OmniFusion: Simultaneous Multilingual and Multimodal Translations via Modular Fusion

Anonymous ACL submission

Abstract

Open-source text-only translation Large Language Models (LLMs) are rapidly improving in multilingual coverage and translation quality, but their unimodal nature limits their use in multimodal translation. In speech translation (ST), they must operate in cascaded pipelines using automatic speech recognition followed by translation introducing additional latency, especially for simultaneous ST (SimulST). In both speech and caption translation, they cannot exploit visual context for disambiguation. In contrast, pretrained multimodal foundation models (MMFMs) offer strong cross-modal perception and reasoning but lack the multilingual depth and translation performance of specialized translation LLMs. This mismatch motivates combining the strengths of both model types. We propose OmniFusion, an end-to-end framework that fuses MMFMs with translation LLMs via a novel multi-layer hidden-state fusion strategy, enabling joint training while remaining data and resource efficient. Built with Omni 2.5–7B as the MMFM and SeedX PPO–7B as the translation LLM, OmniFusion handles speech-to-text, speech-and-image-to-text, and text-and-image-to-text translation within a single architecture. Experiments show that OmniFusion effectively leverages audio and visual cues, reduces SimulST latency by one second compared to cascaded pipelines, and improves overall translation quality.[1]

1 Introduction

In recent years, research in natural language processing (NLP) has increasingly shifted from unimodal foundation models (Touvron et al., 2023; Grattafiori et al., 2024) to multimodal foundation models (MMFMs) (Abdin et al., 2024; Goel et al., 2025; Xu et al., 2025), as many tasks benefit from contextual information derived from multiple sources. Although unimodal systems can perform strongly in certain settings, further progress also requires models to be able to take advantage of additional contextual cues without degrading quality.

Translation is a representative task that can benefit from such multimodal integration. In Speech Translation (ST), the goal is to translate spoken input in a source language into a target language (Ahmad et al., 2024; Agostinelli et al., 2025). While recent advances have led to substantial improvements, the use of multimodal context such as visual information from images or presentation slides accompanying speech remains relatively underexplored (Liu and Niehues, 2024; Gaido et al., 2024a; Sinhamahapatra and Niehues, 2025). Another relevant scenario is caption translation, where the textual input is typically short and closely tied to the image, making visual information particularly useful for disambiguation and grounding.

Although MMFMs can process multimodal inputs inputs, their multilingual translation capabilities are generally weaker than those of specialized translation large language models (LLMs) (Alves et al., 2024; Cheng et al., 2025). Consequently, directly applying MMFMs often results in inferior translation quality and limited language coverage. In the context of ST, A practical solution is to adopt a cascaded pipeline, performing automatic speech recognition (ASR) followed by machine translation (MT) (Ahmad et al., 2024; Agostinelli et al., 2025; Koneru et al., 2025). While this modular design allows independent optimization, it prevents the translation model from leveraging multimodal context and introduces additional latency compared to end to end (E2E) approaches. This drawback is particularly critical for simultaneous ST (SimulST) settings (Agostinelli et al., 2025; Papi et al., 2025a).

These limitations motivate the development of an E2E multilingual and multimodal translation system that can support multiple use cases, including ST, image-aware ST, and caption translation. A unified model can further benefit from shared

[1] Code is attatched as zip file.

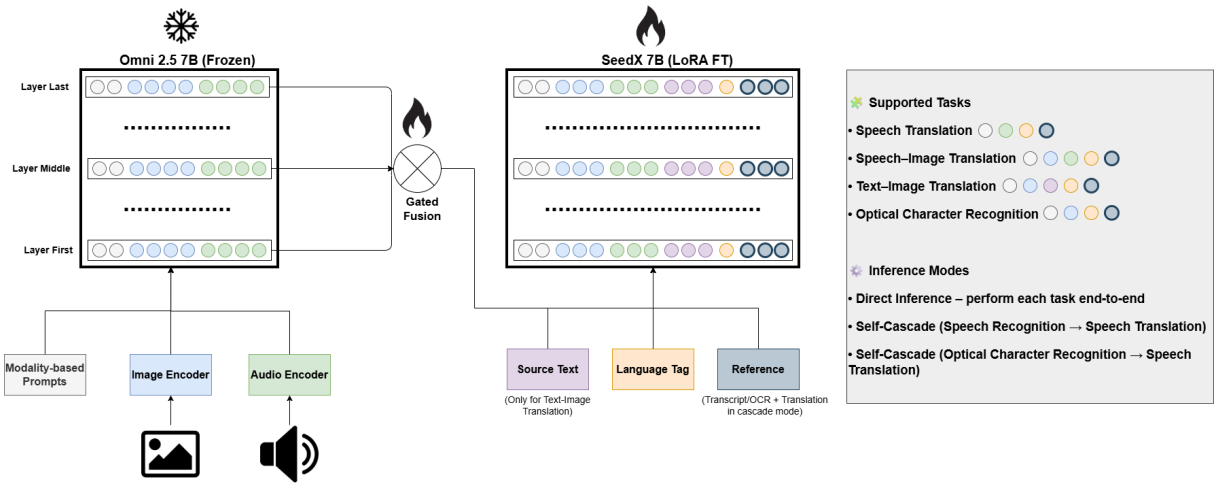


Figure 1: Architecture of OmniFusion with supported tasks and inference modes.

data across tasks and simplify deployment. One promising direction is to extend the capabilities of a translation LLM to handle additional modalities. However, only a limited number of prior works have explored incorporating audio or image inputs into multilingual LLMs (Ambilduke et al., 2025; Viveiros et al., 2025). These approaches typically rely on training from scratch or on using only pretrained audio or vision encoders, without fully exploiting the higher-level representational capabilities of existing MMFMs. As a result, training is often inefficient, since the base LLM must learn modality alignment largely from scratch.

In this work, we propose to leverage pretrained MMFMs to introduce multimodal capabilities into a multilingual LLM within a unified E2E architecture for ST, image-aware ST, and caption translation. To this end, we present a novel architecture that tightly couples the two models during training. In addition, we adopt a multi-task training objective, augmented with auxiliary tasks such as Optical Character Recognition (OCR) and self-cascading, to further improve multimodal alignment and translation performance.

Our main contributions are summarized as follows:

- We introduce **OmniFusion**, an E2E multimodal and multilingual translation model that supports translation across combinations of audio, image, and text modalities.
- We propose a gated fusion architecture coupled with a multi-task training framework that integrates multimodal representations from pretrained MMFMs into a translation LLM, enabling effective reuse of perceptual and interpretive knowledge and improved multimodal alignment.
- We empirically demonstrate that **OmniFusion** reduces inference latency by approximately 1 second compared to cascaded systems in SimulST (Figure 2), lowers major and critical error rates for offline ST (Table 2), and achieves state of the art performance on the CoMMuTE benchmark (Table 3).

