Towards Visually-Guided Movie Subtitle Translation for Indic Languages

Tarun Chintada[*], **Kshetrimayum Boynao Singh**[*], **Asif Ekbal**
Department of Computer Science and Engineering
Indian Institute of Technology Patna, India
{tarunchintada1, boynfrancis, asif.ekbal}@gmail.com

Abstract

Movie subtitle translation is inherently multimodal, yet text-only systems often miss visual cues needed to convey emotion, action, and social nuance, especially for low-resource Indic languages (English to Hindi, Bengali, Telugu, Tamil and Kannada). We present a case study on five full-length films and compare two lightweight visual grounding strategies: structured attribute summaries from a 5-minute sliding window and free-text summaries of inter-subtitle visual gaps. Our analysis shows that temporal misalignment between subtitles and frames is a major obstacle in long-form video, often rendering indiscriminate visual grounding ineffective. However, oracle selective[1] grounding, which replaces only the lowest-quality 20-30% of baseline segments with visual-enhanced outputs, consistently improves COMET over the text-only baseline while requiring far less visual processing. Among the two approaches, coarse attribute-based visual context summarization is more robust, capturing scene-level emotion and contextual subtle cues that text alone often misses.

1 Introduction

The global demand for movie subtitle translation has grown exponentially with the rise of streaming platforms and international releases. Subtitles must convey meaning under strict spatial and temporal constraints, often compressing conversational speech, idiomatic expressions, and cultural references into short, timed segments. For low-resource Indian languages characterised by rich morphology (Singh et al., 2025a), diglossia, and sparse parallel corpora, these challenges are magnified, and text-only machine translation (MT) systems frequently produce translations that are either literal or contextually inadequate (Singh et al., 2025b; Artetxe et al., 2020).

Films are inherently multimodal: meaning is distributed across dialogue, visual scenes, character actions, and emotional cues. In principle, incorporating visual context could disambiguate references, resolve honorifics, and ground translations in the on-screen situation (Elliott et al., 2016). Yet, unlike traditional multimodal machine translation (MT) tasks (e.g., translating image captions), movie subtitle translation presents two distinctive difficulties.

First, most subtitle segments do not rely on visual information. The majority of dialogues are conversational and can be translated accurately from text alone. Visual grounding is beneficial only in a minority of cases-action cues (Gain et al., 2025; Kumar et al., 2025), emotion-driven (Shen et al., 2025) exchanges, or references to visible objects. For instance, translating the line *"He's coming!"* requires knowing whether the threat is a person, animal, or vehicle; such information is visually available. In contrast, a typical exchange like *"How was your day?"* gains little from the accompanying visuals. This asymmetry makes indiscriminate application of computationally expensive visual processing inefficient and often unnecessary.

Second, visual and textual streams in long-form movies are often misaligned. Subtitles are generated independently of video frames, and cumulative

[*]Equal contribution.

[1]Oracle selective refers to a specialized, often theoretical, method used to achieve the absolute best outcome

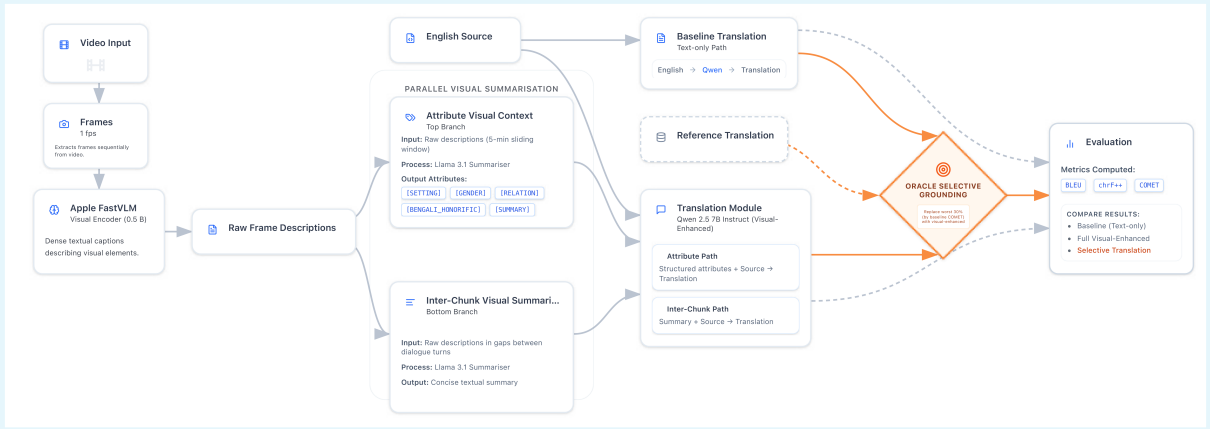

Figure 1: Architecture of the multimodal subtitle translation pipeline

temporal drift can cause a substantial fraction of subtitles to be paired with irrelevant or misleading visuals (Wang and Zhao, 2024). Over a 180-minute film, drift as small as one second per hour accumulates to a three-minute mismatch, affecting a notable portion of subtitle segments. When visual context is misaligned, it ceases to be helpful and can actively degrade translation quality (Appicharla et al., 2024), a phenomenon rarely discussed in the multimodal MT literature (Radinger, 2025).

