OmniFusion: Simultaneous Multilingual and Multimodal Translations via Modular Fusion

Anonymous ACL submission

Abstract

Open-source text-only translation Large Language Models (LLMs) have rapidly improved in multilingual coverage and translation quality, but their unimodal design limits their applicability to multimodal translation. In speech translation (ST), they typically operate in cascaded pipelines with automatic speech recognition followed by translation, introducing additional latency particularly for simultaneous ST (SimulST) and they cannot exploit visual context for disambiguation in either speech or caption translation. In contrast, pretrained multimodal foundation models (MMFMs) provide strong cross-modal perception and reasoning but lack the multilingual depth and translation performance of specialized translation LLMs. We introduce OmniFusion, a unified architecture that combines a multimodal foundation model with a translation-specialized LLM through a lightweight gated fusion module. Built with Qwen2.5-Omni-7B as the MMFM and Seed-X-PPO-7B as the translation LLM, OmniFusion supports 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 and reduces SimulST latency by one second compared to cascaded pipelines while maintaining competitive translation quality.[1]

1 Introduction

In recent years, NLP research has 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 across sources. While unimodal systems can perform well in certain settings, further progress requires models that exploit additional cues without degrading quality.

Translation is a representative task where multimodal integration is beneficial. In Speech Translation (ST), the goal is to convert spoken input in a source language into a target language (Ahmad et al., 2024; Agostinelli et al., 2025). Despite recent advances, leveraging visual context such as images or presentation slides accompanying speech remains underexplored (Liu and Niehues, 2024; Gaido et al., 2024a; Sinhamahapatra and Niehues, 2025). Similarly, caption translation involves short, image-linked text, where visual information can aid disambiguation and grounding.

Although MMFMs can process multimodal inputs, their multilingual translation performance is generally weaker than that of specialized translation LLMs (Alves et al., 2024; Cheng et al., 2025), leading to lower translation quality and limited language coverage. In ST, a common approach is a cascaded pipeline: automatic speech recognition (ASR) followed by machine translation (MT) (Ahmad et al., 2024; Agostinelli et al., 2025; Koneru et al., 2025). While this modular approach allows independent optimization, it prevents the translation model from leveraging multimodal context and adds latency, a limitation especially critical in simultaneous ST (SimulST) scenarios (Agostinelli et al., 2025; Papi et al., 2025a).

These limitations motivate the development of an end-to-end multilingual and multimodal translation system supporting ST, image-aware ST, and caption translation. A unified model can exploit shared data across tasks and simplify deployment. Extending a translation LLM to handle additional modalities is one approach, but prior work (Ambilduke et al., 2025; Viveiros et al., 2025) often trains from scratch or relies solely on pretrained encoders, without fully leveraging the higher-level representations of existing MMFMs, making training inefficient.

We propose leveraging pretrained MMFMs to introduce multimodal capabilities into a multilin-

[1] Code is attached as a zip file.

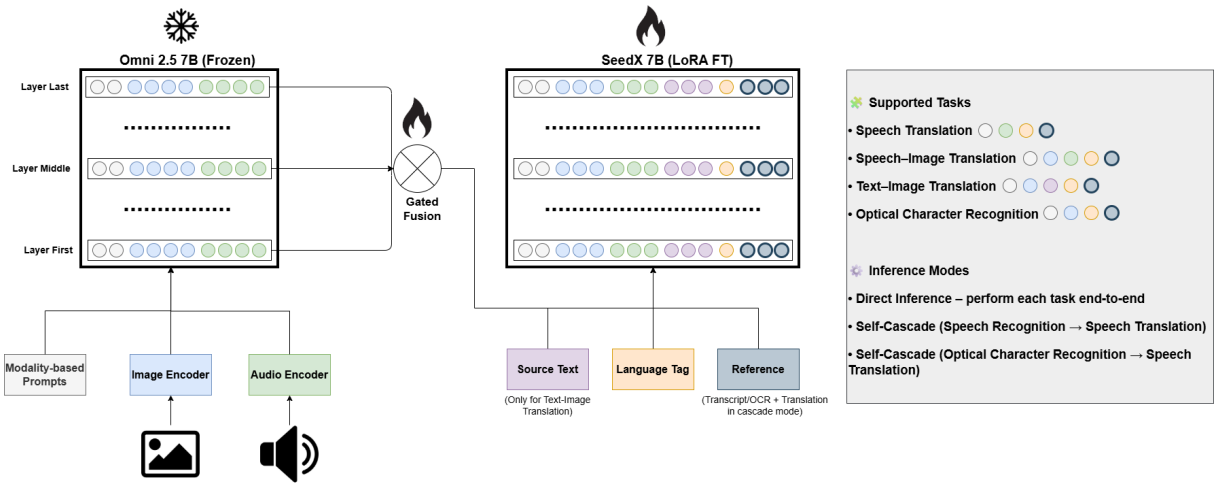


Figure 1: Architecture of OmniFusion with supported tasks and inference modes.

