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

Boyu Guan

State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS)

Institute of Automation, Chinese Academy of Sciences, Beijing, China

guanboyu2022@ia.ac.cn

Outline

1.The VMT Task

2.Motivation

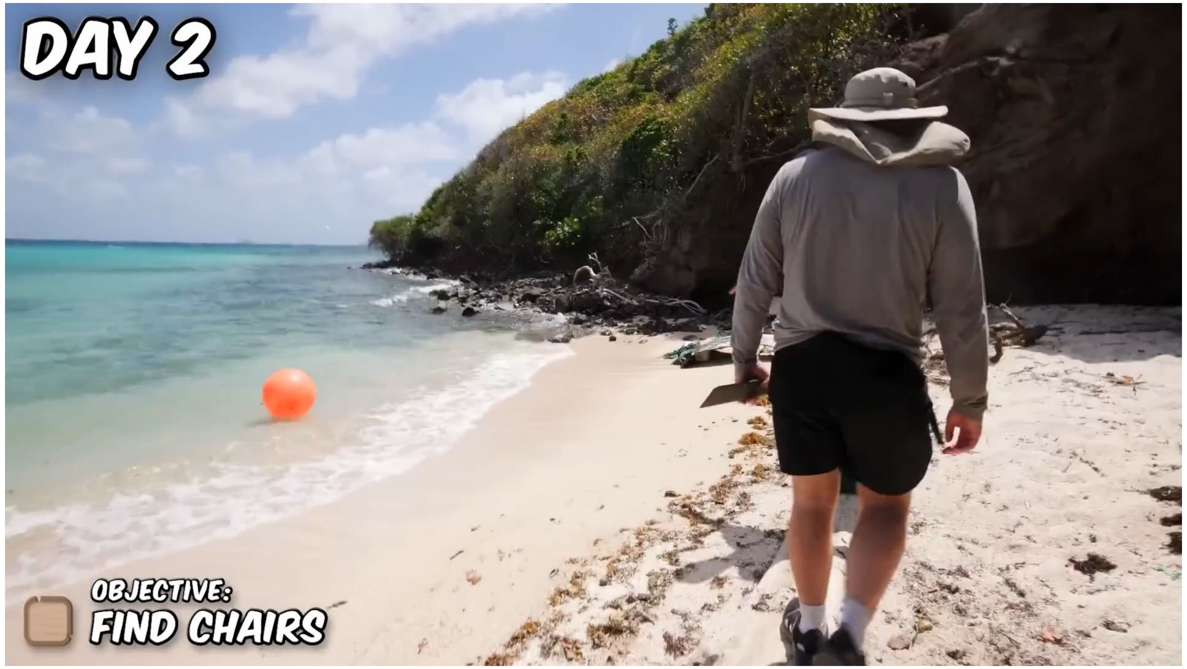
3.TriFine

4.FIAT

5.References

1. The VMT Task

◆ Video-guided Machine Translation(VMT)



SRC : A lot of **bugs**!

NMT: 很多**错误**! ✗
(A lot of **errors**!)

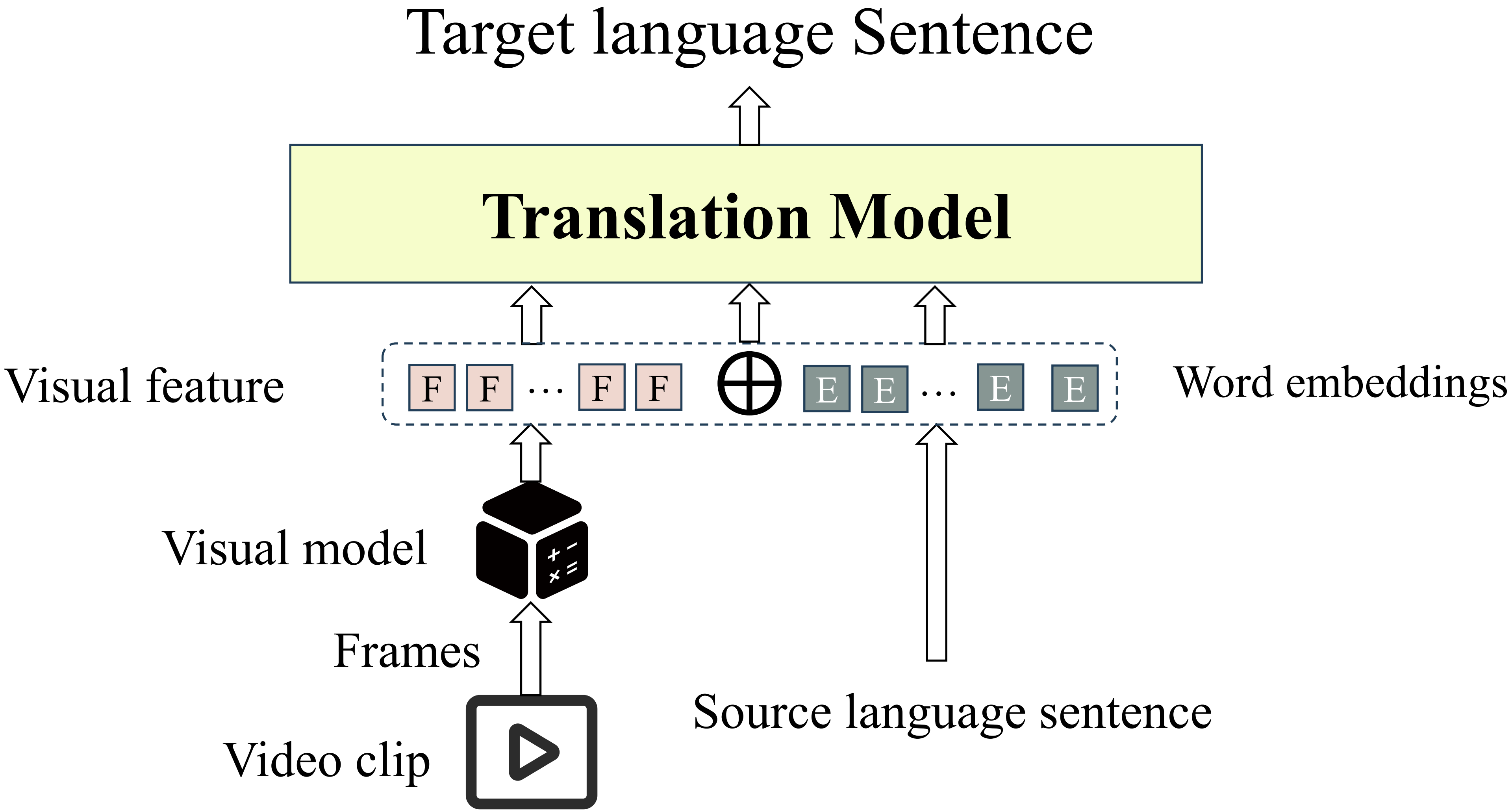
VMT: 很多**虫子**! √
(A lot of **insects**!)