2 Approach

We propose to fuse a MMFM with a task-specific translation LLM to leverage their complementary strengths and enable E2E training for improved latency. Furthermore, we aim to incorporate not only the perception capabilities of the MMFM but also its multimodal reasoning and understanding.

First, we describe the overall architecture and training process of our approach, considering a general input and output sequence independent of the specific task. We formally define the information flow, training loss, and the fusion of MMFM layers. Next, we detail the training tasks used to align multimodal representations from both image and audio modalities to the translation LLM. Finally, we discuss the training prompts that enable a self-cascading inference mode, allowing the model to enhance translation quality at the cost of increased latency. An overview of the architecture and supported tasks is shown in Figure 1.

2.1 Architecture

Let us denote the specialized translation LLM as $\mathcal{M}_{\text{NMT}}$ and the MMFM as $\mathcal{M}_{\text{MMFM}}$, where both

use a language model as their backbone. The $\mathcal{M}_{\text{NMT}}$ is trained for translation via supervised fine-tuning with the following loss function:

$$\mathcal{L}_{\text{Trans}} = -\frac{1}{T_y}\sum_{t=1}^{T_y} \log P_\theta\big(y_t \mid y_{<t},\, X\big), \qquad (1)$$
$$X = (x_1, \ldots, x_{L_x}), \quad y = (y_1, \ldots, y_{T_y})$$

where $X = (x_1, \ldots, x_{L_x})$ is the sequence of source tokens of length L_x, and $y = (y_1, \ldots, y_{T_y})$ is the sequence of target tokens of length T_y.

For $\mathcal{M}_{\text{MMFM}}$, the image and audio modalities are first processed by modality-specific encoders and then projected to match the embedding dimension of $\mathcal{M}_{\text{MMFM}}$, denoted as D_{MMFM}. This projection allows the encoder outputs to be concatenated as a sequence of tokens, enabling the MMFM to operate similarly to a standard LLM, such as $\mathcal{M}_{\text{NMT}}$.

To fuse both the perception and higher-level understanding of the MMFM, we hypothesize that lower layers predominantly capture perceptual features, while higher layers encode more abstract reasoning as computation progresses. Accordingly, we extract the first, middle, and last hidden layers from the MMFM, denoted as:

$$\begin{aligned} h_t^{(1)} &= \text{MMFM}_{\text{layer 1}}(x_t), \\ h_t^{(\text{mid})} &= \text{MMFM}_{\text{layer mid}}(x_t), \\ h_t^{(\text{last})} &= \text{MMFM}_{\text{layer last}}(x_t), \\ h_t^{(1)},\, h_t^{(\text{mid})}, h_t^{(\text{last})} &\in \mathbb{R}^{L_m \times D_{\text{MMFM}}}, t = 1, \ldots, L_m \end{aligned} \qquad (2)$$

where L_m is the length of the MMFM hidden states.

Directly concatenating these multi-dimensional representations into $\mathcal{M}_{\text{NMT}}$ poses three challenges. First, the sequence length triples, significantly increasing computational and memory requirements. Second, the relevance of each layer may vary depending on the task, necessitating selective weighting. Third, the embedding dimension of $\mathcal{M}_{\text{MMFM}}$ may not match that of $\mathcal{M}_{\text{NMT}}$.

To overcome these challenges, we introduce a gated fusion layer that determines the contribution of each hidden state along the sequence length. Each hidden state is multiplied by its corresponding gate values, summed across layers, and then projected through an MLP to align the embedding dimension with that of the translation model:

$$\begin{aligned} W_{\text{gate}} &\in \mathbb{R}^{3\times 3D_{\text{MMFM}}}, \\ [g_1, g_{\text{mid}}, g_{\text{last}}]_t &= \text{softmax}\big(W_{\text{gate}} \cdot h_t^{(1)} \oplus h_t^{(\text{mid})} \oplus h_t^{(\text{last})}\big), \\ m_t' &= g_1 \odot h_t^{(1)} + g_{\text{mid}} \odot h_t^{(\text{mid})} + g_{\text{last}} \odot h_t^{(\text{last})}, \\ m_t &= \text{MLP}(m_t') \in \mathbb{R}^{L_m \times D_{\text{Trans}}}, \quad t = 1, \ldots, L_x \end{aligned} \qquad (3)$$

where D_{Trans} is the embedding dimension of $\mathcal{M}_{\text{NMT}}$ and m_t is the fused multimodal representation at time step t.

The fused multimodal embeddings m_t can now be treated like additional token embeddings and concatenated with the $\mathcal{M}_{NMT}$ token embeddings:

$$\begin{aligned} \tilde{x}_t &= m_t \oplus x_t, \\ \tilde{X} &= (\tilde{x}_1, \ldots, \tilde{x}_{L_m+L_x}) \end{aligned} \qquad (4)$$

Finally, the multimodal translation loss is computed by conditioning on the concatenated sequence $\tilde{X}$ while predicting the target tokens:

$$\mathcal{L}_{\text{MM-Trans}} = -\frac{1}{T_y}\sum_{t=1}^{T_y} \log P_\theta\big(y_t \mid y_{<t},\, \tilde{X}\big) \qquad (5)$$

Where $\tilde{X}$ is the multimodal input to the translation model.

2.2 Training

We have shown how internal representations from MMFMs can be extracted and integrated into a downstream translation model. However, effectively leveraging these representations also requires a training strategy that explicitly aligns multimodal features, enabling the LLM to exploit them during generation. Importantly, such a strategy should avoid massive pretraining or extensive fine-tuning, and remain feasible in modest computing environments such as university clusters.

In this work, our goal is to build a translation model that can leverage both audio and image context while remaining compute- and data-efficient. To this end, we adopt two complementary training strategies: multi-task learning and self-cascading.

2.2.1 Multi-task Learning

We train the model on three primary multimodal translation tasks:

- **Speech Translation (ST):** Generate a target-language translation given only source-language audio.

- **Speech–Image Translation (SIT):** Generate a target-language translation given source-language audio together with an image providing additional context, such as a presentation slide aligned with the spoken segment.

- **Text–Image Translation (TIT):** Generate a target-language translation given source text and an aligned image, as in caption translation scenarios.

Jointly training on these tasks encourages the model to handle different modality combinations and promotes shared representations across translation settings.

2.2.2 Self-Cascading (ASR and OCR)

In addition to multi-task learning, we adopt a self-cascading training strategy that decomposes translation into intermediate generation steps. For instance, English to German ST is trained by first generating the ASR transcription and then continuing generation to produce the final translation conditioned on the ASR hypothesis. This decomposition simplifies the learning problem and provides explicit alignment signals between audio representations and the LLMs semantic space.

