GenTranslate: Large Language Models are Generative Multilingual Speech and Machine Translators

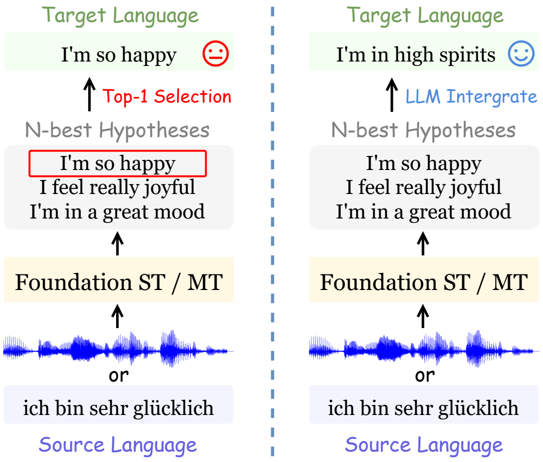
Yuchen Hu[1] **Chen Chen**[1] **Chao-Han Huck Yang**[2,3]
Ruizhe Li[4] **Dong Zhang**[5] **Zhehuai Chen**[3] **Eng Siong Chng**[1]
[1]Nanyang Technological University [2]Georgia Institute of Technology [3]NVIDIA
[4]University of Aberdeen [5]Fudan University

Abstract

Recent advances in large language models (LLMs) have stepped forward the development of multilingual speech and machine translation by its reduced representation errors and incorporated external knowledge. However, both translation tasks typically utilize beam search decoding and top-1 hypothesis selection for inference. These techniques struggle to fully exploit the rich information in the diverse N-best hypotheses, making them less optimal for translation tasks that require a single, high-quality output sequence. In this paper, we propose a new generative paradigm for translation tasks, namely "GenTranslate", which builds upon LLMs to generate better results from the diverse translation versions in N-best list. Leveraging the rich linguistic knowledge and strong reasoning abilities of LLMs, our new paradigm can integrate the rich information in N-best candidates to generate a higher-quality translation result. Furthermore, to support LLM finetuning, we build and release a HypoTranslate dataset that contains over 592K hypotheses-translation pairs in 11 languages. Experiments on various speech and machine translation benchmarks (*e.g.*, FLEURS, CoVoST-2, WMT) demonstrate that our GenTranslate significantly outperforms the state-of-the-art model[1].

1 Introduction

Recent advances in large language models (LLMs) have attracted a surge of research interest due to their strong abilities in logical reasoning and language generation (OpenAI, 2022, 2023; Touvron et al., 2023a,b). These models have achieved surprisingly wide-ranging success across various natural language processing (NLP) tasks (Brown et al., 2020; Wang et al., 2022; Wei et al., 2022a,b; Ouyang et al., 2022).



(a) Typical Translation (b) GenTranslate (ours)

Figure 1: Illustration of (a) Typical seq2seq translation with beam search decoding and top-1 hypothesis selection, (b) our "GenTranslate" with LLM integration.

In the realm of NLP, the translation tasks, which encompasses speech and machine translation (ST & MT), hold significant practical importance for global communication. Similar to other NLP tasks, translation tasks also gain a notable progress thanks to the recent advancement of LLMs (Zhang et al., 2023a; Lyu et al., 2023). In the domain of speech translation, Whisper (Radford et al., 2023) demonstrates superior performance by collecting 680K-hour data for web-scale model training. AudioPaLM2 (Rubenstein et al., 2023) integrates both text- and speech-based language models into a unified architecture to process and generate text and speech, thereby augmenting speech translation performance to a great extent. On the other hand, LLMs also show remarkable ability in machine translation. NLLB (Costa-jussà et al., 2022) is the first to extend LLMs' linguistic capability to over 200 languages. BigTranslate (Yang et al., 2023b) is finetuned on LLaMA (Touvron et al., 2023a) with multilingual instruction tuning, which achieves comparable performance to ChatGPT (OpenAI, 2022) and Google Translate. Most recent work

[1] This work is open sourced at: `https://github.com/YUCHEN005/GenTranslate`

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 74–90
August 11-16, 2024 ©2024 Association for Computational Linguistics

proposes SeamlessM4T (Barrault et al., 2023a), a foundational multilingual and multitask model that can translate across speech and text, which achieves the state-of-the-art on both ST and MT tasks on various public datasets.

