# S2ST-Omni: Hierarchical Language-Aware SpeechLLM Adaptation for Multilingual Speech-to-Speech Translation

**Anonymous ACL submission**

## Abstract

Despite recent advances in speech-to-speech translation (S2ST), it remains difficult to achieve both high translation accuracy and practical flexibility. In this paper, we present *S2ST-Omni*, a compositional S2ST framework that integrates a high-accuracy speech-to-text translation (S2TT) frontend with a modular, plug-and-play text-to-speech (TTS) backend, enabling independent optimization of translation and synthesis. On the S2TT side, we introduce a hybrid adapter that follows a "local-then-global" strategy to bridge the pretrained Whisper encoder and Qwen3 LLM, yielding a hierarchical acoustic-to-semantic abstraction. Building on this bridge, we further propose a hierarchical language-aware architecture that injects source-language information at two complementary levels. At the acoustic level, *Language-Aware Dual-CTC* operates on intermediate adapter features and employs FiLM-style feature modulation with a learnable gate, encouraging the model to learn language-specific but content-faithful acoustic representations. At the linguistic level, *Language-Aware Prompting* dynamically constructs source-language-conditioned prompts that activate language-specific translation knowledge in the LLM. To enable efficient optimization, we design a task-specific progressive fine-tuning strategy that first stabilizes speech-text alignment and then improves translation via LoRA on top of this converged foundation. The TTS backend remains fully modular and can be instantiated with any state-of-the-art synthesizer without retraining the S2TT frontend. Experiments on CVSS-C show that *S2ST-Omni* consistently achieves the best BLEU and ASR-BLEU across French, German, and Spanish to English directions, outperforming strong recent S2ST baselines.

## 1 Introduction

Multilingual speech-to-speech translation (S2ST) aims to directly convert utterances from multiple source languages into target language speech, which plays a pivotal role in cross-language communication, healthcare and education (Kikui et al., 2003; Lee et al., 2022b).

Recent years have witnessed rapid progress in S2ST, with methods broadly falling into two families: end-to-end (E2E) and compositional. E2E approaches, exemplified by Translatotron series (Jia et al., 2019, 2022a) and SeamlessM4T (Barrault et al., 2023), map source speech to target speech within a single unified network. While conceptually attractive, E2E systems must jointly handle speech understanding, cross-lingual translation, and speech generation, which often leads to trade-offs that limit optimization of any individual sub-task. Moreover, the lack of an intermediate textual representation complicates error analysis and correction, and the scarcity of large-scale speech-speech parallel data further constrains scalability (Fang et al., 2024).

To mitigate these issues, compositional S2ST architectures decompose the problem into an S2TT frontend and a TTS backend. For example, DASpeech (Fang et al., 2023) adopts a two-pass design in which a linguistic decoder predicts target-side representations and an acoustic decoder (FastSpeech 2 (Ren et al., 2020)) generates speech conditioned on them. These models improve controllability and latency, but are still trained as direct S2ST systems and heavily rely on costly source speech to target speech pairs. In contrast, ComSpeech (Fang et al., 2024) proposes a genuinely compositional architecture that explicitly integrates pretrained S2TT and TTS models via CTC-based vocabulary adaptors, thereby leveraging abundant S2TT/TTS corpora and exposing an interpretable intermediate text. However, its vocabulary adaptors and phoneme-level embeddings yield relatively shallow, weakly contextualized interfaces, limiting robustness on semantically complex or ambiguous utterances. More critically, existing systems offer little explicit modeling of source language informa-

tion at either the acoustic or linguistic level, despite its central role in accurate multilingual translation.

In this paper, we propose **S2ST-Omni**, a compositional S2ST framework that couples a SpeechLLM-based S2TT frontend with a plug-and-play TTS backend. Our key idea is to bridge the representation gap between a pretrained Whisper-large-v3-based speech encoder (Radford et al., 2023) and a Qwen3 LLM (Yang et al., 2025) through structured adaptation and hierarchical language-aware architecture, thereby better exploiting their cross-modal and multilingual capabilities. First, we introduce a hybrid speech adapter that follows a "local-then-global" strategy: depthwise separable convolutions first aggregate local acoustic patterns, learnable downsampling then shortens the sequence while preserving salient information, and global self-attention finally models long-range semantic dependencies on the compressed features. Second, to explicitly exploit source language information, we propose a **hierarchical language-aware** (HLA) architecture that operates jointly at acoustic and linguistic levels. On the acoustic side, **Language-Aware Dual-CTC** (LA-Dual-CTC) comprises a source-CTC and a target-CTC branch. The target-CTC is a standard CTC head supervising target-language alignment, whereas the source-CTC applies FiLM-based language conditioning (Perez et al., 2018) on intermediate adapter features, using feature-wise affine modulation and a learnable gate to inject language-specific acoustic biases while preserving source-content fidelity. On the linguistic side, **Language-Aware Prompting** (LAP) dynamically constructs prompts based on the detected source language, explicitly marking the language at the LLM input and triggering language-specific translation behavior to resolve cross-lingual ambiguities. Importantly, language identity is obtained without introducing a dedicated classifier: during training we use ground-truth labels, and at inference we reuse Whisper encoder features and its decoder for language prediction. Third, to enable efficient optimization of the S2TT frontend, we present a task-specific **progressive fine-tuning** (PFT) strategy. In Stage I, we freeze the Whisper encoder and Qwen3, and train only the hybrid adapter with LA-Dual-CTC to establish reliable speech-text alignment. In Stage II, we insert LoRA (Hu et al., 2022) adapters into the LLM and continue to update the LA-Dual-CTC-augmented adapter with reduced learning rates, injecting translation capability on top of a converged alignment. On the synthesis side, *S2ST-Omni* can seamlessly integrate any state-of-the-art (SOTA) TTS model (Chen et al., 2024; Peng et al., 2025; Zhou et al., 2025a,b), without retraining the S2TT frontend.