We extend self-cascading to visual inputs by treating OCR as an intermediate step before translation. In visually grounded scenarios (e.g., slide-based ST), the model first generates OCR text and then produces the translation conditioned on this output, providing strong alignment signals between visual representations and text.

During training, we stochastically apply self-cascading based on task-specific probabilities and modality paths, and use modality-specific prompts to handle different input combinations (Algorithm 1). Finally, we freeze the MMFM and fine-tune the translation LLM with LoRA (Hu et al., 2021), substantially reducing memory and compute costs while preserving pretrained multimodal capabilities.

3 Experimental Setup

3.1 Training Data

We aim to train the fused model to perform translation tasks across both speech and image modalities. An overview of the data sources is provided in Table 1.

For ST, we use subsets of Europarl-ST (Iranzo-Sánchez et al., 2020) and Covost-2 (Wang et al., 2020), sampling across multiple target languages. We limit training data to these subsets due to computational constraints and to demonstrate that full datasets are not strictly necessary.

To enable the model to exploit images when translating audio, we use the M3AV corpus (Chen et al., 2024), which contains English scientific talks paired with corresponding slides. The slides also include Paddle-OCR (Cui et al., 2025) predictions, allowing the model to be trained either to predict translations directly or via an intermediate OCR step for better alignment.

Since M3AV has limited language coverage, we augment the data by synthetically translating ASR transcripts into additional target languages (**M3AV-Aug**). This enables the model to handle ST with slides and supports the self-cascade mode, performing ASR first and then translation.

Finally, for TIT, we employ the Multi30k dataset (Elliott et al., 2016), which contains images paired with captions in multiple languages. This task encourages the model to go beyond OCR and to leverage visual content, including objects and scene understanding, for translation.

3.2 Models

Several MMFMs have been developed rapidly in recent years, showing significant improvements in both quality and efficiency (Abdin et al., 2024; Team et al., 2025). In our scenario, we require a model capable of processing both audio and image modalities, while also supporting cross-modal abilities. Accordingly, we select Qwen Omni 2.5B[2] (Xu et al., 2025), due to its strong multimodal capabilities and the potential to incorporate video in future experiments.

For the translation LLM, we choose the recently proposed SeedX PPO 7B[3] (Cheng et al., 2025), which provides broad language coverage and strong translation performance compared to closed-source alternatives.

We use the training data mentioned above and denote the fine-tuned model as **OmniFusion** for the rest of the paper. While we report results using these specific models, it is important to note that our approach is not limited to these base LLMs. The training hyper-parameters are specified in Table 11. For validation, we used Covost-2 validation data and select the checkpoint with the least loss in 20k steps.

[2] Qwen/Qwen2 Audio.5-Omni-7B
[3] ByteDance-Seed/Seed-X-PPO-7B

Dataset	# Examples	Language	Modality	Task
M3AV	122k	*en*	Speech, Image	ASR, OCR
M3AV - Aug	122k	*zh, fr, de, it, ja, ko, pt, ru, es, vi, ar, cs, hr, da, nl, fi, hu, id, ms, nb, no, pl, ro, tr*	Speech, Image	ST, OCR
Europarl	260k*	*de, es, fr, it, nl, pl, pt, ro*	Speech	ST
Covost2		*de, ja, tr, zh*	Speech	ST
Coco-Captions	87k	*cs, de, fr*	Image	TIT

Table 1: Overview of data sources collected for E2E training. * We sample together 260k from the concatenated data uniformly across the language pairs. All the speech and text data are only available with English as the source language. Language code mapping is presented in Table 10.

3.3 Evaluation: Data & Metrics

Evaluating multimodal SimulST requires aligned audio–visual translation data. We exclude M3AV from testing due to overlap with training data and its monolingual nature, and instead assess robustness on unseen domains. We therefore use the MCIF test set (Papi et al., 2025b), which contains audio–video recordings of ACL talks translated from English into three target languages. We report latency using Average Lagging (Ma et al., 2020) and translation quality on this dataset.

For TIT, we focus on cases where visual context is necessary for disambiguation, using the CoMMuTE dataset (Futeral et al., 2023, 2025), a targeted multilingual TIT benchmark.

Translation quality is measured with XCOMET-XL (Guerreiro et al., 2024) for ST, enabling fine-grained error analysis, and with COMET[4] (Rei et al., 2022) for TIT using aligned references.

4 Results

Our primary goal is to develop a high-quality, low-latency E2E translation system for SimulST that effectively leverages image context. We first evaluate latency–quality trade-offs in simultaneous settings to validate improvements over cascaded systems. We then report offline translation quality and conduct an error analysis comparing E2E and cascaded models fine-tuned on the same data. Next, we assess performance in TIT scenarios, benchmarking against off-the-shelf multimodal translation models. Finally, we present an ablation study to analyze the contribution of individual MMFM layers. Additional ablation study on training tasks is in Section A.1 due to space constraints.

We further validate the proposed approach with an alternative model pairing (Qwen-Audio (Chu et al., 2023) and Tower-Instruct (Alves et al., 2024)). Additional analyses on layer selection and post-editing are reported in Appendix Tables 5 and 6.