Multimodal machine translation (MMT) enhances the quality of translations by integrating contextual information derived from complementary modalities in addition to textual input.

Video-guided machine translation is a subtask of MMT, which utilizes corresponding video clips to translate video subtitles.

1. The VMT Task

- General Paradigm in VMT Tasks



Outline

1.The VMT Task

2.Motivation

3.TriFine

4.FIAT

5.References

2. Motivation

◆ Two Limitations in Current VMT Research

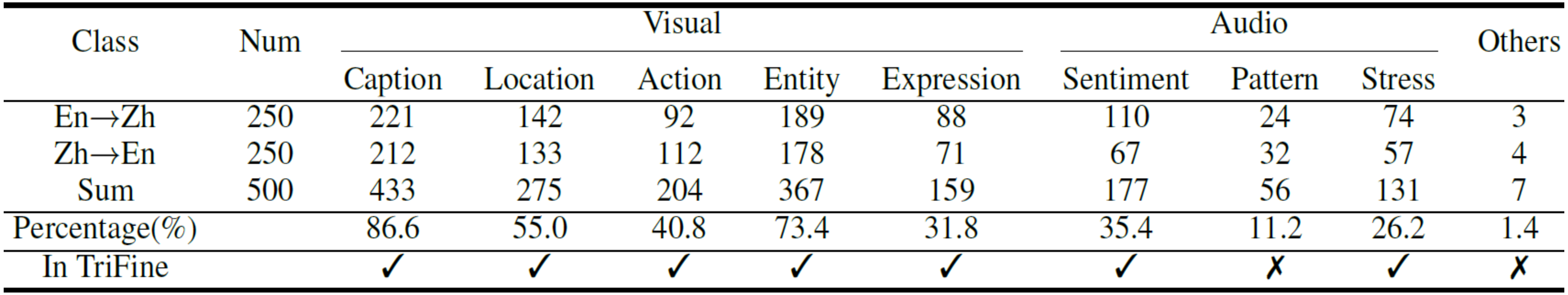
1. **Information redundancy and high computational overhead.** The existing approaches require selecting multiple frames (30-50) from video to extract coarse-grained visual features. This not only decelerates the processing speed but also introduces information redundancy that is irrelevant to the translation 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.

2. Motivation

◆ Human Evaluation

- We selected 500 sentence pairs requiring video-assisted translation to evaluate the role of various fine-grained multimodal information in VMT.



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		✓	✓	✓	✓	✓	✓	✗	✓	✗

- Finally, we annotated and analyzed seven types of fine-grained multimodal labels in TriFine: video caption, location, action, entity, facial expression, audio sentiment, and stress.