In summary, we introduce *S2ST-Omni*, a compositional S2ST framework that integrates a SpeechLLM-based S2TT frontend with a plug-and-play TTS backend. Our approach bridges Whisper and Qwen3 via a local-then-global hybrid adapter, and injects source-language information through a hierarchical language-aware architecture (LA-Dual-CTC and LAP). To ensure stable optimization, we propose a task-specific progressive fine-tuning strategy that establishes a robust speech–text alignment foundation before refining translation capabilities via LoRA. Experiments on CVSS-C (Jia et al., 2022b) show that *S2ST-Omni* achieves the best BLEU and ASR-BLEU scores among eight strong E2E and compositional baselines on French, German, and Spanish → English translation.

## 2 BACKGROUND

### 2.1 Speech-to-Speech Translation

Existing S2ST systems broadly fall into two paradigms: E2E and compositional. E2E models such as Translatotron series (Jia et al., 2019, 2022a) and SeamlessM4T (Barrault et al., 2023) map source speech directly to target speech within a single network, but must jointly learn speech understanding, translation, and generation, which complicates optimization and typically demands large amounts of speechspeech parallel data. Compositional approaches instead factorize S2ST into an S2TT frontend plus a TTS backend. ComSpeech (Fang et al., 2024) links pretrained S2TT and TTS models via CTC-based vocabulary adapters, while StreamSpeech (Zhang et al., 2024) builds a streaming cascaded system with online S2TT and expressive TTS. Despite their advantages in interpretability and modularity, these systems still rely on relatively shallow adapters and task-specific S2TT training, and they do not explicitly model source-language information at the acoustic or linguistic level, leaving room for more language-aware and SpeechLLM-centric designs.

### 2.2 Speech Large Language Models

Recent SpeechLLM frameworks extend text LLMs to audio either via discrete tokens (e.g., SpeechGPT (Zhang et al., 2023), AudioPaLM (Rubenstein et al., 2023)) or con-

tinuous adapters that connect speech encoders such as Whisper to LLM backbones (e.g., SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), Qwen2-Audio (Chu et al., 2024)). These models achieve strong results on ASR and spoken understanding and can perform speech-to-text translation as part of a multitask setup. However, they are not tailored for high-accuracy multilingual S2TT and lack explicit mechanisms to inject source-language information into either the acoustic representations or the prompting interface, which is precisely the gap our HLA-based Speech-LLM framework is designed to address.

### 2.3 CTC for Speech Processing

Connectionist Temporal Classification (CTC) (Graves et al., 2006) provides a principled framework for sequence-to-sequence learning without explicit alignment supervision, which has been widely adopted as an auxiliary objective to improve speech-text alignment in encoder-decoder models (Kim et al., 2017; Watanabe et al., 2017; Yang et al., 2023). Given an input sequence $\mathbf{x} = (x_1, \ldots, x_T)$ and target sequence $\mathbf{y} = (y_1, \ldots, y_U)$ where $U \leq T$, CTC marginalizes over all valid alignments $\pi$ that collapse to $\mathbf{y}$:

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\pi \in \mathcal{B}^{-1}(\mathbf{y})} P(\pi|\mathbf{x}) \tag{1}$$

where $\mathcal{B}$ denotes the collapsing function that removes blanks and repeated tokens.

## 3 METHODOLOGY

### 3.1 System Overview

As illustrated in Fig. 1, *S2ST-Omni* adopts a compositional architecture that decomposes S2ST into two largely independently optimizable modules: a high-accuracy S2TT frontend and a flexible TTS backend. The S2TT frontend comprises four core components: (i) a Whisper Large-v3 encoder that provides robust multilingual speech representations; (ii) a Qwen3-4B LLM that serves as the translation backbone; (iii) a hybrid speech adapter that bridges Whisper and Qwen3 via a hierarchical "local-then-global" strategy; and (iv) a HLA architecture consisting of an acoustic-level LA-Dual-CTC and a linguistic-level LAP. The TTS backend is kept fully modular and can be instantiated with any SOTA synthesizer.

### 3.2 Hybrid Speech Adapter

Adapting continuous speech representations to an LLM requires modeling information across multiple scales, from phonetic-level acoustic patterns to sentence-level semantics. Conformer-style adapters (Dubey et al., 2024; Xu et al., 2025) that interleave convolution and attention in every layer offer a general solution; however, early global mixing can dilute fine-grained local cues, and computing self-attention on long speech sequences is computationally expensive. At the other extreme, MLP-style adapters (Fang et al.; He et al., 2025) that apply only per-frame linear projections are parameter-efficient but lack explicit temporal modeling, thereby shifting the burden of capturing short-term phonetic patterns onto the LLM, which is primarily designed for higher-level semantic reasoning.