4.1 Simultaneous Speech Translation

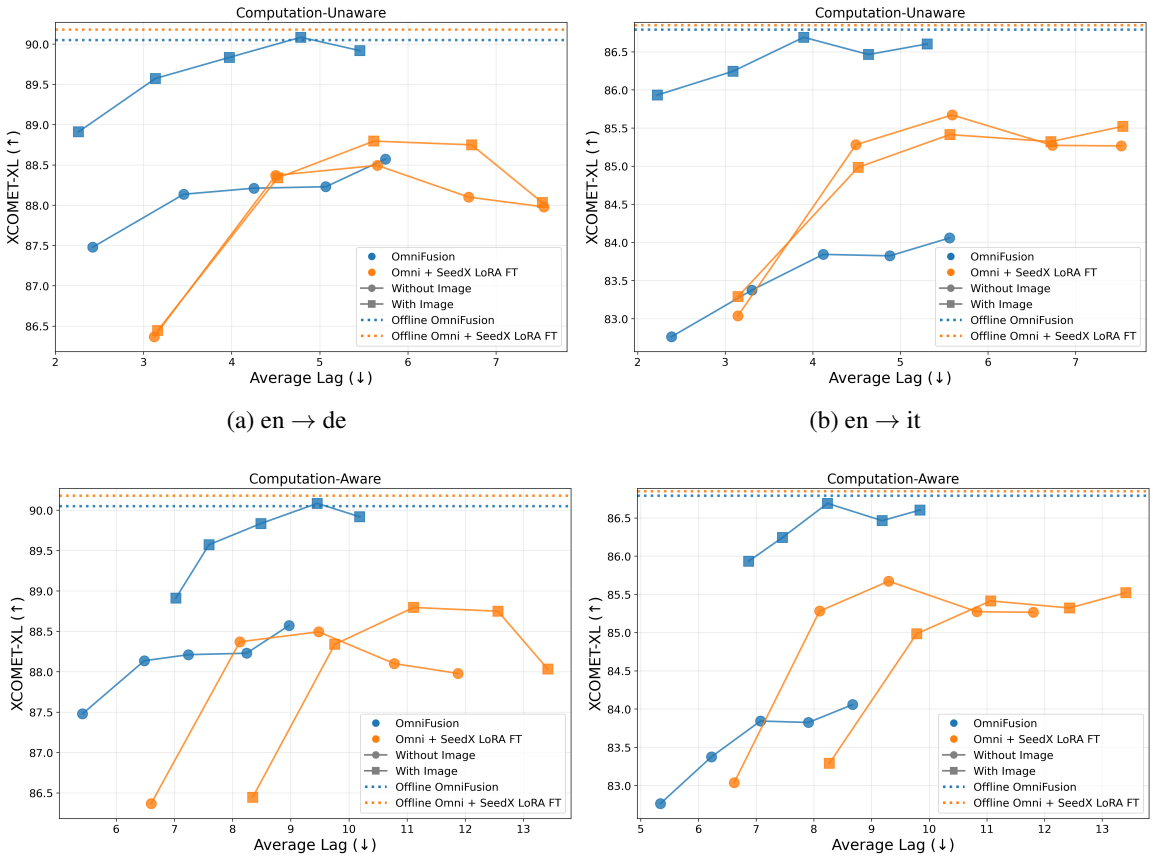
A SimulST system must translate incrementally as audio arrives, requiring a *policy* to decide when to read or commit output. As our E2E model does not natively support simultaneous decoding, we adopt the Local Agreement policy (Liu et al., 2020), which commits the longest common prefix between consecutive hypotheses. Audio is processed in fixed-size chunks, and committed tokens are force-decoded in subsequent steps.

For a fair comparison, we evaluate OmniFusion against a strong cascaded baseline built from the same pretrained components and fine-tuned on the same data using LoRA adapters (Omni + SeedX). We experiment with chunk sizes of $[1, 1.5, 2, 2.5, 3]$ seconds, with and without images, and report computation-aware and -unaware Average Lagging together with XCOMET-XL scores for $en \rightarrow de$ and $en \rightarrow it$ in Figure 2.

Offline ST with images yields the highest quality, as expected. In simultaneous settings, OmniFusion is consistently about 1 second faster than the cascaded system across all chunk sizes, with comparable or slightly lower quality in the audio-only condition. When images are provided, OmniFusion achieves both the lowest latency and the highest quality, indicating that visual context is more effectively exploited in the E2E model, enabling earlier and more stable commitments. In computation-aware settings, image processing introduces additional delay, highlighting a quality–latency trade-off.

We attribute OmniFusion's advantage to (i) our fine-tuning strategy, including an OCR objective that tightly aligns audio and visual representations, and (ii) the efficiency of the E2E architecture, which avoids the delayed translation onset inherent to cascaded pipelines. Overall, OmniFusion delivers superior latency–quality trade-offs for multimodal SimulST, making it well suited for real-world scenarios such as live presentations and meetings.

[4] Unbabel/wmt22-comet-da



(a) en → de (b) en → it

(c) en → de (d) en → it

Figure 2: SimulST performance on en→de and en→it with chunk sizes of $[1, 1.5, 2, 2.5, 3]$ seconds. Average Lag (AL, lower is better) vs. XCOMET-XL quality score (higher is better). Results compare OmniFusion (E2E) against the fine-tuned cascaded baseline and the offline upper bound (**with Image**). Top figures are computation-unaware and bottom are computation-aware.

4.2 Offline Speech Translation

Although SimulST is our primary focus, we also evaluate all systems in an offline setting where latency is unconstrained and translation quality is maximized. In this scenario, we additionally leverage the self-cascade inference mode (Section 2.2) to further improve performance. To isolate the effects of fine-tuning and visual context, we evaluate the cascaded baseline in four configurations: base vs. LoRA fine-tuned, each with and without images. For OmniFusion, we train an additional mid-only fusion variant (without gating) to assess the contribution of the gating module. We compare all E2E configurations with and without images and self-cascade inference. Results are averaged over $en \rightarrow de, it, zh$ and reported using XCOMET-XL in Table 2, including fine-grained error categories.

The results yield several insights. In the cascaded pipeline, LoRA fine-tuning yields consistent gains (e.g., 85.94 → 86.59 with images), whereas adding images provides only marginal benefit. In contrast, the E2E OmniFusion model benefits substantially from both gated fusion (outperforming mid-only fusion) and image context showing the need for gating module and additional contextual information. Although the direct E2E translation score remains slightly below the best cascaded system, enabling self-cascade inference closes this gap almost entirely (cascaded LoRA FT: 86.59 vs. E2E gated fusion + self-cascade: 86.57). More importantly, error analysis reveals that our best E2E configuration produces the fewest critical and major errors (e.g., 719 vs. 751.3 major error for best cascaded). These findings demonstrate that OmniFusion, when equipped with image context and self-cascade inference, matches fine-tuned cascaded system in offline ST quality while offering significantly fewer severe translation errors.

4.3 Text-Image Translation

Beyond ST, caption translation also depends on visual context, where cascaded systems often fail

Category	Config	Image	XCOMET-XL	Minor	Major	Critical	Total
Cascade Systems (ASR → MT) Omni + SeedX	No FT	✗	85.88	1212.3	773.0	74.3	2059.7
	No FT	✓	85.94	1191.3	759.3	74.7	2025.3
	LoRA FT	✗	86.42	1239.0	751.3	75.0	2065.3
	LoRA FT	✓	**86.59**	1202.3	766.3	55.7	2024.3
End-to-End (Direct ST) OmniFusion	Mid Fusion	✗	83.57	**1114.0**	1188.3	72.3	2374.7
	Mid Fusion	✓	85.98	1256.3	749.0	61.7	2067.0
	Gated Fusion	✗	83.98	1119.3	1188.0	62.7	2370.0
	Gated Fusion	✓	86.24	1255.7	736.0	57.0	2048.7
	Gated Fusion + Self-Casade	✓	86.57	1231.0	**719.0**	**55.3**	**2005.3**

Table 2: Offline ST results averaged across en→zh, en→de, and en→it on the MCIF test. ✓ indicates models using image input. Bold indicates best XCOMET-XL and lowest error counts. Language-wise results are additionaly reported in Table 7, 8 and 9.