Motivated by these practical realities, we conduct a case study that systematically compares two summarization-based strategies for integrating visual context into subtitle translation for five Indian languages: Hindi, Bengali, Telugu, Kannada, and Tamil. These languages represent a range of morphological complexity and cultural nuances, making them ideal for studying multimodal subtitle translation in low-resource settings (Eberhard et al., 2025). The two strategies are:

1. **Attribute Visual Context (Attr-VC)**: Aggregates a 5-minute sliding window of raw visual descriptions and summarizes them using Llama 3.1 into structured attributes (e.g., setting, gender, honorifics, emotional intent).

2. **Inter-Chunk Visual Summarization (Inter-VS)**: Summarizes the visual content occurring between dialogue turns (the gaps) into a free-text description using Llama 3.1.

Using subtitles from five full-length movies spanning diverse genres, we evaluate these methods under realistic conditions. A central finding is that applying visual context indiscriminately often degrades performance due to temporal misalignment. However, an *oracle selective grounding* that replaces the worst 20-30% of baseline segments (by baseline COMET) with visual-enhanced translations consistently improves semantic adequacy (COMET) over the text-only baseline, recovering most of the gain while using only a fraction of the visual processing. Coarse attribute-based summarization proves particularly robust, capturing emotional tone and scene-level cues that text alone cannot convey. The key attribute of this work are:

1. **A comparative case study** of two visual summarization strategies for low-resource subtitle translation.
2. **Identification and quantification** of temporal misalignment as a major practical obstacle in long-form multimodal MT.
3. **Empirical evidence** that coarse attribute summarization is resilient to drift and that selective grounding can recover most of the gain.

2 Dataset Preparation and Resources

To evaluate multimodal subtitle translation in realistic settings, we curate a dataset derived from five commercially released movies selected to ensure diversity in genre, narrative style, and visual complexity: *Titanic* (1997), *Skyfall* (2012), *Oppenheimer* (2023), *Spider-Man 2* (2004), and *Avatar 2* (2022). These films span romance, action, historical drama, science fiction, and superhero genres, providing a wide range of dialogue types and visually grounded scenes.

2.1 Movies and Visual Data

For each movie, we extract video frames at 24-fps and align them with subtitle timestamps. Table 1 summarizes the visual data, including total dura-

Movie	Genre	Duration	Total Frames	Extracted Frames
Titanic (1997)	Romance, Drama, Epic	3:14:54	280,305	11,694
Skyfall (2012)	Action, Adventure, Spy Thriller	2:23:10	205,952	8,589
Oppenheimer (2023)	Biographical, History, Drama	3:00:22	259,472	10,822
Spider-Man 2 (2004)	Superhero, Action, Sci-Fi	2:15:48	195,360	8,148
Avatar 2 (2022)	Sci-Fi, Action, Adventure	3:12:38	277,117	11,558

Table 1: Visual data statistics for the five movies. Extracted frames are sampled within the time span of each subtitle segment.

Movie	Total Pairs	Avg Words	Avg Chars
Titanic	1,991	5.64	29.57
Skyfall	1,140	5.70	29.74
Oppenheimer	3,519	5.68	31.77
Spider-Man 2	1,049	5.83	30.64
Avatar 2	1,444	6.44	32.86
Total	**9,143**	**5.86**	**30.92**

Table 2: Subtitle statistics per movie. Total pairs represent the number of English subtitle segments.

tion, total frames, and the number of frames extracted within subtitle time spans (i.e., frames that fall within the time window of a subtitle segment). The extracted frames serve as the visual input for our multimodal methods.

2.2 Subtitle Corpora

We extract subtitles from publicly available sources[2] and temporally align them with the corresponding video segments. All subtitle pairs are preprocessed to remove noise, normalize punctuation, and filter excessively long or short segments. Table 2 provides detailed statistics for each movie, including the number of subtitle pairs (English source and target language), average English source length in words, and average English source length in characters. Parallel subtitles are available for Hindi (all movies except Avatar 2), Bengali and Telugu (all movies), Kannada (all movies except Spider-Man 2), and Tamil (all movies except Titanic). This selection ensures coverage of linguistically diverse Indian languages while respecting the availability of high-quality parallel subtitles. All movies were legally purchased as DVDs. Frame extraction for research constitutes fair use and follows standard practice in video-language benchmarks.

2.3 Data Release

To foster reproducibility and further research, we will release the curated movie-subtitle-visual alignment data for all five languages under a fair-use educational/research license. The release includes the English source, reference translations, and extracted visual descriptions. The code and instructions for reproducing the experiments will also be made publicly available.

[2]http://subtitlecat.com

3 Methodology

Our methodology is designed for real-world applicability: all models are used off-the-shelf in a zero-shot setting, no fine-tuning or training is performed, and the pipeline is fully reproducible. We use Qwen-2.5-7B-Instruct (Qwen et al., 2025) as the translation model, Llama-3.1-8B-Instruct (Meta, 2024) for summarization (Datta et al., 2025), and Apple FastVLM-0.5B (Vasu et al., 2025) for visual description extraction. The pipeline is depicted in Figure 1. For the text-only baseline, Qwen is prompted with the English source only. This yields the text-only translation against which visual-enhanced methods are compared.