Outline

1.The VMT Task

2.Motivation

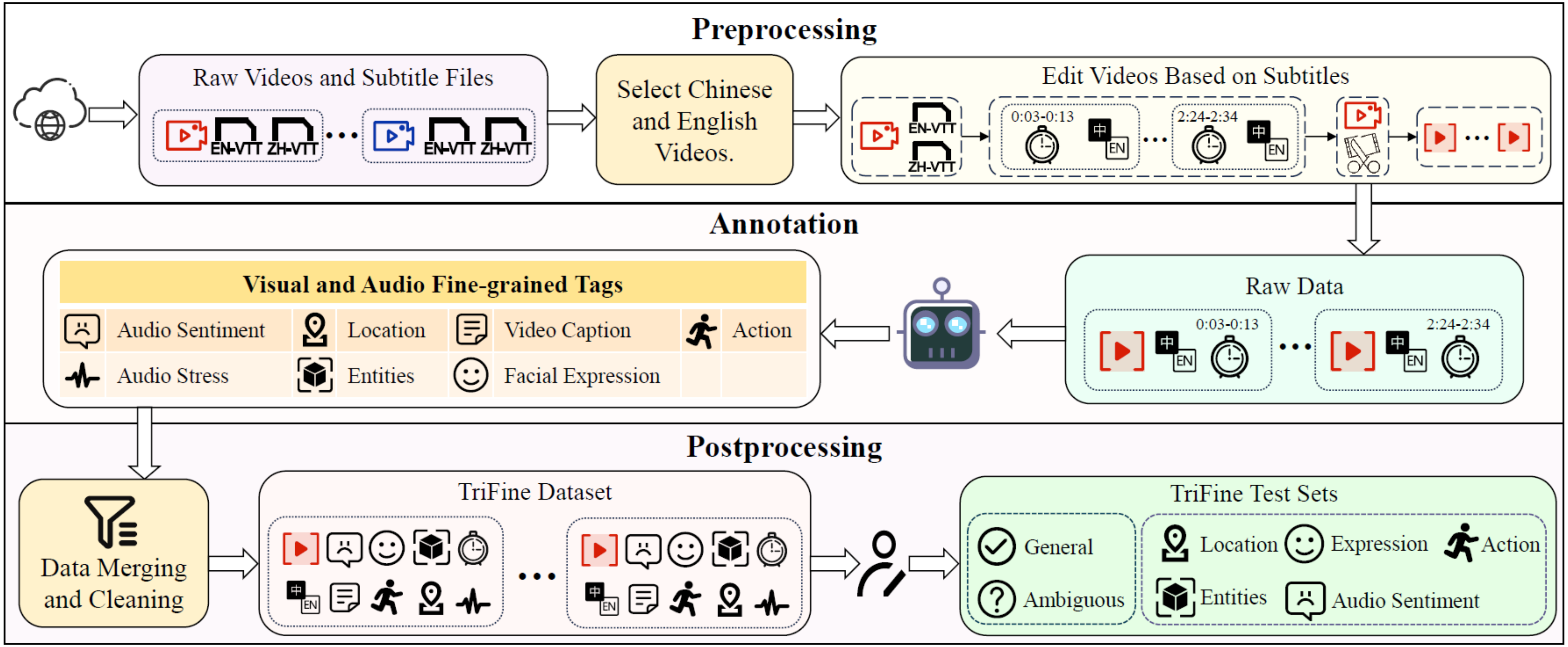
3.TriFine

4.FIAT

5.References

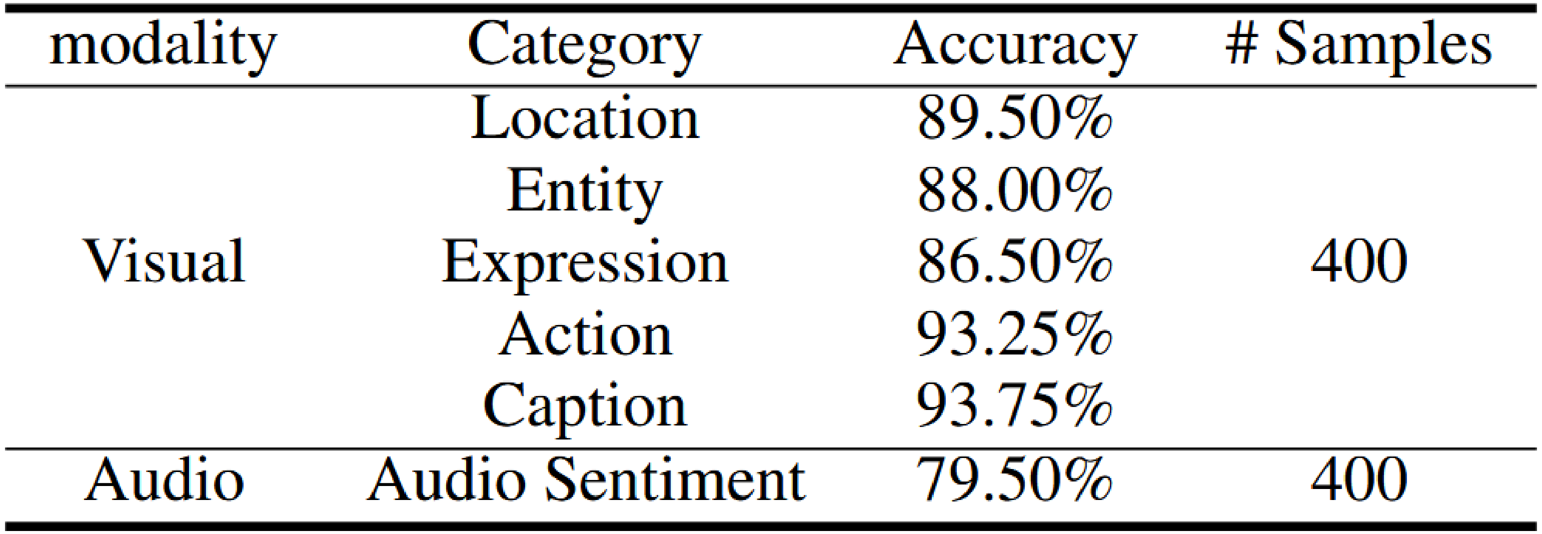
3. TriFine

◆ The whole process of TriFine dataset construction.



