripped** at the end of this.
- **Action Tag**: **exercise**
- **Target Subtitle**: 我们在结束后一定会变得很强壮。
  *(We'll be **strong** when it's over.)*
- **Text-only MT:** 我们会在这一切结束的时候，变得如此撕裂。
  *(We'll be so **torn up** at the end of it all.)*
- **MT + Action Tag**: 我们到最后时一定会变得很强壮。
  *(We will be **strong** in the end.)*

Figure 6: The fine-grained multimodal information of the action "exercise" effectively aids in translating the word "ripped" to "strong". SRC is the abbreviation of source.

Figure 6 illustrates a sample from the TriFine dataset. In this example, all words in the source text except for "ripped" can be directly translated using text translation. However, for this particular word, once it is known that the associated action is "exercise," it can be accurately translated as "strong".

B Dataset

B.1 Annotation Guidelines

Classification requires agreement from at least two out of three annotators. All annotators in this study are native Chinese speakers proficient in English, and they have been compensated fairly.

B.2 Visual Fine-grained Tags Annotation

We created bilingual prompts in both Chinese and English for the multimodal Large Language Model (LLM) to generate fine-grained tags consistent with the source language. The detailed methodology for generating individual fine-grained tags is outlined as follows:

Entities tag. We employ a specifically designed prompt, as illustrated in Figure 7, to guide the multimodal large language model in our task. This prompt instructs the model to analyze frames corresponding to the midpoint of subtitle timestamps. The objective is to identify and output the main entities present in these frames. To maintain focus and relevance, we impose a constraint limiting the output to a maximum of three entities.

Location tag. We employed the prompt illustrated in Figure 8 to instruct the multimodal LLM to generate location information for the frame corresponding to the midpoint of each subtitle.

Expression tag. We employed the prompt shown in Figure 9 to enable the multimodal LLM to generate fine-grained expression tags from the facial expressions of the main character in the frame corresponding to the midpoint of the subtitle timing. Furthermore, we instructed the multimodal LLM to select from five categories of expression labels: angry, happy, sad, neutral, and surprised.

Action tag. Since the action in the video is a dynamic process, our annotation method for action fine-grained tagging adopts a two-step staged reasoning: 1) We first use the prompt in the left column of Figure 10 to guide the multimodal LLM in annotating the human actions in the frames corresponding to the midpoint of the subtitle and its adjacent 0.5 seconds, resulting in three action words or phrases. 2) Then we input these three actions along with the middle-time frame of the subtitles into the multimodal LLM, and use the prompt in the right column of Figure 10 to obtain the overall fine-grained action tag.

Video caption tag. Our annotation method for video captions also employs a two-step staged reasoning: 1) We initially use the prompt in the left column of Figure 11 to have a multimodal LLM generate descriptions for the visuals corresponding to the start, middle, and end moments of the caption. 2) We then input these three descriptions, along with the visual from the middle moment of the caption, into the multimodal LLM using the prompt in the right column of Figure 11 to obtain a comprehensive description of the entire video segment.

B.3 Evaluation Rules For Annotated Data

Given that our objective in annotating these fine-grained information is to facilitate the VMT task, we have developed and implemented a set of evaluation criteria that we proposed ourselves, which are highly correlated with the VMT task. These samples were again annotated by three annotators who are native Chinese speakers and proficient in English. For a sample to be ultimately determined as correct, it required confirmation from at least two of the annotators. The specific annotation guide-

Multi-modal LLM Entities Tag Extraction Prompt

你是一个专业的视觉辨别专家，请向我尽可能简洁高效的描述一下这张图片中出现的实体，只回答三个，也即〈实体1〉〈实体2〉〈实体3〉。例如输入图片中出现了猫、树、老鼠，则回答〈猫〉〈树〉〈老鼠〉。如输入图片中出现的实体难以辨别，则回答〈无〉。
请参照以下样例和我输入的图片，按照<实体1>〈实体2〉〈实体3>的格式描述一下这张图片中的。
样例1：〈苹果〉〈篮子〉〈橘子〉
样例2：〈汽车〉〈马路〉〈楼房〉
样例3：〈None〉

As a professional visual identification expert, please provide a concise and efficient description of the entities present in this image. Limit your response to three entities, formatted as <entity1> <entity2> <entity3>. For instance, if the input image contains a cat, a tree, and a mouse, the response should be <cat> <tree> <mouse>. If the entities in the input image are difficult to discern, respond with <nne>.
Please refer to the following examples and the image I have provided, and describe the entities in this image using the format <entity1> <entity2> <entity3>.
Example 1: <apple> <basket> <orange>
Example 2: <car> <road> <building>
Example 3: <none>

Figure 7: Prompt for the multimodal large language model to output fine-grained entities tag.