gual LLM within a unified E2E architecture. Our approach tightly couples the two models during training and employs a multi-task objective augmented with auxiliary tasks, including OCR and ASR with self-cascading, to enhance multimodal alignment and translation performance.

Our contributions are threefold: (1) We present OmniFusion, an E2E model built using translation LLMs and pretrained MMFMs that jointly supports speech, text and image translation tasks. **(2)** We introduce a gated fusion architecture that selectively integrates representations from multiple MMFM layers into a translation LLM, enabling efficient training and inference for context-aware translations; and **(3)** We demonstrate that **OmniFusion** reduces SimulST latency by ~ 1 second (Figure 3), lowers major and critical error rates for offline ST (Table 1), and achieves state-of-the-art results on CoMMuTE (Table 2).

OmniFusion demonstrates that MMFMs and translation LLMs can be effectively combined within a single architecture for multilingual multimodal translation.

2 Approach

We propose fusing a MMFM with a task-specific translation LLM to leverage their complementary strengths. The MMFM provides perceptual representations, while the translation LLM provides multilingual understanding and generation capabilities[2].

We first describe the overall architecture and training process, abstracting over task-specific inputs and outputs. Motivated by a probing experiment, we define the information flow, training loss, and fusion of MMFM layers. We then detail training tasks aligning image and audio modalities to the translation LLM and discuss prompts enabling self-cascading inference, which improves translation quality at higher latency. An overview is shown in Figure 1.

2.1 Architecture

Let $\mathcal{M}_{\text{NMT}}$ be the translation LLM and $\mathcal{M}_{\text{MMFM}}$ the MMFM, both with a language model backbone. $\mathcal{M}_{\text{NMT}}$ is trained via supervised fine-tuning with:

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

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 $\mathcal{M}_{\text{MMFM}}$, image and audio inputs are encoded and projected to the MMFM embedding dimension D_{MMFM}, forming a token sequence compatible with the LLM.

2.1.1 Probing MMFM Representations

To capture both perceptual and higher-level understanding, we hypothesize that the lower MMFM layers encode perceptual features and higher-level abstract reasoning. We validate this by computing cosine similarity between consecutive layer outputs, which is commonly used as a proxy for layer

[2]Using the MMFM as encoder does not provide significant computational overhead as decoders are primary latency bottleneck (Guta et al., 2025).

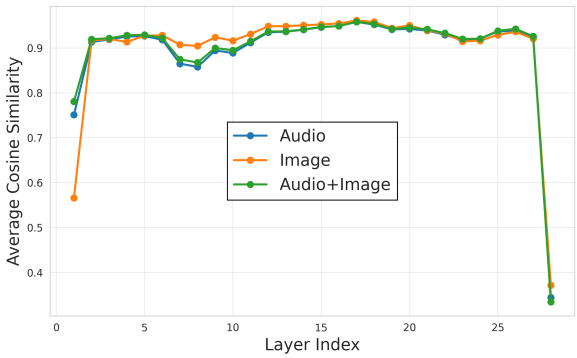


Figure 2: Average cosine similarity between hidden representations of adjacent layers, computed with Qwen-Omni 2.5 7B on the MCIF test set (en → de) under different modality conditions.

importance in model pruning literature. We compute with different modalities and visualize in Figure 2. We see that the first and last layers exhibit the largest changes, while other layers provide small updates. Based on this, we propose using the first, middle, and last layers as inputs to generate multimodal representations as they capture the three regions[3].