We therefore propose a *Hybrid Adapter* that follows a "local-then-global" hierarchy. Given Whisper encoder outputs $\mathbf{H} \in \mathbb{R}^{T \times d_s}$ with frame length $T$ and hidden size $d_s = 1280$, we first project them into an adapter hidden space $d_h = 1024$ via a linear layer with LayerNorm and GELU, yielding $\mathbf{H}^{(0)} \in \mathbb{R}^{T \times d_h}$. On top of $\mathbf{H}^{(0)}$, we stack two depthwise separable convolution blocks. Each block applies LayerNorm, a depthwise-separable 1D convolution with kernel size $k = 7$ to capture local acoustic patterns, followed by a position-wise feed-forward network (FFN) with expansion ratio 4 and a residual connection.

To shorten the sequence before global attention, we apply a strided 1D convolution with kernel size of 5 and stride of 2 to the final output $\mathbf{H}^{(N)}$, followed by LayerNorm and GELU, obtaining a downsampled representation $\mathbf{H}^{\text{down}} \in \mathbb{R}^{T' \times d_h}$ with $T' = \lceil T/2 \rceil$. The learnable downsampling adaptively preserves translation-relevant information while discarding redundant frames. We then feed $\mathbf{H}^{\text{down}}$ into two standard Transformer blocks, each composed of multi-head self-attention with 4 heads and a FFN (expansion ratio 4), both wrapped with LayerNorm and residual connections, to model long-range semantic dependencies and produce a global representation $\mathbf{H}^{\text{glob}} \in \mathbb{R}^{T' \times d_h}$. Finally, we apply LayerNorm and a linear projection $\mathbf{W}_{\text{out}} \in \mathbb{R}^{d_h \times d_l}$ with $d_l = 3584$ to map $\mathbf{H}^{\text{glob}}$ into the LLM hidden space, yielding $\mathbf{Z} \in \mathbb{R}^{T' \times d_l}$. This sequence $\mathbf{Z}$ is fed to the LLM, while the intermediate features $\mathbf{H}^{\text{down}}$ are reused by LA-Dual-CTC for language-aware auxiliary supervision.

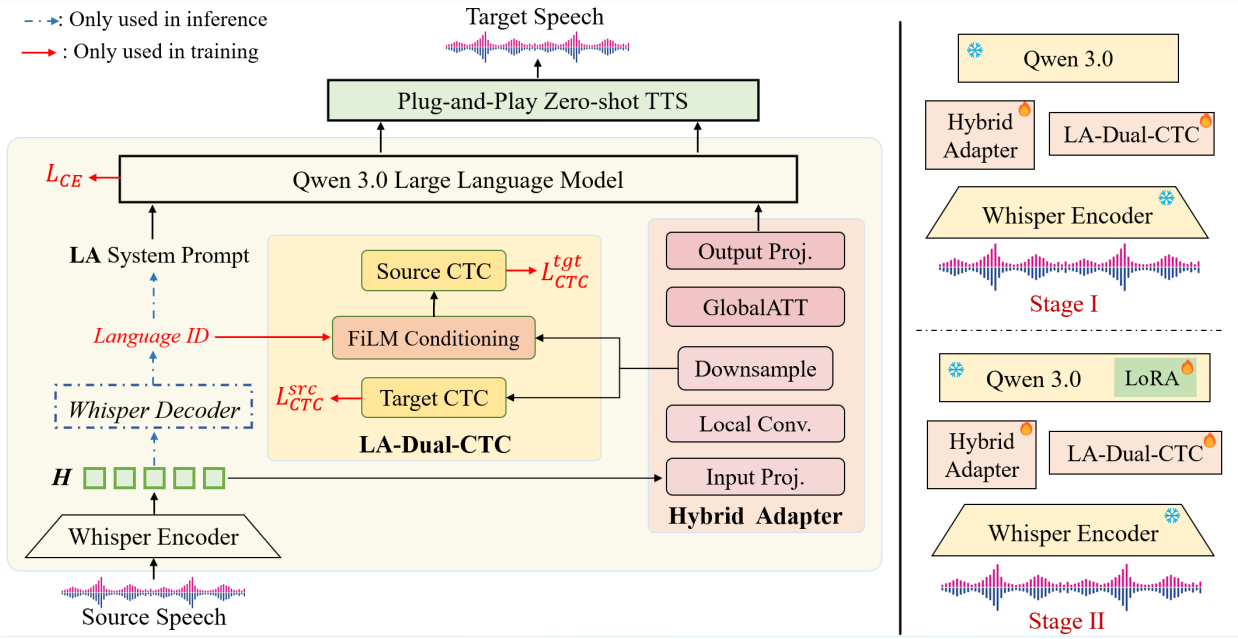


Figure 1: Overall architecture of the proposed S2ST-Omni framework. LA denotes language-aware.