Despite the superior performance, most existing translation models employ the typical beam search algorithm for inference and select the top-1 hypothesis as final output (see Fig. 1 (a)), following that in automatic speech recognition (ASR) (Tsunoo et al., 2021). However, this strategy discards the 2 to N-best hypotheses that could be advantageous to the generation of ground-truth translation. As illustrated in Fig. 2, the discarded 2 to N-best hypotheses contain abundant semantic information that is the key to composite the ground-truth utterance, while the 1-best hypothesis lacks this part of information. As a result, the typical top-1 hypothesis selection is sub-optimal to the translation tasks that require a single informative and high-quality output sequence (Li et al., 2022; Xiao et al., 2022).

Inspired by the recent works on LLMs-enhanced ASR (Ma et al., 2023b; Chen et al., 2023; Yang et al., 2023a; Radhakrishnan et al., 2023), we propose a new generative paradigm for translation tasks, namely GenTranslate (see Fig. 1 (b)). Leveraging the rich linguistic knowledge and strong reasoning ability of LLMs, our paradigm integrates the diverse translation versions in the N-best list from foundation model to generate a higher-quality translation result. Furthermore, in order to support LLM finetuning, we also build and release a HypoTranslate dataset that contains over 592K pairs of N-best hypotheses and ground-truth translation in 11 languages. Experimental evidence on various ST and MT benchmarks (*e.g.*, FLEURS, CoVoST-2, WMT) demonstrate that our proposed GenTranslate significantly outperforms the state-of-the-art model with efficient LLM finetuning.

Our contributions are summarized as follows:

- We propose GenTranslate, a new generative paradigm for translation tasks that leverages LLMs to generate higher-quality translation results from the diverse N-best hypotheses decoded from foundation translation model.
- We release a HypoTranslate dataset to support LLM finetuning, which contains over 592K pairs of N-best hypotheses and ground-truth translation in 11 languages.
- Experiments on various ST and MT benchmarks show that our GenTranslate significantly outperforms the state-of-the-art model.

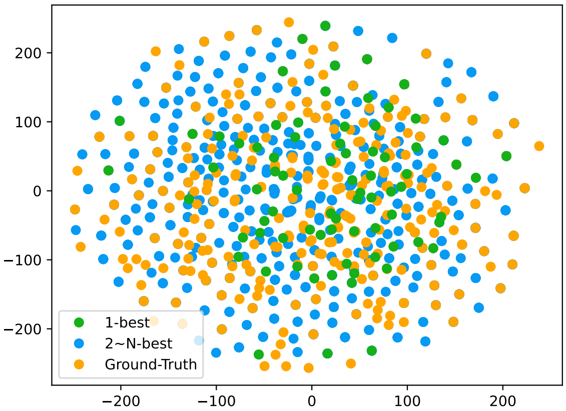


Figure 2: t-SNE visualization of the n-gram tokens (n=1,2,3) in ST 1-best hypothesis (green), 2 to N-best hypotheses (blue), and the ground-truth translation (orange), where the text embeddings are extracted using SBERT (Reimers and Gurevych, 2019). It indicates that the 2 to N-best hypotheses contain richer information than 1-best for generating ground-truth translation.

2 Related Work

2.1 Large Language Models

There is recently a surge of research interests in Transformer-based large language models, such as ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a,b). Benefiting from the giant model size and oceans of training data, LLMs can understand better the linguistic structures and semantic meanings behind raw text, which thus shows remarkable performance on a wide range of natural language processing (NLP) tasks (Brown et al., 2020; Wei et al., 2022a; Ouyang et al., 2022). Thereafter, with techniques like in-context learning (Xie et al., 2021) and efficient finetuning (Hu et al., 2021; Yang et al., 2021b), LLMs further show powerful ability on downstream generative and reasoning tasks (Lampinen et al., 2022; Yang et al., 2023a; Hu et al., 2023b; Zhang et al., 2023b). Our proposed GenTranslate is exactly inspired by the promising generative ability of LLMs.