2.2 Gated Fusion

Accordingly, we extract these layers from the MMFM, denoted as:

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

where L_m is the length of the MMFM hidden states. Directly concatenating these multi-dimensional MMFM representations into $\mathcal{M}_{\text{NMT}}$ is problematic because it triples the sequence length, layer relevance may vary by task, and embedding dimensions may not match. To address this, we use a gated fusion layer that weights each hidden state, sums across layers, and projects the result via an MLP to align with the translation model:

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

[3]We perform ablation using all layers in Table 3 and empirically validate this choice.

where $\oplus$ is concatenation done per each token across layers, 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:

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

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

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

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

2.3 Training

Effectively leveraging MMFM representations requires aligning multimodal features with the translation LLM's token representations. We achieve this through multitask training that includes ST, ASR, and OCR objectives.

2.3.1 Multi-task Learning

We train the model on three primary multimodal translation tasks. In Speech Translation (ST), the model generates a target-language translation from source-language audio. In Speech–Image Translation (SIT), it translates audio together with an image providing additional context, such as a presentation slide aligned with the spoken segment. In Text–Image Translation (TIT), the model translates source text paired with an aligned image, as in caption translation scenarios. Training across these tasks exposes the model to different modality combinations, enabling effective alignment and exploitation of multimodal representations.

2.3.2 Self-Cascading (ASR and OCR)

In addition to multi-task learning, we adopt self-cascading, which decomposes translation into intermediate steps. For English-to-German ST, the model first generates the ASR transcription, then produces the final translation conditioned on it, providing explicit alignment between audio representations and the LLM's semantic space. For visual inputs, OCR text serves as an intermediate

step before translation, aligning visual and textual representations. During training, self-cascading is applied stochastically based on task and modality, with modality-specific prompts (Algorithm 1). We freeze the MMFM and fine-tune the translation LLM with LoRA (Hu et al., 2021), reducing compute costs while retaining pretrained multimodal capabilities. We also find that adding LoRA to the MMFM degrades performance due to catastrophic forgetting.

3 Experimental Setup

3.1 Training Data

We train the fused model to perform translation across speech and image modalities. An overview of data sources is provided in Appendix Table 6. For ST, we use subsets of Europarl-ST (Iranzo-Sánchez et al., 2020) and Covost-2 (Wang et al., 2020), sampling multiple target languages. Training is limited to these subsets to reduce computational costs. To exploit images in audio translation, we use the M3AV corpus (Chen et al., 2024), containing English scientific talks paired with slides. The slides include Paddle-OCR (Cui et al., 2025) predictions, allowing training either for direct translation or via an intermediate OCR step for better alignment. Since M3AV has limited language coverage, we augment ASR transcripts with synthetic translations (**M3AV-Aug**), enabling ST with slides and supporting self-cascade training.

For TIT, we use Multi30k (Elliott et al., 2016), which pairs images with multilingual captions. This task encourages leveraging visual content beyond OCR, including objects and scene understanding, to improve translation.

3.2 Models

Several MMFMs have recently shown strong gains in quality and efficiency (Abdin et al., 2024; Team et al., 2025). As our setting requires processing both audio and image modalities with cross-modal reasoning, we use Qwen Omni 2.5 7B[4] (Xu et al., 2025). For translation, we adopt SeedX PPO 7B[5] (Cheng et al., 2025), which provides broad language coverage and strong translation performance compared to closed-source alternatives. Using the training data described above, we refer to the resulting fine-tuned model as **OmniFusion**. While we report results with these models, the approach is general and applicable to other base LLMs. Training hyperparameters are listed in Table 14. Validation uses Covost-2 dev data, selecting the checkpoint with the lowest loss after 20k steps.

3.3 Evaluation: Data & Metrics

Evaluating multimodal SimulST requires aligned audio–visual translation data. We exclude M3AV from testing due to training overlap and its monolingual nature, and instead assess robustness on unseen domains using the MCIF test set (Papi et al., 2025b), which contains ACL talk recordings translated from English into three target languages, and FLEURS (Conneau et al., 2023) for multilingual coverage. Latency is reported using Average Lagging (Ma et al., 2020), and translation quality is measured on these datasets.

For TIT, we target cases where visual context is crucial for disambiguation, using the CoMMuTE benchmark (Futeral et al., 2023, 2025). Translation quality is evaluated with XCOMET-XL (Guerreiro et al., 2024) for ST, enabling detailed error analysis, and with COMET[6] (Rei et al., 2022) for TIT.