3.1 Visual Context Generation

From each movie, we sample frames at 1-fps and obtain raw textual descriptions using FastVLM-0.5B. These descriptions are then summarized by Llama-3.1 into two distinct forms:

Attribute Visual Context (Attr-VC) A 5-minute sliding window (centered on the subtitle start) is aggregated and summarized into structured attributes: `[SETTING]`, `[GENDER]`, `[RELATION]`, `[HONORIFIC]`, and `[SUMMARY]`. This yields a coarse, high-level scene description.

Inter-Chunk Visual Summarization (Inter-VS) The raw descriptions that fall between the end of the previous subtitle and the start of the current subtitle (the visual gap) are summarized into a free-text description. This captures visual events that occur between dialogue turns. The full prompts used for these summarization tasks are provided in Table 6. Both the summaries are concatenated with the English source using the same prompt template, which instructs the model to ground its translation in the visual context.

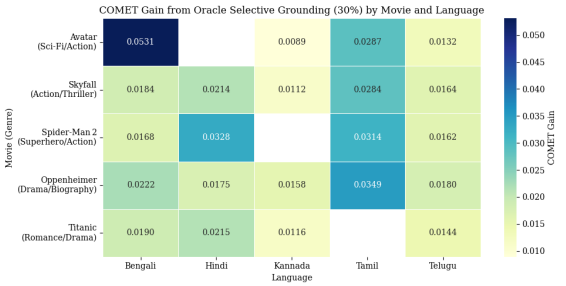

Figure 2: COMET gain from Oracle Selective Grounding (30%) by movie and language. Gain is computed as the difference between the oracle selective translation (replacing the worst 30% of baseline segments by baseline COMET) and the text-only baseline, using the better of the two visual summarisation methods for each pair. Movie names are followed by their genre in parentheses. Darker shades indicate larger gains.

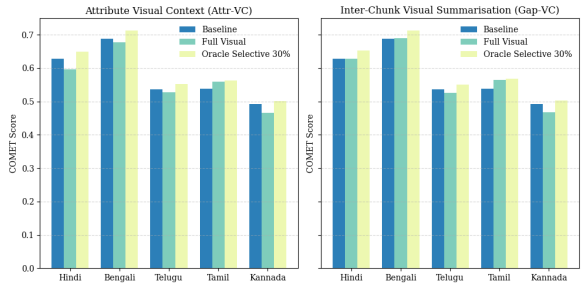

Figure 3: Language-wise COMET scores for the two visual summarization methods. For each method, bars show the baseline (text-only), full visual-enhanced translation, and oracle selective grounding (replacing the worst 30% of baseline segments by baseline COMET). The oracle selective consistently improves COMET over the baseline across all languages, with relative gains of 2-5%.

3.2 Oracle Selective Grounding

To estimate the upper bound of selective visual grounding, we compute per-segment COMET scores for the baseline translations against the reference. We then replace the worst $k\%$ of segments (by baseline COMET) with the corresponding visual-enhanced translation (from either Attr-VC or Inter-VS). We experiment with $k = 20\%$ and 30%. This oracle analysis shows the potential improvement if one could perfectly identify low-quality baseline segments; it does not require any training and represents an upper bound for practical quality-estimation systems.

4 Evaluation Results

4.1 Evaluation Setup

We evaluate on the full curated test set (all aligned subtitle segments) using corpus-level BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and COMET (Rei et al., 2020).

4.2 Results and Analysis

We compare the two visual summarisation strategies *Attribute Visual Context (Attr-VC)* and *Inter-Chunk Visual Summarization (Inter-VS)* against a text-only baseline. We perform experiments with five movies in five Indian languages (Hindi, Bengali, Telugu, Tamil, Kannada) using Qwen-2.5-7B.

Results for full per-movie, per-language are shown in Table 3. Table 4 summarizes the language-wise COMET improvements.

4.3 Overall Observations

Both summarisation methods show that applying visual context indiscriminately often degrades performance compared to the text-only baseline. For example, in several movie-language pairs (e.g., Oppenheimer Hindi, Skyfall Bengali), full VT COMET is lower than the baseline. This is directly attributable to temporal misalignment: when visual frames do not match the spoken dialogue, the model is misled. However, oracle selective grounding consistently improves COMET over the baseline in almost all cases, recovering most of the potential gain while using only 30% of the visual processing. Figure 2 visualises the per-movie, per-language COMET gain from oracle selective grounding, highlighting that action-rich movies (e.g., Skyfall) show larger improvements. Human evaluation with a small set (30 examples, each for Telugu and Hindi) confirmed that selective grounding with oracle significantly improves adequacy: average score increased from 2.9 (baseline) to 4.1 (selective) on a 1-5 scale.

4.4 Comparison of Summarization Methods

Attr-VC, which aggregates a 5-minute window into coarse attributes, proves more robust to drift than Inter-VS. In many cases (e.g., Avatar Bengali, Oppenheimer Telugu, Titanic Hindi), its selective 30% COMET surpasses the full VT of Inter-VS. The attribute-based representation, by ignoring precise timing, effectively filters out irrelevant frames. Inter-VS, while capturing finer-grained visual events, is more sensitive to misalignment; its full VT often underperforms the baseline, but selective grounding still yields gains.