3. TriFine

- TriFine Dataset

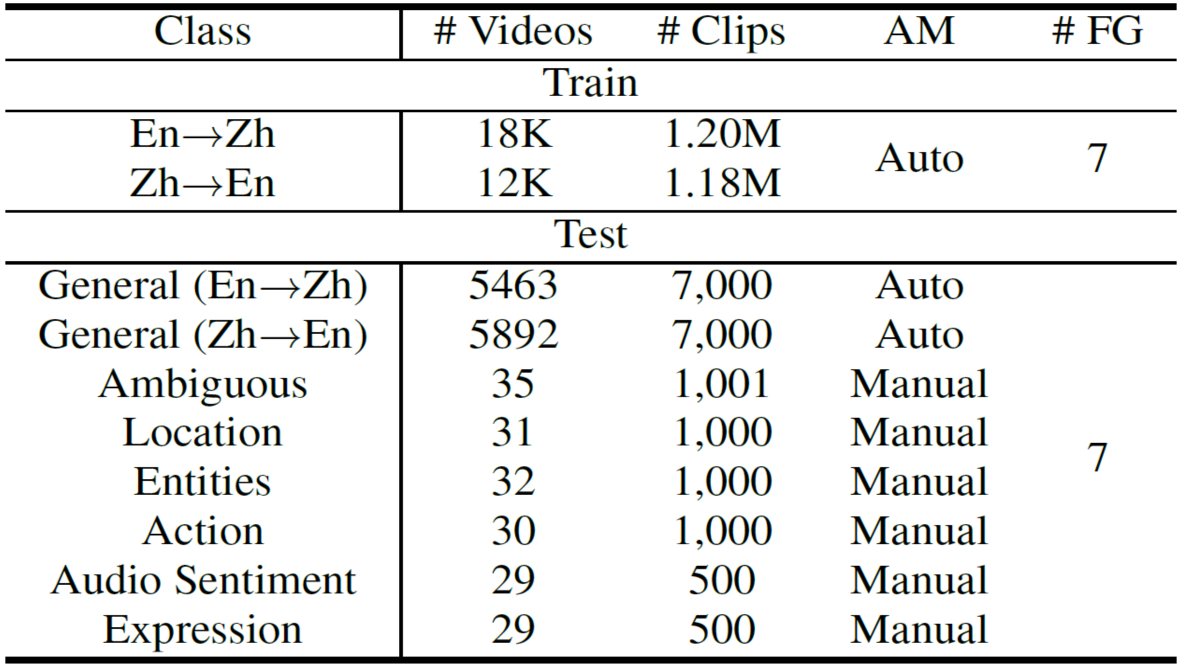


modality	Category	Accuracy	# Samples
Visual	Location	89.50%	400
	Entity	88.00%	
	Expression	86.50%	
	Action	93.25%	
	Caption	93.75%	
Audio	Audio Sentiment	79.50%	400

We randomly selected 400 samples from the automatically annotated data for manual evaluation. With the support of a specific strategy, the annotation accuracy was relatively high.

3. TriFine

- TriFine Dataset



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	

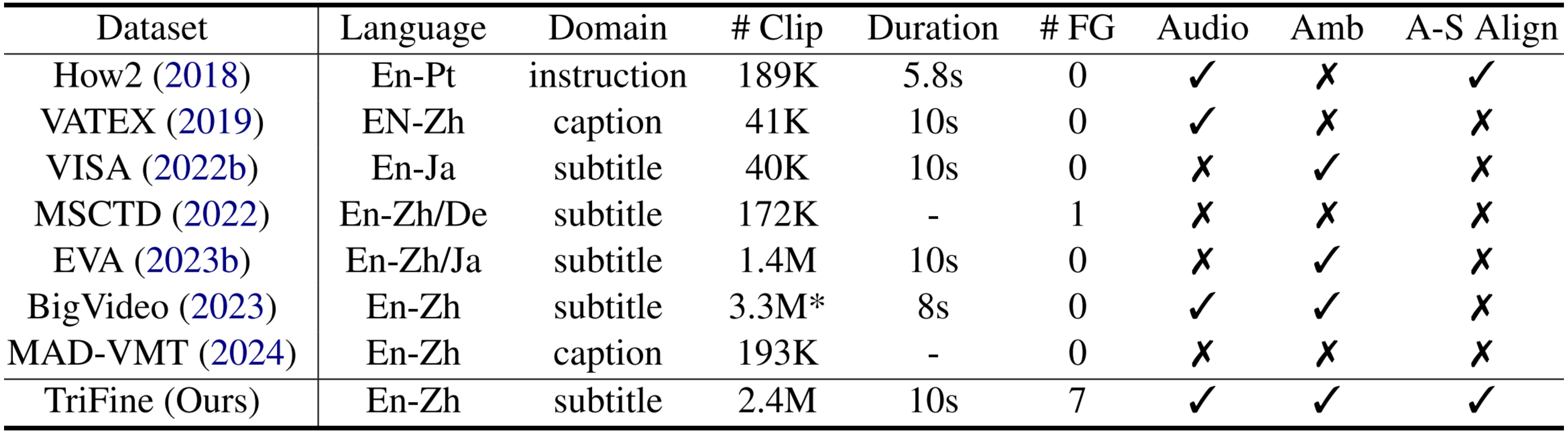
The dataset consists of 1.2 million English→Chinese pairs and 1.18 million Chinese→English pairs.

Each entry contains: source language subtitle, target language subtitle, 10-second video clip (with corresponding audio), seven types of fine-grained multimodal information.

The dataset also includes: a general test set, an ambiguity test set, five specialized test sets enriched with specific information types.

3. TriFine

- Compare With Existing VMT Datasets



Dataset	Language	Domain	# Clip	Duration	# FG	Audio	Amb	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	✓	✓	✓

"# 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.