4 Results

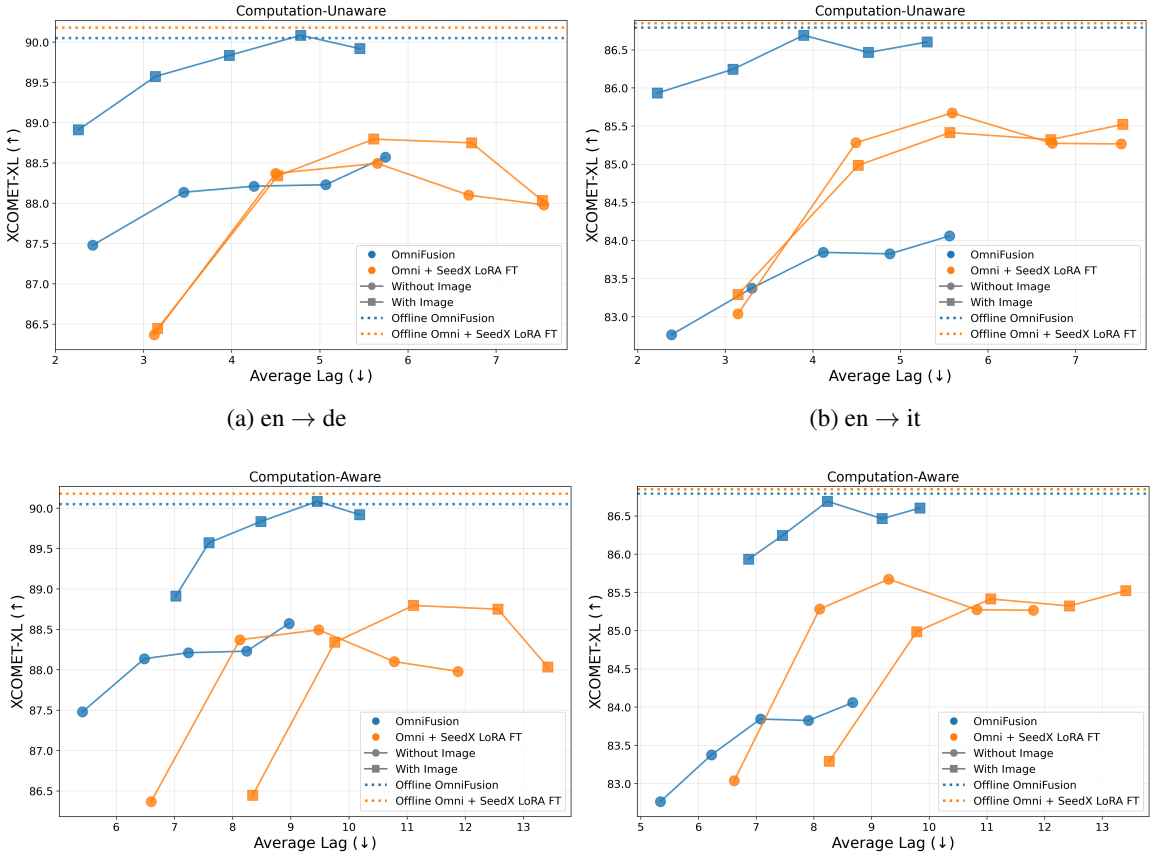
Our goal is to develop a high-quality multimodal translation system with low-latency end-to-end SimulST that effectively leverages image context. We evaluate latency–quality trade-offs in simultaneous settings, compare offline translation quality and errors against cascaded models trained on the same data, assess text-and-image translation performance against multimodal baselines, and present ablations on MMFM layers and fusion mechanisms.

We also report ablations on training strategies to disentangle architectural and training effects (Appendix Table 9). Additionally, we include preliminary experiments with alternative model pairings and layer selections (Qwen-Audio (Chu et al., 2023) and Tower-Instruct (Alves et al., 2024)) as well as post-editing results (Appendix Tables 5 and 8). **We observe similar findings and do not explore additional model combinations due to computational constraints.**

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. Since our E2E model

[4] Qwen/Qwen2.5-Omni-7B
[5] ByteDance-Seed/Seed-X-PPO-7B
[6] Unbabel/wmt22-comet-da



(a) en → de (b) en → it

(c) en → de (d) en → it

Figure 3: 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.

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. As computational time in SimulST is critical, we report both **computation-aware and -unaware** Average Lagging along with XCOMET-XL scores for $en \rightarrow de$ and $en \rightarrow it$ in Figure 3.