### 3.3 HLA Augmentation Architecture

In multilingual S2ST, we suppose that explicit source-language awareness is crucial for accurate translation, as different languages exhibit distinct syntactic structures (e.g., verb-final German clauses, post-nominal French adjectives), word-order patterns, and phonological systems. To capture these differences, we propose a *HLA augmentation* architecture that injects source-language information at two complementary levels: (i) LAP operates at the LLM input to guide high-level translation behavior; and (ii) LA-Dual-CTC operates on adapter features to modulate low-level acoustic representations. Together, they provide bottom-up and top-down language awareness.

#### 3.3.1 LAP

Pretrained LLMs acquire rich cross-lingual knowledge from large-scale multilingual corpora, but effectively leveraging this knowledge requires explicit cues about the input language and task. LAP provides such cues by dynamically constructing language-aware prompts.

Given a source-language identifier $\ell$ (obtained from Whispers built-in language ID during inference, and from dataset labels during training), we instantiate a prompt $\mathcal{P}(\ell)$ using a simple template: *The following is* [LANG] *speech. Translate it accurately into English.*, where [LANG] is the full language name corresponding to $\ell$ (e.g., `fr` $\rightarrow$ *French*). We encode $\mathcal{P}(\ell)$ with the LLM tokenizer to obtain prompt embeddings $\mathbf{E}(\mathcal{P}(\ell))$, which are concatenated with the speech representation $\mathbf{Z}$ and the target text embeddings $\mathbf{E}(\mathbf{y})$ to form the LLM input sequence. By this means, LAP is able to explicitly declare the source language and steer the LLM towards the appropriate language-pair translation mode, without introducing additional cost.

#### 3.3.2 LA-Dual-CTC

Although LAP operates at the linguistic level, we also aim to encode language-specific inductive biases directly into the speech representation. To this end, we design *LA-Dual-CTC*, which injects source-language information into the CTC heads via FiLM-style conditioning.

Detailed, we apply LA-Dual-CTC to the downsampled intermediate features $\mathbf{H}^{\text{down}}$ from the Hybrid Adapter (Section 3.2) rather than to its final output. The intuition behind is that this design ensures that $\mathbf{H}^{\text{down}}$ both preserves the monotonic alignment structure required by CTC and retains sufficient acoustic detail, while decoupling the CTC objective from the subsequent global-attention layers that are primarily responsible for high-level semantic modeling. Then, we obtain a language embedding $\mathbf{e}_\ell = \text{Embed}(\ell)$ and feed it into a small MLP to generate FiLM parameters $(\boldsymbol{\gamma}_\ell, \boldsymbol{\beta}_\ell) \in \mathbb{R}^{d_h} \times \mathbb{R}^{d_h}$, where $\ell$ denotes the source language. To control the overall strength of language conditioning, we introduce a learnable scalar gate $g$. Empirically, we initialize $g = 0.5$,

**Algorithm 1** Task-Specific PFT

**Require:** Dataset $\mathcal{D} = \{(x, y, \ell)\}$, encoder parameters $\theta_{\text{enc}}$, LLM parameters $\theta_{\text{llm}}$
Initialize Hybrid Adapter $\theta_{\text{adpt}}$, LA-Dual-CTC heads $\theta_{\text{ctc}}$, LoRA parameters $\theta_{\text{lora}}$
Freeze $\theta_{\text{enc}}$ and base $\theta_{\text{llm}}$ in all stages
**Stage I (alignment):** optimize $\theta_{\text{adpt}}, \theta_{\text{ctc}}$
**for** $(x, y, \ell) \sim \mathcal{D}$ **do**
  Compute losses $\mathcal{L}_{\text{CE}}, \mathcal{L}_{\text{CTC}}^{\text{src}}, \mathcal{L}_{\text{CTC}}^{\text{tgt}}$
  Update $\theta_{\text{adpt}}, \theta_{\text{ctc}}$ w.r.t. $\mathcal{L}^{(1)}$
**end for**
**Stage II (translation):** enable $\theta_{\text{lora}}$, reduce lr for $\theta_{\text{adpt}}, \theta_{\text{ctc}}$
**for** $(x, y, \ell) \sim \mathcal{D}$ **do**
  Compute losses $\mathcal{L}_{\text{CE}}, \mathcal{L}_{\text{CTC}}^{\text{src}}, \mathcal{L}_{\text{CTC}}^{\text{tgt}}$
  Update $\theta_{\text{adpt}}, \theta_{\text{ctc}}, \theta_{\text{lora}}$ w.r.t. $\mathcal{L}^{(2)}$
**end for**
**Ensure:** Trained S2TT frontend

which provides a moderate degree of modulation at the start of training and allows the model to adaptively increase or decrease its reliance on language-specific conditioning as learning progresses. The gated FiLM transformation produces language-conditioned features

$$\mathbf{Z}^{\text{src}} = (1 + g\,\boldsymbol{\gamma}_\ell) \odot \mathbf{H}^{\text{down}} + g\,\boldsymbol{\beta}_\ell, \qquad (2)$$

on which the *source* CTC head is applied to produce logits $\mathbf{O}^{\text{src}}$. Because the target language is fixed to English, the *target* CTC head operates directly on $\mathbf{H}^{\text{down}}$ without language conditioning to produce logits $\mathbf{O}^{\text{tgt}}$.