3. TriFine

◆ Data Sample

- SRC: A lot of bugs.
- TGT: 很多虫子。

10-second video segment with audio：

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
Video 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.	

Outline

1.The VMT Task

2.Motivation

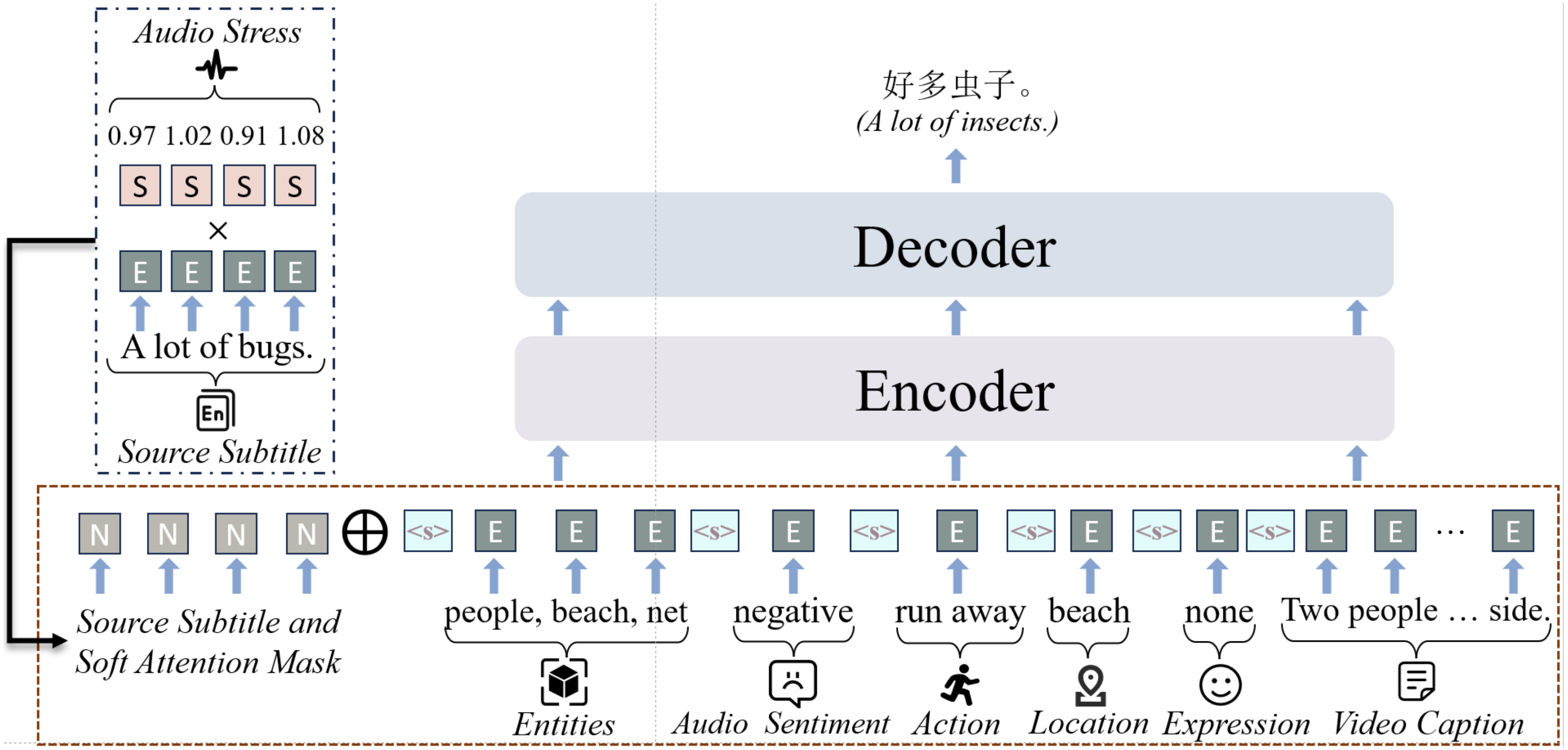
3.TriFine

4.FIAT

5.References

4. FIAT

We propose the first audio-and-visual-aware VMT framework, FIAT (Fine-grained Information-enhanced Approach for Translation), to validate the effectiveness of fine-grained multimodal information in VMT task.



4. FIAT

◆ Baselines

Our experimental evaluation includes two categories of baselines:

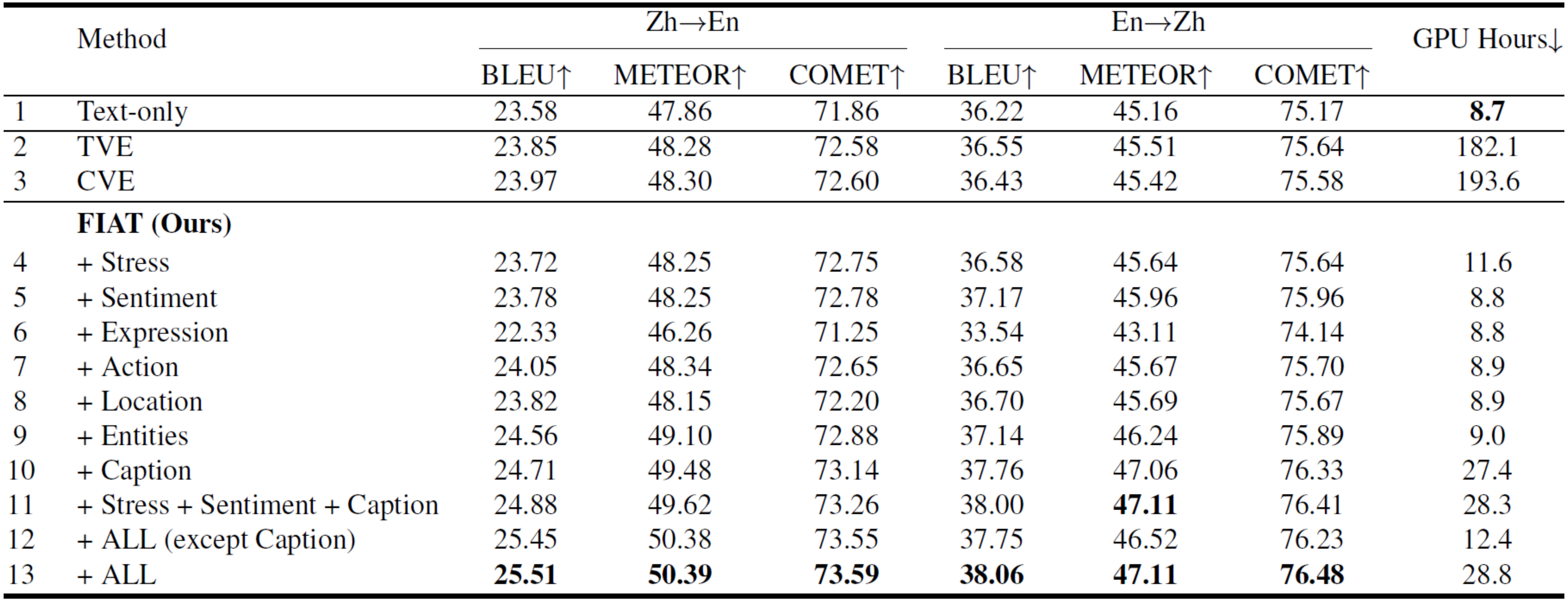
1.A text-only baseline implementing the standard Transformer architecture.

2.Traditional VMT approaches that utilize coarse-grained visual features, specifically TVE and CVE.

The Effects of Pretraining in Video-Guided Machine Translation (Shurtz et al., LREC-COLING 2024)

4. FIAT

◆ Main Results（general test set）

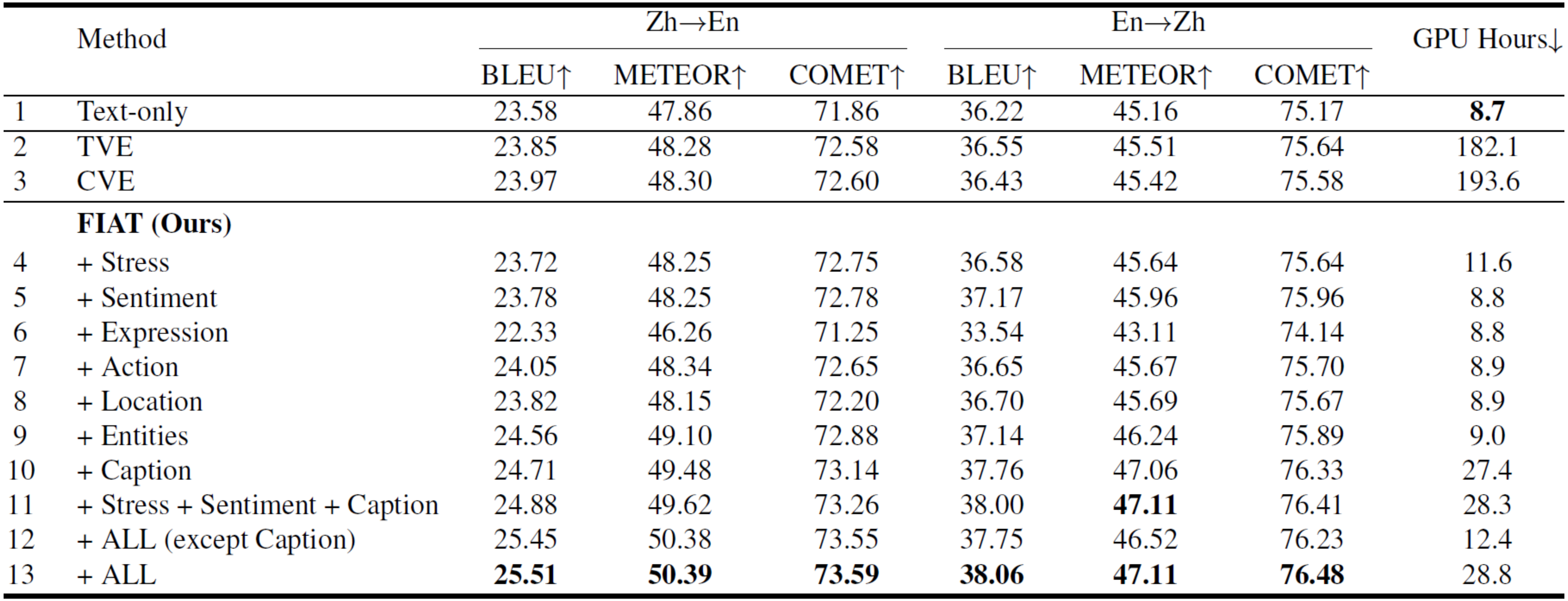


	Method	Zh→En			En→Zh			GPU Hours↓
		BLEU↑	METEOR↑	COMET↑	BLEU↑	METEOR↑	COMET↑	
1	Text-only	23.58	47.86	71.86	36.22	45.16	75.17	**8.7**
2	TVE	23.85	48.28	72.58	36.55	45.51	75.64	182.1
3	CVE	23.97	48.30	72.60	36.43	45.42	75.58	193.6
	FIAT (Ours)							
4	+ Stress	23.72	48.25	72.75	36.58	45.64	75.64	11.6
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

- FIAT surpasses text-only baselines and traditional VMT models with coarse-grained features, while requiring less training time.