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 highest quality, indicating that visual context is more effectively exploited in the E2E model, enabling earlier and more stable commitments. Image processing introduces additional delay, highlighting a latency–quality trade-off. We hypothesize that OmniFusion's advantage stems from (i) our fine-tuning strategy, including an OCR objective that aligns audio and visual representations (Table 9 shows performance degradations without multitask objectives) and (ii) the efficiency of the end-to-end architecture, which prevents the delayed onset of translation inherent in cascaded pipelines. Combined with Local Agreement decoding, OmniFusion achieves improved latency–quality trade-offs compared to cascaded systems.

4.2 Offline Speech Translation

Although SimulST is our primary focus, we also evaluate all systems in an offline setting where la-

Category	Config	Image	XCOMET-XL	Minor	Major	Critical	Total
Omni FT (Direct ST)	LoRA FT	✗	82.30	**1079.0**	852.0	90.3	2021.3
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-Cascade	✓	86.57	1231.0	**719.0**	**55.3**	**2005.3**

Table 1: 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 additionally reported in Table 10, 11 and 12.

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	-	**80.64**	**86.69**	83.54	83.40	83.87	77.85
TowerVision 9B	-	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 2: COMET scores (↑) on the CoMMuTE dataset. Statistically significant top-performing results per language pair are highlighted in **bold**. Qualitative example can be found in Table 16

tency is unconstrained and translation quality is maximized. In this scenario, we additionally leverage the self-cascade inference mode (Section 2.3) to further improve performance. To isolate the effects of fine-tuning and visual context, we evaluate the cascaded baseline in five configurations: base MMFM fine-tuned for ST[7], and base vs. LoRA cascade fine-tuned with and without images. For OmniFusion, we train an additional mid-only fusion variant (without gating) (We find using the middle layer performs best in our preliminary experiments from Table 5) 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 1, including fine-grained error categories. Additionally, inference times are reported in Table 7, although latency is not a primary constraint.

First, directly fine-tuning the MMFM for speech translation performs substantially worse than OmniFusion (82.30 vs 83.98 XCOMET), confirming combining is more effective than adapting the MMFM alone. In the cascaded pipeline, LoRA fine-tuning provides consistent gains (e.g., 85.94 → 86.59 with images), while images offer only marginal improvements. In contrast, the E2E OmniFusion model benefits substantially from both gated fusion (outperforming mid-only fusion) and image context, highlighting the importance of gating and additional contextual information. Although direct E2E translation slightly trails the best cascaded system, enabling self-cascade inference nearly closes the gap (cascaded LoRA FT: 86.59 vs. E2E gated fusion + self-cascade: 86.57). More importantly, error analysis shows that our best E2E configuration produces fewer critical and major errors (e.g., 719 vs. 751.3 major errors for the best cascaded system). These results demonstrate that OmniFusion with image context and self-cascade inference matches fine-tuned cascaded systems in offline ST quality while producing fewer severe translation errors.

4.3 Text-Image Translation

Caption translation relies heavily on visual context, where cascaded systems often fail due to limited disambiguation. To evaluate OmniFusion's ability to leverage images, we report COMET scores on the CoMMuTE test set across multiple language pairs in Table 2. Omni 2.5B exhibits the weakest multilingual performance, motivating its fusion

[7]Even if it is competitive, it is insufficient due to limited language coverage compared to SeedX

Model	Mode	Fusion Layers	en → de		cn → de		jp → ru		fr → ko		ur → cn	
			Chrf	Comet-22	Chrf	Comet-22	Chrf	Comet-22	Chrf	Comet-22	Chrf	Comet-22
Tower-Spire	Direct	✗	56.9	82.54	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Tower-Spire	Self-Cascade	✗	61.9	85.02	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SeamlessM4T v2	Direct	✗	62.4	85.60	35.8	70.22	36.9	71.77	20.8	74.12	13.5	61.11
SeamlessM4T v2	Self-Cascade	✗	63.3	85.81	36.9	71.34	38.2	69.85	21.0	75.22	13.5	63.75
Our Models - OmniFusion												
Attention	Direct	All	60.9	84.02	37.4	72.96	26.6	65.62	21.2	75.36	4.3	43.66
Attention	Direct	3	61.8	86.95	39.4	74.86	27.6	68.81	21.9	78.94	5.5	46.78
Gated	Direct	All	62.3	85.37	39.6	75.91	28.8	69.33	21.8	77.50	5.0	45.78
Gated	Direct	3	**64.0**	**87.12**	**44.5**	**81.82**	**37.2**	**81.95**	**25.5**	**82.74**	**12.5**	**65.60**

Table 3: Ablation study of fusion architecture with non-english zero-shot source languages. Tower-Spire only supports English audio.