### 3.4 Task-Specific PFT

Optimizing the S2TT frontend jointly over a frozen Whisper encoder, a large pretrained LLM, a newly initialized adapter, and auxiliary CTC heads is challenging due to heterogeneous initialization and optimization dynamics. As illustrated in Algorithm 1, we therefore adopt a task-specific two-stage *PFT* strategy: Stage I primarily builds a robust speech–text alignment under strong dual-CTC supervision, while Stage II injects translation capability into the LLM via parameter-efficient adaptation under weaker CTC regularization.

#### 3.4.1 Stage I: Alignment Foundation

In Stage I, we freeze the Whisper encoder and the base Qwen3 parameters, and train only the Hybrid Adapter and LA-Dual-CTC. The LLM thus acts as a fixed conditional language model, while the adapter is optimized to produce features that are jointly aligned with source and target text.

Let $\mathcal{L}_{\text{CE}}$ denote the autoregressive cross-entropy loss of the LLM decoder, and $\mathcal{L}_{\text{CTC}}^{\text{src}}$ and $\mathcal{L}_{\text{CTC}}^{\text{tgt}}$ denote the CTC losses on the source and target branches of LA-Dual-CTC, respectively. The Stage I objective is

$$\mathcal{L}^{(1)} = \mathcal{L}_{\text{CE}} + \alpha_1\,\mathcal{L}_{\text{CTC}}^{\text{src}} + \beta_1\,\mathcal{L}_{\text{CTC}}^{\text{tgt}}, \qquad (3)$$

where we empirically set $\alpha_1 = 0.1$ and $\beta_1 = 0.2$ to emphasize reliable alignment.

#### 3.4.2 Stage II: Translation Enhancement

In Stage II, we introduce parameter-efficient adaptation on the LLM while continuing to train the Hybrid Adapter and LA-Dual-CTC with reduced learning rates. Concretely, we apply LoRA to the query and value projections in the self-attention modules of Qwen3, with rank $r = 8$, scaling factor $\alpha = 32$, and dropout rate $0.1$, and optimize these LoRA parameters with a higher learning rate than the adapter and CTC heads. To shift the focus towards end-to-end translation quality while retaining alignment signals, we down-weight the CTC losses and use

$$\mathcal{L}^{(2)} = \mathcal{L}_{\text{CE}} + \alpha_2\,\mathcal{L}_{\text{CTC}}^{\text{src}} + \beta_2\,\mathcal{L}_{\text{CTC}}^{\text{tgt}}, \qquad (4)$$

with $\alpha_2 = 0.01$ and $\beta_2 = 0.05$. This two-stage schedule first stabilizes speech–text alignment and then refines translation behavior on top of this converged foundation.

### 3.5 TTS Backend

S2ST-Omni supports plug-and-play integration with any SOTA speech synthesizer as the TTS backend. The intermediate English text produced by the S2TT frontend is passed directly to an external TTS model, without any task-specific coupling. In our experiments, we instantiate the backend with IndexTTS 2 (Zhou et al., 2025a), a zero-shot TTS system that can generate natural and fluent speech. This modular design allows practitioners to select or swap TTS systems according to application requirements without any retraining or modifying the S2TT frontend, reducing development and deployment overhead.

## 4 EXPERIMENTS

### 4.1 Experimental Settings

#### 4.1.1 Datasets

We evaluate S2ST-Omni on the CVSS-C benchmark (Jia et al., 2022b), which provides par-

Table 1: Overall performance comparison among evaluated S2ST methods on CVSS-C.

| Models | FR → EN | | ES → EN | | DE → EN | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ |
| Ground Truth | - | 84.52 | - | 88.54 | - | 75.53 | - | - |
| Translatotron | - | 16.96 | - | 8.72 | - | 1.97 | - | 9.22 |
| Translatotron2 | 28.82 | 26.07 | 25.82 | 22.93 | 18.66 | 16.91 | 24.43 | 21.97 |
| S2UT | - | 22.23 | - | 18.53 | - | 2.99 | - | 14.58 |
| UnitY | - | 27.77 | - | 24.95 | - | 18.74 | - | 23.82 |
| DASpeech | - | 25.03 | - | 21.37 | - | 16.14 | - | 20.85 |
| ComSpeech | 30.72 | 28.15 | 26.51 | 24.80 | 19.41 | 18.16 | 25.55 | 23.70 |
| StreamSpeech | 32.60 | 28.45 | 30.35 | 27.25 | 23.36 | 20.93 | 28.77 | 25.54 |
| Hibiki | - | 30.50 | - | - | - | - | - | - |
| Ours | **35.83** | **33.20** | **37.85** | **35.90** | **33.34** | **31.25** | **35.67** | **33.45** |

Table 2: Effect of the proposed hybrid speech adapter against Conformer-based and MLP-based adapters.

| Model | Fr → En | | Es → En | | De → En | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ |
| Ours | **35.83** | **33.20** | **37.85** | **35.90** | **33.34** | **31.25** | **35.67** | **33.45** |
| w/ Conformer | 34.60 | 31.49 | 36.78 | 34.64 | 32.69 | 30.34 | 34.69 | 32.16 |
| w/ MLP | 32.30 | 28.96 | 34.70 | 32.47 | 30.09 | 27.54 | 32.37 | 29.66 |

allel triplets ⟨source speech, target text, target speech⟩ for 21 source languages translated into English. Following the protocol of (Fang et al., 2024; Zhang et al., 2024) for fair comparison, we evaluate S2ST-Omni on three representative directions: French→English (Fr→En), Spanish→English (Es→En), and German→English (De→En). Source speech is resampled to 16 kHz, and target speech to 22.05 kHz.