Multi-modal LLM Location Tag Extraction Prompt

你是一个专业的视觉辨别专家，请向我尽可能简洁高效的描述一下这张图片发生的地点，用<地点>返回画面中的地点信息，如：〈海滩〉。如果画面中没有明确的地点信息，则返回<无>。
请参照以下样例和我输入的图片，按照〈地点>的格式描述一下这张图片发生的地点。
样例1：〈足球场〉
样例2：〈办公室〉
样例3：〈无〉

As an expert in visual recognition, please provide a concise and efficient description of the location depicted in this image. Enclose the location information within <location> tags, such as <beach>. If there is no discernible location information in the image, please respond with <none>.
Please refer to the following examples and the image I have provided, and describe the location of this image using the <location> format.
Example 1: <soccer field>
Example 2: <office>
Example 3: <none>

Figure 8: Prompt for the multimodal LLM to output fine-grained location tag.

lines are as follows:

Audio sentiment and expression. For these two tags, we had already predetermined the categories during the annotation process, thus eliminating the need for additional rule formulation.

Action and location. Each annotator will first independently annotate the fine-grained action and location tags, then determine whether their annotations are consistent with the results of the automatic annotation. Words or phrases with similar or synonymous meanings are considered consistent, such as "office" and "workplace".

Entities. Each annotator will first independently annotate three entities in the video frame. They will then compare their manual annotations with the results of the automatic annotation to determine consistency. The automatic annotation is considered correct only when two or more of the automatically annotated entities match the manual annotation results. Additionally, if an annotator believes that any of the entities would influence the translation, that particular entity must also appear in the automatic annotation results; otherwise, it will be deemed incorrect. As mentioned earlier, words or phrases with similar or synonymous meanings are considered consistent in this evaluation process.

Caption. The automatically generated caption is considered correct only if the annotator determines that it accurately describes the main content of the video and includes any content from the video

Multi-modal LLM Facial Expression Tag Extraction Prompt

你是一个专业的视觉辨别专家，请向我描述一下这张图片中出现的主要人物的面部表情，请从〈生气〉、〈高兴〉、〈伤心〉、〈平静〉、〈惊讶〉五个选项中选择。如输入图片中出现的人物面部表情是生气，则回答〈生气〉。如输入图片中出现的人物面部表情难以辨别，则回答〈无〉。
请参照以下样本和我输入的图片，按照<表情>的格式描述一下这张图片中出现的人物的面部表情，从〈生气〉、〈高兴〉、〈伤心〉、〈平常〉、〈惊讶〉五个选项中选择。
样例1:〈高兴〉
样例2:〈惊讶〉
样例3:〈无〉

You are a professional visual recognition expert. Please describe the facial expressions of the main characters appearing in this image. Choose from the following five options: <angry>, <happy>, <sad>, <neutral>, <surprise>. If the facial expression of the character in the input image is angry, respond with <angry>. If the facial expression of the character in the input image is difficult to discern, respond with <none>.
Please refer to the following samples and the image I input and describe the facial expression of the character appearing in this image in the format of <expression>, choosing from the five options: <angry>, <happy>, <sad>, <neutral>,<surprise>.
Example 1: <happy>
Example 2: <surprise>
Example 3: <none>

Figure 9: Prompt for the multimodal large language model to output fine-grained expression tag.

that may influence the translation (if such content exists).

B.4 Data Cleaning

We have conducted data cleaning on the dataset to improve its quality. We filtered out clips where the Chinese-to-English sentence length ratio was clearly unreasonable, specifically those with a ratio below 0.3 or above 3.0. Next, we used the fast_align (Dyer et al., 2013) tool to align the Chinese and English sentences in the clips, retaining only clips with an alignment ratio of 0.3 or higher. Finally, following BigVideo, we filtered out data with COMET scores below 0.3. We employed rule-based filtering to eliminate 70% of the samples with fewer than 2 valid fine-grained multimodal tags (labeled as "none" or other similarly non-informative categories).

B.5 Selection Criteria for the Ambiguity Test Set

The criteria for selecting instances of ambiguity are twofold: firstly, the text proves challenging to translate accurately even for sophisticated commercial translation systems such as Google Translate[3]; and secondly, the accompanying video clip provides essential contextual information that resolves the ambiguity, facilitating accurate translation.