Model	Config	en → ar	en → cs	en → de	en → fr	en → ru	en → zh
Omni 2.5B	-	78.56	81.03	81.94	79.99	47.27	81.74
SeedX 7B	-	**80.39**	83.56	82.38	80.31	82.13	83.25
ZeroMMT 3.3B (Futeral et al., 2025)	-	**80.64**	**86.69**	83.54	83.40	83.87	77.85
TowerVision 9B (Viveiros et al., 2025)	-	70.85	86.62	**84.97**	**84.29**	**85.81**	**85.24**
OmniFusion 14B	Mid Fusion	**80.98**	**87.49**	**85.14**	**84.97**	**85.69**	**85.15**
OmniFusion 14B	Gated Fusion	**80.81**	**87.58**	**84.88**	**84.28**	84.62	84.34

Table 3: COMET scores (↑) on the CoMMuTE dataset. Statistically significant top-performing results per language pair are highlighted in **bold**. 'Config' distinguishes the two OmniFusion variants (mid-layer vs. gated fusion).

due to limited disambiguation. To demonstrate that OmniFusion effectively exploits image information, we evaluate multiple models on the CoMMuTE test set across several language pairs and report COMET scores in Table 3.

Omni 2.5B shows the weakest multilingual performance, motivating its integration with a stronger translation backbone. SeedX 7B achieves reasonable scores but consistently underperforms vision-aware models, highlighting its limitations without image context. Among multimodal systems, both Mid Fusion and Gated Fusion variants of OmniFusion 14B achieve near-SOTA performance, matching or surpassing vision-only models such as ZeroMMT 3.3B (Futeral et al., 2025) and TowerVision 9B (Viveiros et al., 2025) on most language directions.

In contrast to ST, Mid Fusion slightly outperforms Gated Fusion on CoMMuTE, suggesting that when images provide auxiliary rather than primary information, mid-layer visual integration is sufficient. Notably, unlike TowerVision, which relies on large-scale multimodal pretraining from scratch, OmniFusion builds on strong pretrained components, offering a more data- and training-efficient approach without sacrificing quality. These results highlight the effectiveness of our E2E design for visually grounded translation.

4.4 Contribution of Layers

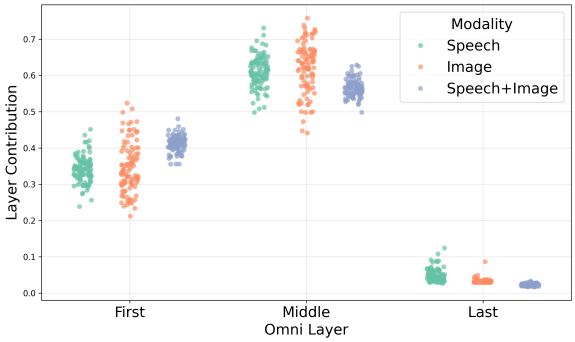


Figure 3: Analysis of the layer contribution from the MMFM in the gating module across modalities.

We introduced a gating module (Section 2.1) to fuse information from multiple MMFM layers. To analyze its behavior, we perform an ablation study by examining the average gate values assigned to the first, middle, and last MMFM layers, noting that gate weights sum to 1 at each timestep. We run inference on 100 MCIF examples under speech-only and speech+image conditions, and on CoMMuTE

for image-only input. The resulting average gate values are shown in Figure 3.

Across all settings, the first and middle layers receive substantially higher weights, while the last layer contributes marginally. This indicates that most transferable multimodal information is captured in earlier MMFM layers. We hypothesize that later layers are more language-specific, whereas middle layers encode more language-agnostic representations, consistent with prior findings (Liu and Niehues, 2025).

5 Related Work

Multimodal MT: Translating captions has been widely explored in the MT field over the last decade (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Several works focus on enabling NMT models to process images, either by training models from scratch (Vijayan et al., 2024) or by augmenting large pretrained models such as NLLB (Costa-Jussà et al., 2022; Futeral et al., 2025). Most recently, and in parallel to our work, Viveiros et al. (2025) extend Tower with image and video modalities through a computationally expensive multi-stage training pipeline.

While these efforts focus primarily on vision, an other line of direction is adding audio to translation LLMs (Ambilduke et al., 2025). A smaller line of work tackles the combination of vision and audio, such as Sinhamahapatra and Niehues (2025), which exploits slides and speech for ASR. Building on this trend, there is also increasing interest in translating audio together with video (Anwar et al., 2023; Choi et al., 2024). In contrast to these works, our approach focuses on jointly using images and audio specifically for ST, while remaining efficient in both training and inference.

Simultaneous ST: Many real-world scenarios require real-time ST without waiting for the full source utterance. However, most research focuses on training models in offline settings and only later adapting them to SimulST. Such adaptation typically relies on predefined or learned translation policies, such as wait-k (Ma et al., 2019; Elbayad et al., 2020), which assumes a fixed token delay before committing to output. Other approaches include Local Agreement (Liu et al., 2020), which selects the most stable prefix across consecutive read steps or attention-based heuristics (Papi et al., 2023). More recently, several works explore leveraging LLMs for SimulST policy learning (Koshkin et al., 2024b,a; Guo et al., 2025; Fu et al., 2025). In our work, we use Local Agreement to validate the effectiveness of our approach and to compare against cascaded baselines. However, our method is compatible in theory with other SimulST policies, and integrating more advanced or LLM-driven strategies is a promising direction for future work.

Modality Fusion: With the rapid development of multimodal foundation models, a key challenge is how to effectively integrate speech and image modalities into text-centric LLMs. Some approaches (Abdin et al., 2024; Xu et al., 2025; Team et al., 2025) process each modality with separate encoders and merge them later within the LLM backbone. In contrast, models such as (Ye et al., 2025) enable early interaction between vision and audio modalities, facilitating better cross-modal alignment. Beyond the timing of modality interaction, ongoing research also investigates the design of modality projectors (Verdini et al., 2024; Gaido et al., 2024b) and the use of continuous versus discrete tokens for non-text modalities (Zhan et al., 2024; Li et al., 2025), balancing the trade-offs between latency and translation quality.

6 Conclusion

We proposed a novel approach for fusing a MMFM with a translation LLM to enable high-quality, multilingual, multimodal, and simultaneous translation. We validated its effectiveness across three key scenarios. First, the fused model achieves lower latency while maintaining strong translation quality in SimulST, even with partial source audio. Second, it produces fewer major and critical errors. Third, it effectively exploits visual context for both speech and text translation, yielding consistent gains in multimodal settings. Our layer level analysis further shows that the first and middle MMFM layers contribute most to semantic multimodal representations, offering insight into where cross-modal information is integrated.