#### 4.1.2 Implementation Details

All experiments are implemented in PyTorch on two NVIDIA A6000 GPUs. In Stage I, we train only the Hybrid Adapter and LA-Dual-CTC for 150k steps using AdamW with a cosine schedule, learning rates of $1 \times 10^{-5}$ (adapter) and $5 \times 10^{-5}$ (LA-Dual-CTC), and 1k warmup steps. In Stage II, we attach LoRA modules and continue training for another 150k steps with learning rates $5 \times 10^{-6}$ (adapter), $1 \times 10^{-6}$ (LA-Dual-CTC), and $5 \times 10^{-5}$ (LoRA). We use a batch size of 4 per GPU with gradient accumulation of 6. For LA-Dual-CTC, we employ 8k and 4k SentencePiece vocabularies for the multilingual source and English target branches, respectively, both trained on the source and target text of our training set.

#### 4.1.3 Evaluation

We compare S2ST-Omni against eight strong S2ST baselines: Translatotron and Translatotron 2 (Jia et al., 2019, 2022a), S2UT (Lee et al., 2022a), DASpeech (Fang et al., 2023), UnitY (Inaguma et al., 2023), ComSpeech (Fang et al., 2024), StreamSpeech (Zhang et al., 2024), and Hibiki (Labiausse et al., 2025).

For ASR-BLEU, we first transcribe the generated speech with a pretrained wav2vec 2.0 ASR model[1], and then compute BLEU (Papineni et al., 2002) using SacreBLEU[2].

### 4.2 Main Results

Table 1 summarizes the overall results on CVSS-C. *S2ST-Omni* achieves the best performance across all three directions among the evaluated systems. Compared with the strongest compositional baseline StreamSpeech, our model improves the average BLEU score from 28.77 to 35.67 and the average ASR-BLEU from 25.54 to 33.45, corresponding to relative gains of approximately $+24\%$ BLEU and $+31\%$ ASR-BLEU. These results indicate that S2ST-Omni produces not only more accurate textual translations, but also synthesized speech whose ASR transcriptions are substantially closer to the ground-truth references.

Breaking down the results by language pair, *S2ST-Omni* achieves consistent improvements. On the relatively more challenging **De→En** track,

[1] https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_vox_960h_pl.pt

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

our system reaches 33.34 BLEU and 31.25 ASR-BLEU, compared with 23.36 and 20.93 for StreamSpeech, yielding large relative gains of about $+43\%$ BLEU and $+49\%$ ASR-BLEU. For **Es→En**, *S2ST-Omni* reaches 37.85 BLEU and 35.90 ASR-BLEU, improving over StreamSpeech (30.35 / 27.25) by approximately $+25\%$ and $+32\%$, respectively. On **Fr→En**, our model obtains 35.83 BLEU and 33.20 ASR-BLEU, which corresponds to relative gains of about $+10\%$ BLEU and $+17\%$ ASR-BLEU over StreamSpeech, and it also surpasses the recently proposed Hibiki that is specifically optimized for ASR-BLEU on this language pair.

Overall, we attribute these consistent gains to the combination of the LA-Dual-CTCaugmented hybrid adapter and the hierarchical language-aware architecture. The hybrid adapter provides a stronger speech-to-LLM interface by first consolidating local acoustic structure and then modeling long-range dependencies on a downsampled sequence, while LA-Dual-CTC and LAP inject complementary acoustic- and linguistic-level language cues. Together with the progressive fine-tuning schedule, these components enable *S2ST-Omni* to more effectively exploit Whisper and Qwen3 for multilingual S2TT and S2ST, yielding best performance on CVSS-C.

## 4.3 Ablation Study

### 4.3.1 Effect of Hybrid Speech Adapter

Table 2 compares our hybrid adapter with Conformer- and MLP-based variants. Replacing the hybrid adapter with the Conformer design results in a consistent drop across all directions: the average BLEU score decreases from 35.67 to 34.69 (about $2.8\%$ relative), and the average ASR-BLEU from 33.45 to 32.16 (about $3.9\%$), with similar trends for each language pair. In contrast, the MLP adapter causes a substantially larger degradation, reducing the average BLEU/ASR-BLEU to 32.37/29.66, corresponding to roughly $9\%$ and $11\%$ relative drops, respectively. Overall, these results suggest that MLP-style adaptation is insufficient to bridge Whisper and Qwen3 for S2ST, and that densely interleaving convolution and attention (Conformer-style) is also suboptimal, highlighting the effectiveness of the proposed hybrid adapter.