Movie	Lang	Baseline			5-Minute Slide Visual Attribute						Inter-Chunk Visual Summarisation					
		BLEU	chrF++	COMET	Visual-Enhanced			Oracle Selective			Visual-Enhanced			Oracle Selective		
					BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Avatar	Ben	5.68	28.71	0.6298	6.95	27.95	**0.7014**	6.92	**29.98**	**0.6829**	**8.10**	28.24	0.7137	6.91	29.80	0.6865
Avatar	Tel	4.30	19.66	0.5257	3.28	18.23	0.5154	4.38	**19.79**	0.5390	**3.67**	18.32	0.5153	**4.67**	19.85	0.5390
Avatar	Tam	3.85	23.49	0.5352	4.08	22.94	0.5545	4.24	24.36	0.5580	**4.57**	**23.62**	**0.5613**	**4.33**	**24.50**	**0.5639**
Avatar	Kan	3.50	18.94	0.4857	2.34	15.20	0.4582	3.39	18.65	0.4933	**2.23**	**15.28**	**0.4612**	**3.33**	18.37	**0.4946**
Oppenh.	Ben	8.05	29.38	0.7026	5.47	25.09	0.6735	8.03	**29.41**	0.7248	**6.37**	26.26	0.6858	**8.08**	29.50	0.7237
Oppenh.	Hin	11.76	31.64	0.6467	8.62	27.28	0.5972	11.83	31.74	0.6642	**9.62**	**28.72**	**0.6297**	**11.86**	**32.06**	**0.6690**
Oppenh.	Tel	4.04	18.86	0.5475	3.29	17.77	0.5387	3.89	**19.13**	**0.5654**	**3.53**	18.05	0.5379	**3.96**	19.21	0.5647
Oppenh.	Tam	3.15	20.60	0.5366	3.15	21.37	0.5654	3.36	21.85	0.5630	**3.47**	**21.63**	**0.5690**	**3.66**	**22.18**	**0.5715**
Oppenh.	Kan	2.95	16.81	0.4938	2.23	14.63	0.4740	2.96	17.20	0.5066	**2.30**	14.25	0.4735	2.93	17.05	**0.5090**
Skyfall	Ben	6.31	27.30	0.6914	4.10	23.51	0.6588	**6.04**	27.39	**0.7098**	**4.55**	**23.68**	0.6612	5.93	27.14	0.7056
Skyfall	Hin	6.31	25.65	0.6026	5.74	25.28	0.5882	**6.53**	**26.68**	**0.6258**	**6.41**	25.86	0.6098	6.65	26.47	0.6245
Skyfall	Tel	2.47	17.68	0.5288	2.13	16.66	0.5248	**2.24**	**18.00**	**0.5478**	1.41	16.84	0.5157	2.16	17.86	0.5454
Skyfall	Tam	2.33	21.09	0.5350	2.22	21.11	0.5581	2.59	21.78	0.5581	**1.83**	20.89	**0.5595**	**2.66**	**22.02**	**0.5639**
Skyfall	Kan	1.59	16.88	0.4920	1.76	14.38	0.4668	1.59	16.90	0.5013	**1.54**	**14.36**	**0.4682**	1.58	16.80	**0.5038**
Spider2	Ben	9.58	26.81	0.7190	6.69	24.44	0.6902	9.10	26.97	0.7359	**8.55**	**25.77**	**0.7021**	**9.47**	**27.18**	**0.7350**
Spider2	Hin	12.33	29.15	0.6459	10.13	27.71	0.6286	12.61	**30.20**	0.6746	**12.19**	29.17	**0.6532**	**12.93**	30.32	**0.6786**
Spider2	Tel	5.22	18.47	0.5407	3.57	17.69	0.5349	5.04	18.84	0.5567	3.40	17.73	0.5342	**5.04**	**18.74**	**0.5569**
Spider2	Tam	4.12	21.65	0.5448	3.27	21.39	0.5601	4.29	22.73	0.5684	**4.01**	**22.09**	**0.5694**	**4.32**	**22.83**	**0.5761**
Titanic	Ben	9.59	25.87	0.6960	7.04	22.97	0.6616	9.55	26.11	0.7130	**8.65**	**24.97**	**0.6849**	**9.82**	**26.41**	**0.7150**
Titanic	Hin	11.98	26.59	0.6152	9.12	24.45	0.5711	11.92	27.12	0.6321	**12.29**	**26.90**	**0.6176**	**12.66**	**27.62**	**0.6367**
Titanic	Tel	5.03	17.82	0.5350	3.85	16.99	0.5211	4.92	**18.01**	0.5481	**4.32**	17.52	0.5296	**5.11**	18.21	0.5494
Titanic	Kan	4.95	17.16	0.4950	3.11	14.04	0.4670	4.73	16.97	0.5037	**3.14**	**14.45**	**0.4665**	**4.86**	**17.03**	**0.5049**

Table 3: Comparison of two visual summarization strategies. *5-Minute Slide Visual Attribute* aggregates a 5-minute sliding window into structured attributes (setting, gender, honorifics, emotion); *Inter-Chunk Visual Summarization* summarizes the visual content between dialogue turns. For each method, we report *Visual-Enhanced* (using the full visual context for all segments) and *Oracle Selective* (replacing the worst 30% of baseline segments by baseline COMET with the visual-enhanced translation). This oracle shows the upper bound of selective grounding. Metrics are corpus-level BLEU, chrF++, and COMET. **Bold** indicates the higher score between the two methods for the same condition (Visual-Enhanced or Oracle Selective).