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 reach near state-of-the-art 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.

Unlike in 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 context-awareness of our E2E design for visually grounded translation**.

4.4 Layer Contribution - Training

We use only the first, middle, and last layers in the gated fusion. To validate this choice, we perform ablation experiments where all layers of the MMFM are used within the gating module. In addition, we explore an attention-based fusion approach, where a learned query token attends over all layer hidden states to produce new contextualized representations. Moreover, we compare our method with other state-of-the-art models and report the results in Table 3 on the FLEURS test set to account for multilinguality.

First, we observe that the gated variants achieve the best performance across all language pairs, including challenging ones such as Urdu to Chinese. Next, using all layers in the gating module leads to performance degradation, which becomes even more pronounced with attention-based fusion. We attribute this primarily to the increased noise introduced when incorporating all layers. Further, attention-based fusion strategies fail compared with gating due to training/test mismatch, especially for non english source languages. In comparison, our proposed approach introduces minimal number of parameters, avoiding reliance on large datasets while still achieving competitive performance.

4.5 Layer Contribution - Inference

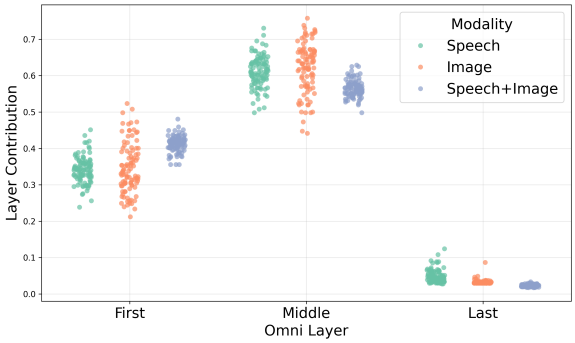


Figure 4: Analysis of the layer contribution from the MMFM in the gating module across modalities.

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 4. Across settings, 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.

Inference-time Gate Force	Speech	Speech + Image
First	55.22	63.02
Middle	Halluc. + ASR	Halluc. + ASR
Last	EOS	EOS
✗	89.90	90.05

Table 4: Inference time control of the gating with forcing first, middle or last layer. We report XCOMET-XL scores on the MCIF en → de test set.

Next, we conduct another ablation by performing inference-time control in the gating module. We ignore the gate's predictions and instead force the model to use only the first, middle, or last layer, and report the results in Table 4. Forcing the first layer degrades translation quality but does not collapse the system and even improves performance when images are present. In contrast, forcing the middle or last layer leads to failure modes. With the middle layer, the system often generates the transcript along with additional noise, while forcing the last layer causes the model to predict only the `eos` token. We hypothesize that the first layer provides reasonably aligned multimodal embeddings, whereas the middle layer contains more processed representations that become ineffective due to a training–test mismatch. Notably, training with only a single layer still produces workable systems, although using the last layer yields the lowest performance and weakest zero-shot transfer, as shown in Table 5.

5 Related Work

Multimodal MT: Translating captions has been widely explored over the past decade (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Several works extend NMT models to process images, either by training from scratch (Vijayan et al., 2024) or by augmenting pretrained models such as NLLB (Costa-Jussà et al., 2022; Futeral et al., 2025). Recently, and in parallel with our work, Viveiros et al. (2025) extend Tower with image and video modalities through a computationally expensive multi-stage training pipeline. Beyond vision, another line of research explores incorporating audio into translation LLMs (Ambilduke et al., 2025). A smaller body of work investigates the joint use of vision and audio, such as Sinhamahapatra and Niehues (2025), which leverages slides and speech for ASR. Recent studies also explore translating with video (Anwar et al., 2023; Choi et al., 2024). In contrast, our approach focuses on jointly leveraging images and audio while remaining efficient for data and latency.

Simultaneous ST: Real-world scenarios require real-time ST without waiting for the full utterance. However, most research trains models in offline settings and later adapts 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, and 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 this work, we adopt Local Agreement to evaluate our approach and compare against cascaded baselines. Our method is, however, compatible with other SimulST policies given its E2E nature.