For future work, the proposed fusion framework naturally extends to additional modalities. Since Omni already supports video, incorporating video embeddings could provide richer contextual cues. Beyond multimodal inputs, multimodal generation can enable the system to produce not only text but also speech, images, or video exploting fusion strategy with pretrained components.

Limitations

A primary limitation of our work is that OmniFusion is trained with fixed task-specific prompts across the three supported translation settings, and therefore does not exhibit general instruction-following capabilities. It remains unclear whether our fusion strategy can be extended to support more flexible, instruction-driven multimodal behavior. In addition, most of our experiments use English as the source language. Although we report zero-shot results in Table 5, we do not evaluate the OmniFusion model with vision inputs in multilingual zero-shot conditions.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, James R Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C T Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *arXiv [cs.CL]*.

Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, and 1 others. 2025. Findings of the iwslt 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, and 25 others. 2024. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G C de Souza, and André F T Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv [cs.CL]*.

Kshitij Ambilduke, Ben Peters, Sonal Sannigrahi, Anil Keshwani, Tsz Kin Lam, Bruno Martins, André F T Martins, and Marcely Zanon Boito. 2025. From TOWER to SPIRE: Adding the speech modality to a translation-specialist LLM. *arXiv [cs.CL]*.

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. In *Proc. Interspeech 2023*, pages 4064–4068.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024. M^3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Bangkok, Thailand. Association for Computational Linguistics.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-X: Building strong multilingual translation LLM with 7B parameters. *arXiv [cs.CL]*.

Jeongsoo Choi, Se Jin Park, Minsu Kim, and Yong Man Ro. 2024. Av2av: Direct audio-visual speech to audio-visual speech translation with unified audio-visual speech representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27325–27337.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, and 1 others. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient wait-k models for simultaneous machine translation. In *Interspeech 2020-Conference of the International Speech Communication Association*, pages 1461–1465.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. 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.

Biao Fu, Donglei Yu, Minpeng Liao, Chengxi Li, Yidong Chen, Kai Fan, and Xiaodong Shi. 2025. Efficient and adaptive simultaneous speech translation with fully unidirectional architecture. *arXiv preprint arXiv:2504.11809*.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2025. Towards zero-shot multimodal machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 761–778, Albuquerque, New Mexico. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024a. Speech translation with speech foundation models and large language models: What is there and what is missing? *arXiv [cs.CL]*.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024b. Speech translation with speech foundation models and large language models: What is there and what is missing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14778.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-Gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv [cs.SD]*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Shoutao Guo, Xiang Li, Mengge Liu, Wei Chen, and Yang Feng. 2025. Streamuni: Achieving streaming speech translation with a unified large speech-language model. *arXiv preprint arXiv:2507.07803*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Sai Koneru, Maike Züfle, Thai-Binh Nguyen, Seymanur Akti, Jan Niehues, and Alexander Waibel. 2025. KIT's offline speech translation and instruction following submission for IWSLT 2025. *arXiv [cs.CL]*.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024a. LLMs are zero-shot context-aware simultaneous translators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1207, Miami, Florida, USA. Association for Computational Linguistics.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024b. TransLLaMa: LLM-based simultaneous translation system. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 461–476, Miami, Florida, USA. Association for Computational Linguistics.

Jindong Li, Yali Fu, Jiahong Liu, Linxiao Cao, Wei Ji, Menglin Yang, Irwin King, and Ming-Hsuan Yang. 2025. Discrete tokenization for multimodal llms: A comprehensive survey. *arXiv preprint arXiv:2507.22920*.

Danni Liu and Jan Niehues. 2024. Recent highlights in multilingual and multimodal speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 235–253, Stroudsburg, PA, USA. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in

fine-tuned LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.

Danni Liu, Gerasimos Spanakis, and Jan Niehues. 2020. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. In *21st Annual Conference of the International Speech Communication Association*, pages 3620–3624.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Javad Mohammad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020. Simuleval: An evaluation toolkit for simultaneous translation. In *Proceedings of the EMNLP*.

Sara Papi, Peter Polák, Dominik Macháček, and Ondřej Bojar. 2025a. How "real" is your real-time simultaneous speech-to-text translation system? *Trans. Assoc. Comput. Linguist.*, 13:281–313.

Sara Papi, Marco Turchi, and Matteo Negri. 2023. Alignatt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. In *Proceedings of INTERSPEECH 2023*, pages 3974–3978.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2025b. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *Preprint*, arXiv:2507.19634.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Supriti Sinhamahapatra and Jan Niehues. 2025. Do slides help? multi-modal context for automatic transcription of conference talks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16111–16121, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv [cs.CL]*.

Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sébastien Bratières, and 1 others. 2024. How to connect speech foundation models and large language models? what matters and what does not. *arXiv preprint arXiv:2409.17044*.

Vipin Vijayan, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Adding multimodal capabilities to a text-only translation model. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–28, Chicago, USA. Association for Machine Translation in the Americas.

André G Viveiros, Patrick Fernandes, Saul Santos, Sonal Sannigrahi, Emmanouil Zaranis, Nuno M Guerreiro, Amin Farajian, Pierre Colombo, Graham Neubig, and André F T Martins. 2025. TowerVision: Understanding and improving multilinguality in vision-language models. *arXiv [cs.LG]*.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *Preprint*, arXiv:2007.10310.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv [cs.CL]*.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9637–9662, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Impact of Training Tasks

To analyze the contribution of different training tasks proposed in Section 2.2, we train two ablation models: one without the OCR task and another without both OCR and self-cascading tasks. We evaluate all models using direct inference on the MCIF test set and report results in Table 4.

Training Configuration	en→de	en→it	en→zh
W/o OCR	87.12	82.42	79.75
W/o (OCR + Self-Cascade)	89.19	84.35	82.14
All Tasks	89.90	86.79	82.03
All Tasks (Self-Cascade Inference)	**90.05**	**87.20**	**82.47**

Table 4: Ablation study of training tasks on MCIF test set. All models use direct inference except the last row, which exploits self-cascading at inference time.

The results reveal several key findings. First, removing the OCR task during training leads to the weakest performance across all language pairs. We hypothesize this is because the model learns to focus on E2E translation without explicitly leveraging the intermediate ASR step and improving direct translation capability.

Second, removing both OCR and self-cascading tasks yields moderate performance. While this configuration cannot exploit self-cascading at inference time.