2.2 Speech and Machine Translation

The advancement of LLMs has notably enhanced the capabilities of translation tasks. In the domain of speech translation (Liu et al., 2021), Whisper (Radford et al., 2023) demonstrates commendable effectiveness, leveraging extensive web-scale data. AudioPaLM2 (Rubenstein et al., 2023) integrates text- and speech-based language models,

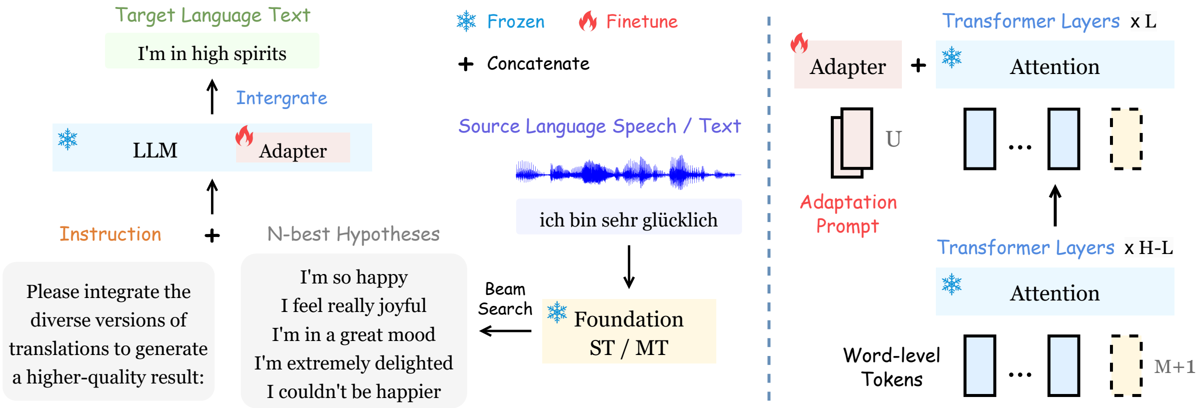


Figure 3: **Left:** Overview of the GenTranslate paradigm (*e.g.*, De→En). **Right:** Details of efficient LLM finetuning.

thereby augmenting the speech translation performance. In the context of machine translation, NLLB (Costa-jussà et al., 2022), a model finetuned on LLMs, extends its linguistic range to over 200 languages. Additionally, BigTranslate (Yang et al., 2023b) utilizes instruction tuning to enhance the translation capabilities of LLMs. The most recent innovation, SeamlessM4T (Barrault et al., 2023a), represents a highly-unified model capable of fluid translation between speech and text, setting new benchmarks in both ST and MT tasks. However, it is noteworthy that the majority of these methodologies rely on beam search decoding (Yang et al., 2021a; Hu et al., 2023a) and top-1 hypothesis selection for inference. How to leverage N-best hypotheses to deliver better translation result remains to be an open question.

2.3 LLMs-Enhanced ASR

Recent works investigate LLMs to enhance the ASR output by error correction (Ma et al., 2023a; Chen et al., 2023), which serves as a post-processing technique to improve the recognition result (Leng et al., 2021). In particular, they leverage LLM finetuning (Zhang et al., 2023b) and in-context learning (Wang et al., 2023) to correct the wrongly recognized tokens in hypotheses by second-pass reasoning, which achieves promising improvement. Inspired by them, in this work we leverage LLMs to integrate the diverse translation versions in N-best list to generate a informative and higher-quality translation result.

3 Methodology

In this section, we introduce the proposed method. First, we describe the latest foundational translation model, SeamlessM4T, which we employ for beam search decoding and hypotheses generation (§3.1). Then, we introduce our LLMs-based GenTranslate paradigm by N-best hypotheses integration (§3.2). Finally, we present the details of our released HypoTranslate dataset for GenTranslate training (§3.3).

3.1 Foundational Translation Model: SeamlessM4T

Recent work (Barrault et al., 2023a,b) proposes SeamlessM4T[2] (Massively Multilingual & Multimodal Machine Translation), a single Transformer-based (Vaswani et al., 2017) model that supports speech-to-speech translation, speech-to-text translation, text-to-speech translation, text-to-text translation, and automatic speech recognition for up to 100 languages. During development process, it is firstly pre-trained on 1 million hours of speech data by self-supervised learning, and it is then finetuned on a 406K-hour multimodal corpus of automatically aligned speech translations named SeamlessAlign. Experiments show that SeamlessM4T yields superior performance on all of the five supported tasks. In particular, it has achieved the state-of-the-art on both ST and MT tasks in terms of BLEU score on various public benchmarks.