Modality Fusion: 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 within the LLM. 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 such interaction, recent works also investigates the design of modality projectors (Verdini et al., 2024; Gaido et al., 2024b) and the use of discrete tokens (Zhan et al., 2024; Li et al., 2025), balancing between latency and translation quality.

6 Conclusion

We propose a novel approach for fusing an MMFM with a translation LLM to enable high-quality, multilingual, multimodal, and simultaneous translation. We validate its effectiveness across three 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.

Limitations

Our training data is primarily English-centric, which may limit performance for other source languages. Additionally, while we evaluate multimodal translation on established benchmarks, broader human evaluation would help better assess the model's ability to utilize visual context. Future work will explore multilingual multimodal training and larger-scale human evaluation Another 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.

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*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

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*.

Andreas Guta, Frithjof Petrick, and Peter Polák. 2025. The green knight: Green machine translation with knowledge-distilled, narrow, inexpensive, greedy, hybrid transformers.

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, 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 the training tasks proposed in Section 2.3, 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 9. The results reveal several insights. Removing the OCR task leads to the weakest performance across all language pairs, likely because the model cannot leverage the intermediate ASR step to improve direct translation. Removing both OCR and self-cascading tasks yields moderate performance, as this configuration cannot exploit self-cascading at inference time. In contrast, training on all tasks achieves the best results, particularly when combined with self-cascading inference, demonstrating that our multi-task training strategy effectively leverages both visual information and intermediate ASR outputs without causing negative interference.

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 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 tar-

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.

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 6: 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 13.

Model	Time(s)	Image	XCOMET-XL
Omni + SeedX FT	2.96	w/o	89.75
OmniFusion	**1.98**	w/o	88.09
OmniFusion SC	3.78	w/o	88.44
Omni + SeedX FT	3.85	w/	**90.18**
OmniFusion	3.15	w/	89.90
OmniFusion SC	4.82	w/	90.05

Table 7: Quality vs latency tradeoffs in the offline setting for English → German on the MCIF test set. SC indicates Self-Cascade mode.

get 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 8. When Whisper+Tower produce the strongest cascaded baseline, using them as hypotheses 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.

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 8: 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.

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 9: Ablation study of training tasks on MCIF test set.

Category	Model Name	Image	XCOMET-XL	Minor	Major	Critical	Total
Omni FT (Direct ST)	LoRA FT	✗	85.90	**2047**	166	173	2386
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 10: 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
Omni FT (Direct ST)	LoRA FT	✗	80.88	**727**	1018	54	1799
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 11: 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
Omni FT (Direct ST)	LoRA FT	✗	80.13	463	1372	**44**	1879
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 12: Offline ST results en→zh on the MCIF test. ✓ indicates models using image input. Bold indicates best XCOMET-XL and lowest error counts.

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*
Urdu	*ur*
Vietnamese	*vi*

Table 13: Language codes denoted in our 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 14: Huggingface trainer hyperparameters used for the model on 4 Nvidia A100 48GB GPU's with DDP training mode. Unspecified hyperparameters are set to default

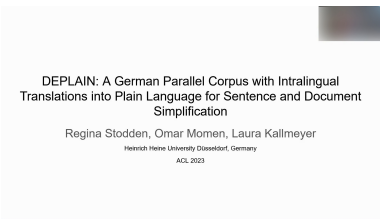
Image	Cascaded FT Output	OmniFusion Output	Reference
	Hallo, willkommen zu unserer Präsentation von **Deplain**, einem neuen Korpus für die Identifizierung deutscher Texte auf Dokumenten- und Satzebene.	Hallo, willkommen zu unserer Präsentation von **DEPLAIN**, einem neuen Korpus für die deutsche Texterkennung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

Table 15: Example illustrating where OmniFusion is better than cascaded variant. Example taken from MCIF segment 1.

Image	SeedX Output	OmniFusion Output	Reference
	AUSGANG	AUSFAHRT	AUSFAHRT
	AUSGANG	AUSGANG	AUSGANG

Table 16: Example illustrating OmniFusion disambiguates source sentence with image context. The word 'EXIT' when translated to German can be 'AUSFAHRT' (as vehicle exit) or 'AUSGANG' (as person exit) and needs the image for disambiguation.