C Experimental Detatils

During the data annotation process, we utilized 10 NVIDIA V100 32GB GPUs, 12 NVIDIA 3090 GPUs, and 2 NVIDIA A100 80GB GPUs. For all experiments, we used a tokenizer based on the Zh-En tokenizer from OPUS-MT (Tiedemann et al., 2023; Tiedemann and Thottingal, 2020), with an additional special token [sep] added to separate text from fine-grained tags of other modalities. Our code is implemented based on PyTorch (Paszke et al., 2019), Huggingface Transformers (Wolf et al., 2020) and DeepSpeed[4] (Rasley et al., 2020). All experiments are done using mixed-precision training (Das et al., 2018) on four NVIDIA L40 GPUs. We use the AdamW Optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, and set the weight decay to 0.1. The learning rate is set to $2 \times e^{-5}$ with a warmup of 4000 steps. The batch size is set to 256, and the experiment runs for 27,000 steps.

D Caption and Other Tags

Video caption shows superior performance across human evaluations and experimental results. This efficacy likely stems from caption's comprehensive nature, often encompassing other elements like entities and location. However, this raises concerns about potential information redundancy in

[3] https://translate.google.com

[4] https://github.com/microsoft/DeepSpeed

🏃 Multi-modal LLM Action Tag Extraction Prompt	
Single Frame Action Tag Extraction Prompt	**Summary Extraction Prompt**
你是一个专业的视觉辨别专家，请向我尽可能简洁高效的描述一下这张图片中人物的动作，用<动作>返回画面中的地点信息，如：<跑步>。如果画面中没有明确的动作信息，则返回<无>。 请参照以下样例和我输入的图片，按照 <动作>的格式描述一下这张图片发生的地点。 样例1：<跳舞> 样例2：<吃饭> 样例3：<无>	你是一个针对总结任务表现的非常好的专家，我将会输入三个动作单词或短语，分别对应一个视频片段的开始、中间、结束时的动作。同时一起输入的还有视频片段中间的画面。 请根据输入的三个动作文本和相应视频的中间画面给出这个视频里的总动作，用<动作>表示。 样例1：（奔跑）（奔跑）（踢球），总动作为：<踢足球>； 样例1：（跑步）（举哑铃）（俯卧撑），总动作为：<健身>； 样例2：（{运动1}）（{运动2}）（{运动3}），总动作为：
As an expert in visual recognition, please provide a concise and efficient description of the action performed by the individual(s) in this image. Utilize the format <action> to convey the location information depicted in the scene, such as <running>. If the image does not contain explicit action information, respond with <none>. Please refer to the following examples and the image I have provided, and describe the location where the action in this image is taking place using the <action> format. Example 1: <dancing> Example 2: <eating> Example 3: <none>	You are an expert highly proficient in summarizing tasks. I will input three action words or phrases corresponding to the actions at the beginning, middle, and end of a video clip. Additionally, I will include a screenshot from the middle of the video. Please identify the overall action in the video based on the three input action texts and the middle screenshot of the video, denoted by <action>. Example 1: (Running) (Running) (Kicking a ball), the overall action is: <Playing Football>; Example 2: (Running) (Lifting dumbbells) (Push-ups), the overall action is: <Working Out>; Example 3: ({Action1}) ({Action2}) ({Action3}), the overall action is:

Figure 10: The two-stage prompts for generating fine-grained action tags using multimodal LLMs.

fine-grained tags of visual modality. Analysis of Zh→En tasks revealed that methods incorporating tags without caption (row 12 in Table 8) outperformed caption-only with audio tags (row 11 in Table 8) approach, while En→Zh tasks showed the opposite trend. Notably, utilizing all fine-grained information consistently yielded optimal results, suggesting complementary benefits of captions and tags in enhancing translation quality.

Multi-modal LLM Caption Tag Generation Prompt	
Single Frame Caption Generation Prompt	**Caption Summary Prompt**
你是一个视觉方面的专家。你能否简洁明了的对输入图片进行有条理的描述？请你描述的尽可能简洁，最好控制在两句话之内。	你是一个针对总结任务表现的非常好的专家，我将会输入三个视频画面描述，分别对应一个视频片段开始、中间、结束时画面的描述。同时一起输入的还有视频片段中间的画面。 请根据这三个描述和相应视频的中间画面总结出这个视频片段的总描述，请让最后的总描述尽可能简洁。 {描述1}、{描述2}、{描述3} ，视频总描述为：
You are an expert in the field of vision. Could you provide a concise and well-organized description of the input image? Please describe it as concisely as possible. Ideally, limit your response to two sentences.	You are an expert with exceptional performance in summarization tasks. I will input three video scene descriptions, corresponding to the beginning, middle, and end of a video segment. Additionally, I will provide the description of the middle scene of the video segment. Based on these three descriptions and the corresponding central image of the video, please synthesize an overall description of this video segment. Ensure that the overall description is as concise as possible. {caption 1}, {caption 2}, {caption 3}, the overall video description is:

Figure 11: The two-stage prompts for generating video caption tags using multimodal LLMs