4. FIAT

◆ Main Results（general test set）

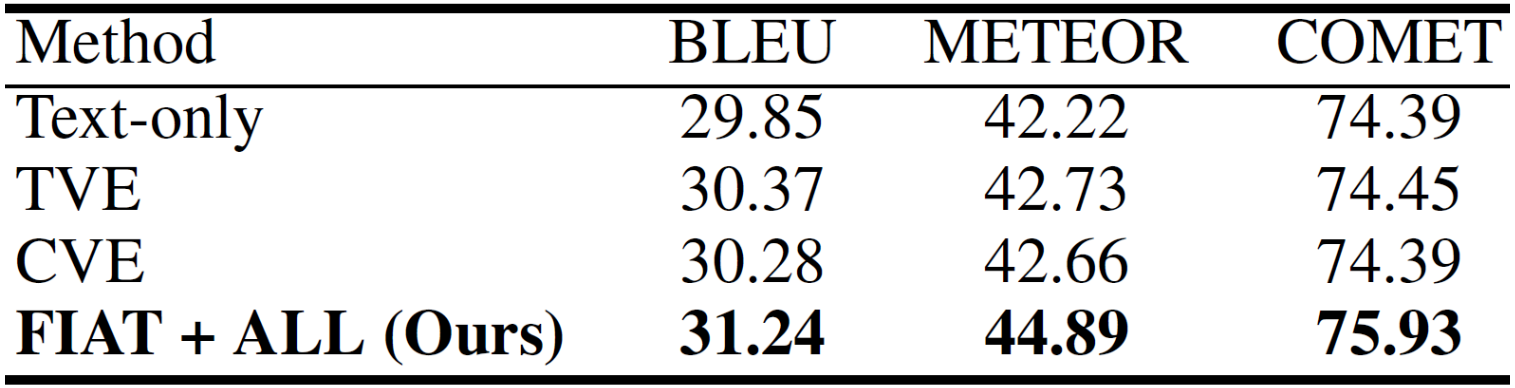


	Method	Zh→En			En→Zh			GPU Hours↓
		BLEU↑	METEOR↑	COMET↑	BLEU↑	METEOR↑	COMET↑	
1	Text-only	23.58	47.86	71.86	36.22	45.16	75.17	**8.7**
2	TVE	23.85	48.28	72.58	36.55	45.51	75.64	182.1
3	CVE	23.97	48.30	72.60	36.43	45.42	75.58	193.6
	FIAT (Ours)							
4	+ Stress	23.72	48.25	72.75	36.58	45.64	75.64	11.6
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

- Audio stress and sentiment improve translation quality. Audio sentiment shows stronger effects in En→Zh than Zh→En translation, reflecting greater emotional variety in English speech.

4. FIAT

◆ Results On Ambiguity Test Set

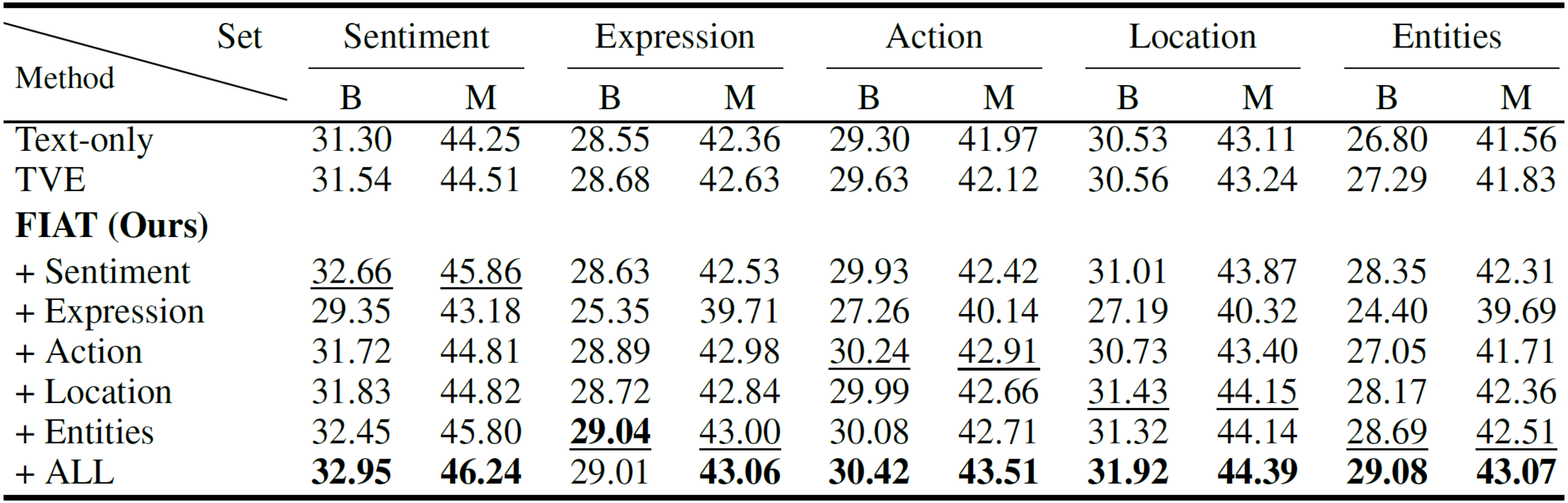


Method	BLEU	METEOR	COMET
Text-only	29.85	42.22	74.39
TVE	30.37	42.73	74.45
CVE	30.28	42.66	74.39
FIAT + ALL (Ours)	**31.24**	**44.89**	**75.93**

In the field of translation disambiguation - a crucial application of VMT - FIAT demonstrates significantly superior performance compared to three baseline methods.