Finally, training on all tasks achieves the best results, particularly when combined with self-cascading inference. This demonstrates that our multi-task training strategy does not introduce negative interference; rather, it enables the model to effectively exploit both visual information and intermediate ASR outputs, leading to consistent improvements across all language pairs.

A.2 Prompt Building

System Prompt for all modalities

You are Qwen, a virtual human developed by the Qwen Team, Alibaba Group, capable of perceiving auditory and visual inputs, as well as generating text and speech.

User Prompt (ST with/without Image)

Transcribe the audio using the image for context, and perform OCR on the image for correct spelling of keywords and names.

User Prompt (TIT)

Translate the following sentence into `{tgt_lang}` using the image for context. `<IMG>` refers to the image. Use the image to determine formality, gender, keywords, and other details important for disambiguation: `{ex[src]}` `<IMG>`

A.3 Layer Analysis, Zero-shot ability and APE - AudioTower

In our preliminary experiments, we focus exclusively on the audio modality and investigate fusing Qwen-Audio 7B (Chu et al., 2023) with Tower-Instruct 7B v0.2 (Alves et al., 2024). We train the fused model for ST using the same setup described earlier, but without gating, and we independently fuse the first, middle, and last layers of Qwen-Audio to assess their individual contributions. We then compare this fused model against several baselines. First, we evaluate both the out-of-the-box and fine-tuned versions of Qwen-Audio (Qwen2 Audio FT), trained end-to-end on the same ST data. We also include a standard cascaded setup using off-the-shelf ASR and MT models. Additionally, we evaluate an Automatic Post-Editing (APE) configuration in which Tower corrects translations generated by Qwen-Audio.

To contextualize performance, we further compare against other off-the-shelf ST systems such as Tower-Spire (Ambilduke et al., 2025) and Seamless-M4T (Barrault et al., 2023), which provides insight into multilingual generalization. Beyond the main fused model, we also train an APE variant of our fusion architecture that post-edits the outputs of cascaded systems. Unlike traditional

Algorithm 1 Online Prompt building for examples in the training data to enable alignment and self-cascading inference mode.

Require: Example ex

```
// SeedX prompt (source text, language tag and reference)
src, tgt, img, aud, ocr, lang ← EXTRACT(ex)
suffix ← ⟨LANG(lang)⟩

if aud and not img then                                          ▷ Speech only
    if RANDOM < 0.1 then
        inp ← ⟨en⟩                                               ▷ English-Audio only in our dataset
        lbl ← ASR(src)+⟨eos⟩+suffix+tgt+⟨eos⟩                    ▷ ASR Self-cascade
    else
        inp ← suffix
        lbl ← tgt+⟨eos⟩
    end if
else if img and not aud then                                     ▷ Image only
    if RANDOM < 0.05 then
        SWAP(src, tgt)
        suffix ← ⟨English⟩
    end if
    src_m ← MASK(src, prob_words = 0.2, mask_chance = 0.2)       ▷ Mask 20% source words for 20% of the time to pay attention to Image
    inp ← ⟨SRC⟩ src_m+⟨eos⟩+suffix
    lbl ← tgt+⟨eos⟩
else if img and aud then                                         ▷ Audio + Image
    if RANDOM < 0.8 then
        if RANDOM < 0.1 then
            inp ← ⟨en⟩
            lbl ← ASR(src)+⟨eos⟩+suffix+tgt+⟨eos⟩
        else
            inp ← suffix
            lbl ← tgt+⟨eos⟩
        end if
    else
        inp ← ⟨OCR⟩                                              ▷ OCR Self-cascade
        lbl ← ocr+⟨eos⟩+suffix+tgt+⟨eos⟩
    end if
else
    error "no modalities"
end if

// Omni prompt (Multimodal prompt builder)
if aud and img then
    user ← "Transcribe with OCR and image context:" || img || aud
else if aud then
    user ← "Transcribe audio:" || aud
else
    user ← "Translate with image:" || img
end if
STORE(user,inp, lbl)
```

Model	*Mode*	*Hypothesis*	*Comet-22 (↑)*					
			en → de	en→ zh	zh → de	es → it	ja → ru	fr → kr
Baselines								
Qwen2 Audio	E2E	_	81.24	78.20	73.59	83.38	70.70	71.26
Qwen2 Audio FT	E2E	_	84.82	85.57	80.22	84.17	67.69	77.09
Qwen2 Audio + Tower	Cascade	_	85.85	**85.97**	82.94	85.36	69.26	**84.78**
Tower APE	Cascade APE	Qwen2 Audio + Tower	85.96	**85.77**	82.59	85.21	47.43	70.67
Off-the-shelf ST models								
Tower-Spire	E2E	_	82.77	81.63	_	_	_	_
Tower-Spire	Self-Cascade	_	85.17	85.73	_	_	_	_
Seamless-M4T v2	E2E	_	85.63	80.05	70.22	77.78	70.58	78.91
E2E models								
AudioTower-Zero	E2E	_	85.13	83.97	76.62	83.98	68.71	81.64
AudioTower-Mid	E2E	_	86.02	85.03	80.62	84.86	74.47	83.26
AudioTower-Last	E2E	_	82.97	81.03	54.12	49.57	50.19	52.31
E2E + Speech-Aware APE model								
AudioTower-Mid	Cascade APE	Qwen2 Audio + Tower	**86.34**	**85.99**	**83.42**	**85.96**	**78.26**	**84.59**
AudioTower-Mid	E2E	Empty	85.77	84.79	80.39	85.27	74.67	82.71

Table 5: Offline ST results on FLEURS test set. Speech-Aware APE models use the hypothesis from a system and perform post-editing. Statistically significant top-performing results per language pair are highlighted in **bold**. Underlined langauges are not seen during the fusing process as source/target.

Model	*Mode*	*Hypothesis*	*Comet-22 (↑)*					
			en → de	en→ zh	zh → de	es → it	ja → ru	fr → kr
Qwen2 Audio + Tower	Cascade	_	85.85	85.97	82.94	85.36	69.26	84.78
Whisper + Tower	Cascade	_	86.71	**86.31**	82.23	**86.54**	83.56	**85.29**
AudioTower-Mid	Cascade APE	Qwen2 Audio + Tower	86.34	85.99	**83.42**	85.96	78.26	84.59
AudioTower-Mid	Cascade APE	Whisper + Tower	**87.09**	**86.34**	82.81	**86.60**	**84.38**	84.95
Character-Error-Rate (↓)								
Whisper	ASR	_	3.14	3.15	25.37	2.11	7.67	3.28

Table 6: Analysis of system combination by using Whisper to generate the hypothesis for post-editing on FLEURS test set. Statistically significant top-performing results per language pair are highlighted in **bold**. Underlined langauges are not seen during the fusing process as source/target.