Language	Attr-VC		Inter-VS	
	Full	Sel30	Full	Sel30
Hindi	-5.0%	+3.4%	0.0%	+3.9%
Bengali	-1.6%	+3.7%	+0.3%	+3.7%
Telugu	-1.6%	+3.0%	-1.7%	+2.9%
Tamil	+4.0%	+5.9%	+5.0%	+5.8%
Kannada	-5.1%	+2.3%	-4.9%	+2.4%

Table 4: Language-wise average COMET improvement (Δ) over baseline for each method and condition. Positive values indicate improvement. Oracle Selective replaces the worst 30% of baseline segments by baseline COMET.

4.5 Language-Wise Trends

Table 4 aggregates COMET improvement over the baseline for each language. For Attr-VC selective, all languages show positive gains (range +2.3% to +5.9%). The gains are larger for morphologically rich languages (Bengali, Tamil, Kannada) where visual cues (e.g., honorifics, emotional tone) help resolve pragmatic ambiguities. Inter-VS selective also yields consistent improvements, though slightly lower for some languages. The COMET improvements across languages are further illustrated in Figure 3.

4.6 Summary of Key Findings

Our case study yields three actionable insights:

1. **Coarse attribute-based summarization is robust to temporal drift** By aggregating visual information over a 5-minute window into structured attributes, Attr-VC achieves pragmatic gains without being misled by misaligned frames.

2. **Selective visual grounding can recover most of the gain** An oracle that replaces only the worst 20-30% of baseline segments with visual-enhanced translations consistently improves COMET over baseline, using a fraction of the visual processing.

3. **Alignment quality often outweighs architectural complexity** Fine-grained summarization methods that rely on precise temporal alignment are fragile. For real-world deployment, drift-tolerant architectures are preferred.

5 Discussion

Our case study reveals a central tension: while visual context can provide critical grounding for a small subset of segments, it is irrelevant or even harmful for the majority. This asymmetry, combined with temporal misalignment, shapes the relative performance of the two summarization strategies and the effectiveness of selective grounding.

5.1 When Visual Context Helps and When It Does Not

Only a minority of subtitle segments truly depend on visual information. Action cues ("He's coming!"), emotion-driven exchanges ("I'm so sorry"),

and references to on-screen objects ("That one") require visual grounding to resolve ambiguity. For the vast majority of conversational dialogue, text alone is sufficient, and adding visual context adds no benefit. In our dataset, we estimate that fewer than 15% of subtitles are visually grounded in this sense. This explains why even the best-performing method oracle selective grounding-achieves only a modest overall COMET gain (2-5%) while replacing only 20-30% of segments.

5.2 Why Attribute Summarization Outperforms Gap Summarization?

Attr-VC aggregates a 5-minute sliding window into high-level attributes. This coarse representation is rarely misleading for neutral segments and provides valuable pragmatic context for the minority that need it. Moreover, it is inherently robust to misalignment because it aggregates over longer time windows, effectively ignoring irrelevant frames. Inter-VS summarizes the visual content between dialogue turns. This finer-grained representation captures visual events that may be directly relevant, but it is more sensitive to drift. When visual frames are misaligned, Inter-VS can introduce misleading information, causing its full visual-enhanced translations to sometimes underperform the baseline. However, when applied selectively (only to the worst baseline segments), Inter-VS still yields substantial gains on the replaced set.

5.3 The Role of Oracle Selective Grounding

Our oracle analysis replaces the worst 20-30% of baseline segments (by baseline COMET) with the corresponding visual-enhanced translation. This shows the upper bound of what could be achieved with a perfect quality-estimation model. For both methods, selective grounding consistently lifts COMET above the baseline, often matching or even exceeding the full visual-enhanced performance of the other method.

5.4 Insights for Low-Resource Indian Languages

For morphologically rich languages such as Bengali and Kannada, the small subset of visually grounded segments often includes honorifics, implicit referents, or emotional tone areas, where text-only models are weakest. Coarse attribute summarisation helps in these cases without harming the rest, yielding clear COMET gains in the selective setting (e.g., +3.7% for Bengali, +2.3% for Kannada). This suggests that for low-resource settings, investing in reliable, drift-tolerant visual abstractions is more practical than pursuing fine-grained fusion or high-frequency visual processing.

5.5 Practical Implications for Movie Localisation

Our findings lead to three actionable recommendations for deploying multimodal subtitle translation:

1. **Be selective:** Not every subtitle needs visual context. An oracle study shows that replacing only the worst 20-30% of baseline segments can recover most of the gain. A practical system would use a quality-estimation model (e.g., based on sentence length, emotion words, or visual confidence) to trigger visual enhancement only when needed.

2. **Prefer robust architectures:** Attribute-based summarization (5-minute sliding window condensed into structured cues) offers a lightweight, drift-tolerant solution suitable for production pipelines.

3. **Fix alignment before using fine-grained context:** If a more detailed visual context is desired (e.g., gap-based summarization), pre-processing steps such as dynamic time warping (DTW) or audio-visual synchronisation are essential to avoid performance degradation.

6 Conclusion

In this paper, we compare two summarization strategies for integrating visual context into subtitle translation for five low-resource Indian languages. We find that temporal misalignment- a common real-world issue- causes full visual-enhanced translation to often underperform the text-only baseline. However, an oracle selective grounding that replaces only the worst 20-30% of baseline segments with visual-enhanced translations consistently improves semantic adequacy (COMET) across all the languages, recovering most of the potential gain while using a fraction of the visual processing. Coarse attribute-based summarization proves particularly robust to drift, capturing emotional tone and scene-level cues that text alone cannot convey. Our results underscore that alignment quality often outweighs architectural complexity and that selective visual grounding offers a practical path to efficient, deployable multimodal subtitle translation.