### 4.3.2 Effect of HLA Architecture

Table 3 examines the contribution of our HLA architecture. Removing either LAP or LA-Dual-CTC leads to small but consistent performance drops: removing LAP reduces the average BLEU/ASR-BLEU from 35.67/33.45 to 35.25/32.89, while removing LA-Dual-CTC yields 35.16/32.72. In contrast, removing both components (w/o HLA) causes a much larger degradation, bringing the averages down to 33.68 BLEU and 30.95 ASR-BLEU, which corresponds to roughly $5.6\%$ and $7.5\%$ relative drops. These results indicate that linguistic-level prompting (LAP) and acoustic-level FiLM conditioning (LA-Dual-CTC) are complementary: each provides modest gains on its own, while jointly encoding source-language information at both acoustic and linguistic levels yields a stronger inductive bias for multilingual S2ST.

### 4.3.3 Effect of Task-Specific PFT

Table 4 evaluates the effect of the proposed two-stage PFT strategy. The Stage-I-only variant (w/o PFT-II), which omits LoRA-based LLM adaptation, exhibits the largest degradation: the average BLEU/ASR-BLEU drops from 35.67/33.45 to 33.33/30.63, with a particularly pronounced decline on De→En (33.34/31.25 → 31.23/28.74). Collapsing the two phases into a single-stage training (w/o PFT) mitigates the drop but still underperforms the full schedule, reducing the average BLEU/ASR-BLEU to 34.37/31.81 (about $3.6\%$ and $4.9\%$ relative drops). These results suggest that optimizing alignment and translation simultaneously can introduce harmful gradient interference, whereas the proposed two-stage schedule: first emphasizing dual-CTC alignment, then refining translation with LoRA-based LLM adaptation, provides a more stable and effective optimization path for the S2TT frontend.

## 4.4 TTS Backend Analysis

Table 5 examines how the choice of TTS backend affects synthesis quality when the S2TT frontend of *S2ST-Omni* is fixed and only the TTS module is swapped. Overall, the ASR-BLEU scores are relatively stable across five SOTA TTS systems: the average gap between the best configuration (IndexTTS2, 33.45) and the weakest (VoxCPM1.5, 32.09) is within roughly $1.4$ points (about $4\%$ relative), and all backends fall into a narrow band on each language pair. Notably, even the weakest backend (VoxCPM1.5) achieves an average ASR-BLEU of 32.09, which remains higher than StreamSpeech under the same evaluation setting.

These results indicate that, given a strong SpeechLLM-based S2TT frontend, the synthesis

Table 3: Effect of the proposed HLA architecture.

| Model | Fr → En | | Es → En | | De → En | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ |
| Ours | **35.83** | **33.20** | **37.85** | **35.90** | **33.34** | **31.25** | **35.67** | **33.45** |
| w/o LAP | 35.25 | 32.36 | 37.37 | 35.30 | 33.13 | 31.01 | 35.25 | 32.89 |
| w/o LA-Dual-CTC | 35.16 | 32.11 | 37.22 | 35.09 | 33.10 | 30.97 | 35.16 | 32.72 |
| w/o HLA | 33.49 | 30.26 | 35.96 | 33.77 | 31.57 | 28.82 | 33.68 | 30.95 |

Table 4: Effect of the task-specific PFT strategy. "w/o PFT-II" denotes removing Stage II LoRA-based LLM adaptation; "w/o PFT" represents training in a single stage without the proposed two-stage schedule.

| Model | Fr → En | | Es → En | | De → En | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ | BLEU↑ | ASR-BLEU↑ |
| Ours | **35.83** | **33.20** | **37.85** | **35.90** | **33.34** | **31.25** | **35.67** | **33.45** |
| w/o PFT-II | 33.17 | 29.78 | 35.58 | 33.36 | 31.23 | 28.74 | 33.33 | 30.63 |
| w/o PFT | 34.24 | 31.12 | 36.41 | 34.24 | 32.46 | 30.08 | 34.37 | 31.81 |

Table 5: Effect of different TTS backends on the ASR-BLEU metric.

| Models | Fr→En | Es→En | De→En | Avg. |
|---|---|---|---|---|
| IndexTTS2 | 33.20 | 35.90 | 31.25 | 33.45 |
| FireRedTTS2 | 31.55 | 35.18 | 30.94 | 32.56 |
| CosyVoice3 | 31.56 | 35.15 | 30.36 | 32.36 |
| ZipVoice | 31.48 | 35.08 | 30.46 | 32.34 |
| VoxCPM1.5 | 31.35 | 34.82 | 30.20 | 32.09 |

performance is largely insensitive to the specific choice of modern TTS backend. In practice, *S2ST-Omni* can thus be paired with diverse off-the-shelf TTS systems. Practitioners can choose the backend based on latency, speaker style, or deployment constraints, without redesigning or retraining the translation module.

### 4.5 Latency Analysis

As shown in Table 6, we report the latency of the S2TT frontend of *S2ST-Omni* on the CVSS-C test sets. For each utterance, we measure the latency of (i) Whisper encoding, (ii) Whisper decoding for language identification, (iii) the hybrid adapter, and (iv) LLM prefill and generation with batch size 1 on a single RTX A5000 GPU.