APE, this Speech-Aware APE allows the model to use the source audio when correcting errors, enabling more informed revisions. For both APE variants, we generate synthetic APE tuples by running Qwen-Audio on the training data and fine-tune with LoRA.

We report results on the FLEURS test set in Table 5, including evaluations on zero-shot language directions not present on either the source or target side during training. First, while fine-tuning improves Qwen2 Audio, it still falls short of the out-of-the-box cascaded setup across all languages, underscoring the strength of translation-centric LLMs. Tower-based APE is competitive for seen languages but struggles in harder zero-shot directions, such as *fr→kr*. Off-the-shelf end-to-end ST models lag significantly behind cascaded systems as well.

Our proposed fused model (AudioTower) with middle-layer fusion achieves the strongest overall performance and even surpasses the cascaded baseline in several languages. Notably, both the first and middle layers show strong transfer to unseen languages, whereas the last layer collapses entirely—supporting the hypothesis that upper layers encode highly language-specific features, making them less suitable for cross-lingual contextual embeddings.

Finally, the APE variant of the fused model delivers the best results overall by combining the strong hypotheses of the cascaded pipeline with the ability to exploit the raw audio when correcting errors. We also test this APE model with empty hypotheses and observe performance comparable to AudioTower-Mid, indicating its flexibility for multiple use cases.

Given the strong APE results, we additionally explore whether the model can correct outputs from different ASR systems. Using Whisper (Radford et al., 2023) for ASR, we evaluate the fused APE model on Whisper+Tower outputs and report results in Table 6. When Whisper+Tower produce the strongest cascaded baseline, using them as hy-

Category	Model Name	Image	XCOMET-XL	Minor	Major	Critical	Total
Cascade Systems (ASR → MT) Omni + SeedX	No FT	✗	89.66	2172	132	68	2372
	No FT	✓	89.49	**2115**	133	76	2324
	LoRA FT	✗	89.75	2209	128	67	2404
	LoRA FT	✓	**90.18**	2135	118	**56**	**2309**
End-to-End (Direct ST) OmniFusion	Mid Fusion	✗	88.44	2126	157	74	2357
	Mid Fusion	✓	90.03	2256	104	68	2428
	Gated Fusion	✗	88.09	2133	181	87	2401
	Gated Fusion	✓	89.90	2217	106	60	2383
	Gated Fusion + Self-Cascade	✓	90.05	2191	**94**	62	2347

Table 7: Offline ST results en→de on the MCIF test. ✓ indicates models using image input. Bold indicates best XCOMET-XL and lowest error counts.

Category	Model Name	Image	XCOMET-XL	Minor	Major	Critical	Total
Cascade Systems (ASR → MT) Omni + SeedX	No FT	✗	86.03	850	932	51	1833
	No FT	✓	86.11	862	891	67	1820
	LoRA FT	✗	86.76	884	855	54	1793
	LoRA FT	✓	86.85	868	855	55	1778
End-to-End (Direct ST) OmniFusion	Mid Fusion	✗	82.80	754	1065	93	1912
	Mid Fusion	✓	85.75	899	903	64	1866
	Gated Fusion	✗	84.06	**745**	1085	56	1886
	Gated Fusion	✓	86.79	914	847	48	1809
	Gated Fusion + Self-Cascade	✓	**87.20**	879	**838**	**44**	**1761**

Table 8: Offline ST results en→it on the MCIF test. ✓ indicates models using image input. Bold indicates best XCOMET-XL and lowest error counts.

Category	Model Name	Image	XCOMET-XL	Minor	Major	Critical	Total
Cascade Systems (ASR → MT) Omni + SeedX	No FT	✗	81.94	615	1255	104	1974
	No FT	✓	82.23	597	1254	81	1932
	LoRA FT	✗	**82.74**	624	1271	104	1999
	LoRA FT	✓	**82.74**	604	1326	**56**	1986
End-to-End (Direct ST) OmniFusion	Mid Fusion	✗	79.47	**462**	1343	50	1855
	Mid Fusion	✓	82.15	614	1240	53	1907
	Gated Fusion	✗	79.78	480	1298	**45**	**1823**
	Gated Fusion	✓	82.03	636	1255	63	1954
	Gated Fusion + Self-Cascade	✓	82.47	623	**1225**	60	1908

Table 9: Offline ST results en→zh on the MCIF test. ✓ indicates models using image input. Bold indicates best XCOMET-XL and lowest error counts.

potheses likewise yields the highest APE performance—highlighting that Speech-Aware APE can be beneficial not only for E2E setups but also for offline cascaded scenarios.

Language	Code
Arabic	*ar*
Chinese	*zh*
Croatian	*hr*
Czech	*cs*
Danish	*da*
Dutch	*nl*
Finnish	*fi*
French	*fr*
German	*de*
Hungarian	*hu*
Indonesian	*id*
Italian	*it*
Japanese	*ja*
Korean	*ko*
Malay	*ms*
Norwegian	*no*
Norwegian Bokmål	*nb*
Polish	*pl*
Portuguese	*pt*
Romanian	*ro*
Russian	*ru*
Spanish	*es*
Turkish	*tr*
Vietnamese	*vi*

Table 10: Language codes denoted in paper following Cheng et al. (2025).

Hyperparameter	Value
Gradient checkpointing	True
Gradient checkpointing kwargs	`{'use_reentrant': False}`
Save steps	1000
Evaluation steps	1000
Save strategy	steps
Evaluation strategy	steps
Logging steps	5
Logging strategy	steps
Learning rate	1e-4
Train batch size (per device)	4
Eval batch size (per device)	4
Gradient accumulation steps	2
Weight decay	0.01
Save total limit	3
Max steps	20000
bf16	True
Push to hub	False
Metric for best model	loss
Remove unused columns	False
Label names	`["labels"]`
Dataloader num workers	1
DDP find unused parameters	True

Table 11: Huggingface trainer hyperparameters used for the model on 4 Nvidia A100 48GB GPU's with DDP training mode. Unspecified hyperparameters are set to default

Image	Cascaded FT Output	OmniFusion Output	Reference
	Hallo, willkommen zu unserer Präsentation von **Deplain**, einem neuen Korpus für die Identifizierung deutscher Texte auf Dokumenten- und Satzebene.	Hallo, willkommen zu unserer Präsentation von **DEPLAIN**, einem neuen Korpus für die deutsche Texterkennung auf Dokumentebene und auf Satzebene.	Herzlich willkommen zu unserer Präsentation von **DEPLAIN**, einem neuen Korpus für die deutsche Texterkennung auf Dokument- und Satzebene.

Table 12: Example illustrating where OmniFusion is better than cascaded variant. Example taken from MCIF segment 1.