Limitations

Our study is limited to five movies and five languages; the proportion of visually grounded segments may vary by genre and across different types of audiovisual content. The oracle selective grounding demonstrates an upper bound; future work should develop automatic quality-estimation models that can identify low-quality baseline segments without reference translations, enabling practical selective grounding. Human evaluation of pragmatic adequacy (e.g., honorifics, emotional tone) would complement automatic metrics to better capture the subtle benefits of selective visual grounding. Additionally, exploring more sophisticated alignment techniques (e.g., dynamic time warping) could further reduce temporal misalignment and improve the robustness of fine-grained visual context.

Acknowledgements

The authors would like to express their sincere gratitude to the project, Centre of Indian Language Data (COIL-D) under Bhashini, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India for its generous support.

References

Appicharla, Ramakrishna, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July. Association for Computational Linguistics.

Datta, Debtanu, Shounak Paul, Kshetrimayum Boynao Singh, Sandeep Kumar, Abhinav Joshi, Shivani Mishra, Sarika Jain, Asif Ekbal, Pawan Goyal, Ashutosh Modi, and Saptarshi Ghosh. 2025. Findings of the JUST-NLP 2025 shared task on summarization of Indian court judgments. In *Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025)*, pages 5–11, Mumbai, India, December. Association for Computational Linguistics.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig. 2025. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 28th edition.

Elliott, Desmond, 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.

Gain, Baban, Dibyanayan Bandyopadhyay, Samrat Mukherjee, Chandranath Adak, and Asif Ekbal. 2025. Impact of visual context on noisy multimodal nmt: An empirical study for english to indian languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(8), August.

Kumar, Deepak, Baban Gain, Kshetrimayum Boynao Singh, and Asif Ekbal. 2025. Does vision still help? multimodal translation with CLIP-based image selection. In Nakazawa, Toshiaki and Isao Goto, editors, *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 115–123, Mumbai, India, December. Association for Computational Linguistics.

Meta. 2024. Llama 3.1: The llama 3.1 collection of multilingual large language models. `https://huggingface.co/meta-llama/Llama-3.1-8B`, July. Model release date: July 23, 2024. Accessed: 2026-03-27.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Radinger, Anke. 2025. Subtitling in audiovisual translation studies. In *Researching Subtitling Processes*, pages 29–53. Springer.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online, November. Association for Computational Linguistics.

Shen, Zhiyu, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025. CoE: A clue of emotion framework for emotion recognition in conversations. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23548–23563, Vienna, Austria, July. Association for Computational Linguistics.

Singh, Kshetrimayum Boynao, Asif Ekbal, and Partha Pakray. 2025a. Evaluating IndicTrans2 and ByT5 for English–Santali machine translation using the ol chiki script. In Shukla, Ankita, Sandeep Kumar, Amrit Singh Bedi, and Tanmoy Chakraborty, editors, *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pages 95–100, Mumbai, India, December. Association for Computational Linguistics.

Singh, Kshetrimayum Boynao, Deepak Kumar, and Asif Ekbal. 2025b. Evaluation of LLM for English to Hindi legal domain machine translation systems. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 823–833, Suzhou, China, November. Association for Computational Linguistics.

Vasu, Pavan Kumar Anasosalu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. 2025. Fastvlm: Efficient vision encoding for vision language models.

Wang, Yuqing and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand, August. Association for Computational Linguistics.

A Appendix

A.1 Ethical Considerations

Multimodal systems may inherit and amplify biases present in movies, where visual cues can reinforce stereotypes. Temporal misalignment between audio and visuals may also lead to inappropriate translations of culturally sensitive content. For Indian languages, honorifics and regional variations require careful handling to ensure accuracy. We advocate for human oversight, transparent reporting, and evaluation frameworks that assess cultural appropriateness. As selective grounding shows that only some segments need visual context, applying visual enhancement through a bias-aware filter can help reduce potential harms.

A.2 Visual Feature Extraction

1. The original `.mkv` video is compressed to $\approx$ 1-GB and converted to `.mp4`.

2. Frames are extracted at 1-fps and passed to Apple FastVLM-0.5B to obtain a raw textual description per frame.

3. Raw descriptions are cleaned (removing repeated phrases, normalising punctuation) to produce a final description per frame.

For each subtitle, the visual context is formed by concatenating the cleaned descriptions of all frames whose timestamps fall within the subtitle's time window. When multiple frames belong to the same segment, the descriptions are aggregated into a single summary (for the gap-based method, the window is the interval between the previous subtitle's end and the current subtitle's start).

A.3 Context-Aware Translation Pipeline

The translation pipeline uses the same Qwen-2.5-7B-Instruct model for both baseline and visual-enhanced translations. The only difference is the input prompt. For the baseline, the model receives only the English source. For visual-enhanced, we prepend the summarised visual context (Attr-VC or Inter-VS) to the source using the visual-enhanced prompt template shown in Table 5. The prompt instructs the model to ground its translation in the visual scene, paying attention to gender, honorifics, and emotional tone. The same greedy decoding parameters described in hyperparameterss are used for all runs.