4. FIAT

Results on The Information-rich Test Sets

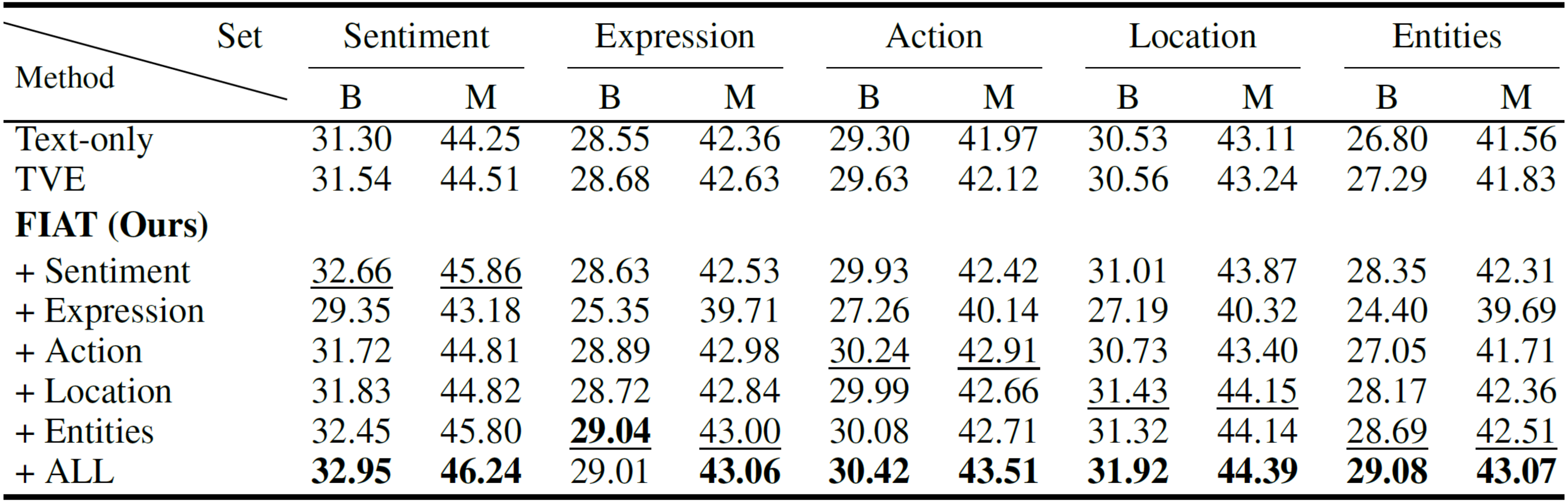


Set	Sentiment		Expression		Action		Location		Entities	
Method	B	M	B	M	B	M	B	M	B	M
Text-only	31.30	44.25	28.55	42.36	29.30	41.97	30.53	43.11	26.80	41.56
TVE	31.54	44.51	28.68	42.63	29.63	42.12	30.56	43.24	27.29	41.83
FIAT (Ours)										
+ Sentiment	32.66	45.86	28.63	42.53	29.93	42.42	31.01	43.87	28.35	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**

- On test sets rich in emotion sentiment, actions, locations, and entities, the FIAT method utilizing only the corresponding fine-grained information significantly outperforms approaches that solely rely on other fine-grained information. Moreover, its performance approximates that of the +ALL method which utilizes all fine-grained information.

4. FIAT

◆ Results on The Information-rich Test Sets



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**

- Consistent with previous results on general test sets, the FIAT method plus only expression performs significantly worse than the baseline in all tests.

Outline

1.The VMT Task

2.Motivation

3.TriFine

4.FIAT

5.References

5. References

1. MSCTD: A Multimodal Sentiment Chat Translation Dataset (Liang et al., ACL 2022)
2. Video-Helpful Multimodal Machine Translation (Li et al., EMNLP 2023)
3. BigVideo: A Large-scale Video Subtitle Translation Dataset for Multimodal Machine Translation (Kang et al., ACL Findings 2023)
4. The Effects of Pretraining in Video-Guided Machine Translation (Shurtz et al., LREC-COLING 2024)
5. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. (Meng et al., arXiv preprint 2020).
6. VALHALLA: Visual Hallucination for Machine Translation (Li et al., CVPR 2023)
7. SpeechBrain: A general-purpose speech toolkit. (Ravanelli et al., arXiv 2021)
8. VoxLingua107: a dataset for spoken language recognition. (Valk et al., SLT 2021)
9. Speech Emotion Diarization: Which Emotion Appears When? (wang et al., arXiv 2023)
10. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. (Yao et al., arXiv 2024)
11. Entity-level cross-modal learning improves multimodal machine translation. (Huang et al., EMNLP Findings 2021)

Thanks!