Table 6: Average latency (ms) of the S2TT frontend of *S2ST-Omni* on CVSS-C (Fr/Es/De) test sets.

| Module | Whisper Enc. | LID Dec. | Adapter | LLM | Total |
|---|---|---|---|---|---|
| Latency | 118.4 | 22.8 | 6.8 | 555.8 | 703.8 |

On average, the S2TT frontend requires about 704 ms per utterance across Fr/Es/De test sets. LLM decoding is the dominant cost (roughly 79% of the total time), Whisper encoding is the second largest contributor (about 17%). Language-ID decoding and the hybrid adapter together contribute only about 4.2% to the overall latency. Additionally, since the TTS backend in *S2ST-Omni* is fully modular, practitioners can select a synthesizer that best matches their latency and quality requirements without modifying or retraining the S2TT frontend.

## 5 Conclusion

We presented *S2ST-Omni*, an accurate and flexible hierarchical language-aware framework for multilingual speech-to-speech translation. By coupling a SpeechLLM-based S2TT frontend with a plug-and-play TTS backend, S2ST-Omni decouples translation from synthesis, enabling independent optimization of each module while preserving an interpretable text intermediate. At the core of the frontend, a hybrid "local-then-global" speech adapter, augmented with LA-Dual-CTC and LAP, provides a strong speechLLM interface that jointly models local acoustic patterns, long-range semantics, and explicit source-language cues. A task-specific two-stage PFT strategy further stabilizes training by first establishing robust speech-text alignment and then improving translation capability via LoRA-based LLM adaptation. Experiments on CVSS-C demonstrate that S2ST-Omni outperforms eight strong E2E and compositional S2ST baselines on Fr→En, Es→En, and De→En, achieving the best BLEU and ASR-BLEU scores. Ablation studies further validate the effectiveness and contribution of each proposed component.

## 6 Limitations

While S2ST-Omni achieves strong performance on CVSS-C, several limitations remain. First, our empirical study is limited to three high-resource source languages (French, Spanish, German) with English as the only target language, and to a single benchmark. This leaves open questions about the robustness of the proposed hybrid adapter, HLA architecture, and PFT strategy in low-resource settings, many-to-many S2ST, or more diverse domains and speaking styles. Second, the current instantiation relies on large backbone models (Whisper Large-v3 and Qwen3-4B), which require nontrivial computational resources for both training and inference. While the hybrid adapter and PFT are designed to be parameter-efficient, deploying S2ST-Omni on edge devices or under strict latency/energy constraints may remain challenging and may require model compression or distillation. Third, our hierarchical language-aware design assumes access to a reliable source-language identifier. At inference, we reuse Whispers encoder and decoder for language prediction, which adds negligible latency and introduces a small amount of noise. In our experiments, language ID accuracy is 99.76% (De), 99.67% (Fr), and 99.80% (Es). Finally, our evaluation focuses on automatic metrics (BLEU and ASR-BLEU) and does not include human judgments of translation adequacy, naturalness, or speaker similarity, nor a detailed analysis of end-to-end latency and computational cost across deployment settings. A more comprehensive evaluation along these dimensions is left for future work.

## References


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*.

Sijing Chen, Yuan Feng, Laipeng He, Tianwei He, Wendi He, Yanni Hu, Bin Lin, Yiting Lin, Yu Pan, Pengfei Tan, and 1 others. 2024. Takin: A cohort of superior quality zero-shot speech generation models. *arXiv preprint arXiv:2409.12139*.

Yunfei Chu and 1 others. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Yunfei Chu and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models, 2024. *URL https://arxiv. org/abs/2409.06666*.

Qingkai Fang, Shaolei Zhang, Zhengrui Ma, Min Zhang, and Yang Feng. 2024. Can we achieve high-quality direct speech-to-speech translation without parallel speech data? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7264–7277, Bangkok, Thailand. Association for Computational Linguistics.

Qingkai Fang, Yan Zhou, and Yang Feng. 2023. Daspeech: Directed acyclic transformer for fast and high-quality speech-to-speech translation. *Advances in Neural Information Processing Systems*, 36:72604–72623.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML*.

Yingxu He, Zhuohan Liu, Geyu Lin, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy Chen, and Aiti Aw. 2025. Meralion-audiollm: Advancing speech and language understanding for singapore. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 22–30.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2023. UnitY: Two-pass direct speech-to-speech translation with discrete units. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15655–15680, Toronto, Canada. Association for Computational Linguistics.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022a. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022b. CVSS corpus and massively

multilingual speech-to-speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France. European Language Resources Association.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*.

Gen-ichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *INTERSPEECH*, pages 381–384.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *ICASSP*.

Tom Labiausse, Laurent Mazaré, Edouard Grave, Alexandre Défossez, and Neil Zeghidour. 2025. High-fidelity simultaneous speech-to-speech translation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 32116–32129. PMLR.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022a. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022b. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, and 1 others. 2025. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *AAAI*.

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.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuguang Yang, Yu Pan, Jingjing Yin, Jiangyu Han, Lei Ma, and Heng Lu. 2023. Hybridformer: Improving squeezeformer with hybrid attention and nsr mechanism. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Dong Zhang and 1 others. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2024. StreamSpeech: Simultaneous speech-to-speech translation with multi-task learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8964–8986, Bangkok, Thailand. Association for Computational Linguistics.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025a. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, and 1 others. 2025b. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*.

# A Example Appendix

This is an appendix.