A.4 Hyperparameters

All inference runs use greedy decoding to ensure reproducibility.

For the translation model (Qwen-2.5-7B-Instruct):

- `max_new_tokens = 100`
- `do_sample = False`
- `repetition_penalty = 1.1`
- temperature and top-p left at default (1.0 and 1.0, effectively greedy).

For the summarization model (Llama-3.1-8B-Instruct):

- `max_new_tokens = 256`
- `do_sample = False`
- temperature and top-p at default (1.0 and 1.0).

A.5 Oracle Selective Grounding

Per-segment COMET scores for the baseline translation (against the reference) are computed using the `Unbabel/wmt22-comet-da` model. The worst $k\%$ of segments (by baseline COMET) are replaced with the corresponding visual-enhanced translation (from either Attr-VC or Inter-VS). We report results for $k = 20\%$ and 30%. Appendix 8

B Additional Analysis

B.1 Window Size Considerations

The choice of a 5-minute sliding window for Attribute Visual Context was guided by the need to capture enough local scene context while remaining robust to temporal drift. A larger window (e.g., 10 minutes) would aggregate more visual information, but it could also include more irrelevant frames, potentially increasing the risk of hallucination when the visual context does not align with the dialogue. A smaller window (e.g., 2 minutes) would be more sensitive to misalignment. The 5-minute window offers a reasonable trade-off, as evidenced by the improvement in COMET over the baseline for most language-movie pairs.

B.2 Oracle Selective Grounding

The full oracle selective results for both summarization methods are presented in Table 8. Replacing only the worst 20-30% of baseline segments consistently lifts COMET above the baseline, demonstrating that most of the gain can be achieved with a fraction of the visual processing.

C Additional Analysis

The 5-minute sliding window offers a reasonable trade-off between context and robustness; a larger window would risk including irrelevant frames, a smaller window would be more sensitive to misalignment. Oracle selective grounding (Table 8) shows that replacing only the worst 20-30% of baseline segments recovers most of the gain with minimal visual processing. Fine-grained fusion methods (visual prefixing and cross-attention fusion) failed under misalignment (e.g., "He is very kind" $\rightarrow$ "He drives fast") because they assume perfect alignment, whereas coarse attribute summarization ignores irrelevant frames. This reinforces the idea that alignment quality often outweighs architectural complexity, and that selective grounding offers a practical path to efficient visual-guided translation.

C.1 Clarifications

Movie subtitle translation is inherently difficult (fragmented speech, cultural references, zero-shot domain adaptation); our baseline scores reflect this challenge, and our contribution lies in the *relative* COMET gain, not absolute SOTA. We use Qwen-2.5-7B-Instruct because it supports zero-shot instruction following without fine-tuning, unlike dedicated Indic models (e.g., IndicTrans2) that are optimised for sentence-level translation and do not accept multi-field visual context prompts. FastVLM-0.5B prioritises efficiency (85× faster than LLaVA); using a smaller VLM makes gains harder to achieve, so our positive results demonstrate robustness. FastVLM outputs raw descriptions; we summarise them with Llama 3.1 to obtain structured attributes or free-text gap summaries, because direct prompting of FastVLM for structured output is not feasible. Our experiments compare visual-enhanced translation against an identical text-only baseline, isolating the effect of visual context; comparing across different MT systems would confound architectural differences.

C.2 Evaluation Metrics

We report corpus-level BLEU (Papineni et al., 2002) using SacreBLEU and chrF++ (Popović, 2015) with `word_order=2`. COMET (Rei et al., 2020) is computed with the `wmt22-comet-da` model using default settings.

C.3 Data and Code Availability

To foster reproducibility, the curated movie-subtitle-visual alignment data for all five languages will be released under a fair-use educational/research license. The release includes English sources, reference translations, and extracted visual descriptions. All the codes and datasets used for extraction, summarization, translation, and evaluation are available at GitHub [3] and our group page.[4]

[3] `https://github.com/Tarunc224/visually-guided-subtitle-translation`

[4] https://ai-nlp-ml.github.io/resources.html

Baseline Prompt (text-only)

```
<|im_start|>system
You are a translation expert. Translate dialogue from English to {target_language}.
RULES:
- Provide ONLY the translated {target_language} dialogue.
- DO NOT include explanations, or English text.
<|im_end|>
<|im_start|>user
[SOURCE]: "{row['english_dialogue']}"
[TASK]: Translate to {target_language} dialogue.
<|im_end|>
<|im_start|>assistant
```

Visual-Enhanced Prompt

```
<|im_start|>system
You are a cinematic multimodal translator specializing in English-to-{target_language}.
Your goal is to provide a "grounded translation" where the choice of words depends on the visual scene.

RULES:
1. GENDER: Use the Visual Context to identify speaker/listener gender.
2. HONORIFICS: Determine social hierarchy from the scene (Formal vs. Informal).
3. LOOSE MEANING: Prioritize emotional intent and natural {target_language} flow.
4. Output ONLY the translated {target_language} dialogue text. No names, no English.
<|im_end|>
<|im_start|>user
[VISUAL CONTEXT]: {row['visual_context']}
[ENGLISH SOURCE]: "{row['english_source']}"
[TASK]: Based on the visual scene, provide the most natural {target_language} translation.
<|im_end|>
<|im_start|>assistant
```

Table 5: Comparison of baseline and visual-enhanced prompts.

Summarisation Method	Prompt Template

Attribute Visual Context (Attr-VC)	`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`
	`Identify these cinematic attributes to guide {target_language} translation:`
	`[SETTING]: (e.g., Formal, Public, Intimate)`
	`[GENDER]: (Speaker/Listener gender)`
	`[RELATION]: (e.g., Stranger, Family, Hostile)`
	`[HONORIFIC]: (language-specific, e.g., APNI/TUMI for Bengali)`
	`[SUMMARY]: (One sentence factual summary with emotional intent)`
	`Output ONLY these tags.<|eot_id|><|start_header_id|>user<|end_header_id|>`
	`Visual Data: {sample[:3000]}<|eot_id|>`
	`<|start_header_id|>assistant<|end_header_id|>`

Inter-Chunk Visual Summarisation (Inter-VS)	`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`
	`You are a movie analyzer. Summarize the following visual descriptions`
	`from {start_sec}s to {end_sec}s of the movie into 2-3 sentences.`
	`Focus ONLY on the current location and character actions.`
	`Do not use introductory filler.`
	`<|eot_id|><|start_header_id|>user<|end_header_id|>`
	`Visual Data: {text_blob[:2500]}`
	`<|eot_id|><|start_header_id|>assistant<|end_header_id|>`

Table 6: Prompt templates used for visual summarization. Both the prompts are given to Llama-3.1-8B-Instruct. The attribute prompt outputs structured tags; the inter-chunk prompt outputs a free-text sentence. Placeholders in braces are replaced with actual data at runtime.

Scenario	Source	Baseline (literal)	Visual-Enhanced	Reference	Explanation
			Visual context helps		
Emotion	"I'm so sorry."	"I am sorry."	"I am truly sorry."	"I am truly sorry."	Facial expression conveys deeper remorse, captured by visual summary.
Action	"He's coming!"	"He is coming."	"The man is coming!"	"The man is coming!"	Visual identifies gender and urgency.
Honorific	"Please sit."	"Sit."	"Please sit (formal)."	"Please sit (formal)."	Formal setting triggers correct honorific.
			Visual context backfires (temporal misalignment)		
Misalign	"He is very kind."	"He is very kind."	"He drives fast."	"He is very kind."	Car chase frame (drift) causes hallucination.
Misalign	"I'm flying."	"I'm flying."	"I'm swimming."	"I'm flying."	Calm ocean scene (drift) misleads.
			Visual context backfires (irrelevant visuals)		
Neutral	"How was your day?"	"How was your day?"	"How was your day?"	"How was your day?"	No improvement; adds processing overhead.

Table 7: When visual context helps (emotion, action, honorifics) and when it backfires (temporal misalignment, irrelevant visuals). Visual-enhanced outputs are from the better of the two summarization methods for each case. The misalignment examples are drawn from actual data where cumulative drift paired a kind remark with a car chase, and a flying statement with a calm ocean scene.

Movie	Language	Baseline COMET	5-Minute Slide Visual Attribute		Inter-Chunk Visual Summarisation	
			Oracle selective 20%	Oracle selective 30%	Oracle selective 20%	Oracle selective 30%
Avatar	Bengali	0.6298	0.6682	0.6829	0.6709	0.6865
Avatar	Telugu	0.5257	0.5354	0.5390	0.5361	0.5390
Avatar	Tamil	0.5352	0.5580	0.5650	0.5569	0.5639
Avatar	Kannada	0.4857	0.4933	0.4943	0.4928	0.4946
Oppenhimer	Bengali	0.7026	0.7224	0.7248	0.7210	0.7237
Oppenhimer	Hindi	0.6467	0.6617	0.6642	0.6643	0.6690
Oppenhimer	Telugu	0.5475	0.5604	0.5654	0.5600	0.5647
Oppenhimer	Tamil	0.5366	0.5630	0.5715	0.5639	0.5715
Oppenhimer	Kannada	0.4938	0.5066	0.5096	0.5065	0.5090
Skyfall	Bengali	0.6914	0.7064	0.7098	0.7044	0.7056
Skyfall	Hindi	0.6026	0.6223	0.6258	0.6216	0.6245
Skyfall	Telugu	0.5288	0.5430	0.5478	0.5415	0.5454
Skyfall	Tamil	0.5350	0.5581	0.5659	0.5561	0.5639
Skyfall	Kannada	0.4920	0.5013	0.5038	0.5014	0.5038
Spider 2	Bengali	0.7190	0.7337	0.7359	0.7305	0.7350
Spider 2	Hindi	0.6459	0.6684	0.6746	0.6717	0.6786
Spider 2	Telugu	0.5407	0.5527	0.5567	0.5527	0.5569
Spider 2	Tamil	0.5448	0.5684	0.5753	0.5700	0.5761
Titanic	Bengali	0.6960	0.7114	0.7130	0.7120	0.7150
Titanic	Hindi	0.6152	0.6293	0.6321	0.6320	0.6367
Titanic	Telugu	0.5350	0.5456	0.5481	0.5465	0.5494
Titanic	Kannada	0.4950	0.5037	0.5065	0.5049	0.5063

Table 8: Oracle selective COMET scores for the two summarization methods. "Oracle selective 20%" and "Oracle selective 30%" replace the worst 20% and 30% of baseline segments (by baseline COMET) with the corresponding visual-enhanced translation. The full visual-enhanced results are reported in Table 3.