Considering its effectiveness, generality and popularity, we employ SeamlessM4T as the foundation model for both speech and machine translation in our system, as depicted in the left part of Fig. 3. Given an input speech S^{src} or text T^{src} in source language (*e.g.*, German), SeamlessM4T translates it into target language (*e.g.*, English) text by beam search decoding, which generates N-best hypotheses list $\mathcal{T}_N^{\text{tgt}} = \{T_1^{\text{tgt}}, T_2^{\text{tgt}}, \cdots, T_N^{\text{tgt}}\}$.

[2] https://github.com/facebookresearch/seamless_communication

3.2 GenTranslate

3.2.1 Overall Framework

To solve the information loss in typical top-1 hypothesis selection, we leverage LLMs to generate a final translation result based on the decoded N-best hypotheses. Since each candidate in N-best list represents one unique version of translation for source language input, our GenTranslate can integrate their rich information to generate a higher-quality translation result, thanks to the strong linguistic and reasoning ability of LLMs. This new generative paradigm can be formulated as:

$$T^{\text{tgt}} = \mathcal{M}_{\text{GT}}(\mathcal{T}_N^{\text{tgt}}, \mathcal{I}), \tag{1}$$

where $\mathcal{I}$ is a proper instruction for LLM prompting. The goal of GenTranslate is to learn a mapping $\mathcal{M}_{\text{GT}}$ from N-best hypotheses to the true translation. Following typical sequence-to-sequence learning strategy, we employ the ground-truth translation $T^{\text{tgt*}}$ as supervision signal and optimize the LLM to learn $\mathcal{M}_{\text{GT}}$ in an auto-regressive manner. The cross-entropy-based training loss is defined as:

$$\mathcal{L}_{\text{GT}} = \sum_{l=1}^{L} -\log \mathbb{P}_\theta(t_l^{\text{tgt*}} | t_{l-1}^{\text{tgt*}}, \cdots, t_1^{\text{tgt*}}; \mathcal{T}_N^{\text{tgt}}, \mathcal{I}), \tag{2}$$

where $t_l^{\text{tgt*}}$ is the l-th token of $T^{\text{tgt*}}$, L denotes the sequence length, and θ denotes the learnable parameters in LLM (*i.e.*, adapter).

3.2.2 Efficient LLM Finetuning

Considering the giant scale of LLMs, we adopt the popular efficient finetuning strategy, LLaMA-Adapter (Zhang et al., 2023b), which is comparable to LoRA tuning (§4.3.4). As shown in Fig. 3 (right), it inserts a set of learnable adaptation prompts into the top-L of total H Transformer layers in a pre-trained LLM to learn high-level semantics. Denote the prompt for l-th layer as $\mathcal{P}_l \in \mathbb{R}^{U \times D}$, where U is prompt length and D is embedding size.

Assume we gain M tokens including instruction and already generated response, *i.e.*, $T_l \in \mathbb{R}^{M \times D}$, now we aim to predict the $(M+1)$-th token as response. The learnable adaptation prompt is concatenated with T_l as prefix, *i.e.*, $[\mathcal{P}_l; T_l] \in \mathbb{R}^{(U+M) \times D}$, which provides learned instruction knowledge to guide the subsequent response generation.

Furthermore, considering the prompt $\mathcal{P}_l$ is randomly initialized and thus could disturb the LLM tuning at early training stage, a zero-initialized attention mechanism is devised to mitigate such disturbance. Denote the current M-th token as $T_l^{(M)} \in \mathbb{R}^{1 \times D}$, in attention there are three projection layers to generate query, key and value:

$$\begin{aligned} Q_l &= \text{Linear}_q(T_l^{(M)}), \\ K_l &= \text{Linear}_k([\mathcal{P}_l; T_l]), \\ V_l &= \text{Linear}_v([\mathcal{P}_l; T_l]), \end{aligned} \tag{3}$$

Then the attention score is calculated as $A_l = Q_l \cdot K_l / \sqrt{D} \in \mathbb{R}^{1 \times (U+M)}$, which captures the correlation between current token and the history tokens as well as prompts to predict the next token. Therefore, it can be split into two parts accordingly:

$$A_l = [A_l^{\mathcal{P}}; A_l^{T}]^T, \tag{4}$$

where $A_l^{\mathcal{P}} \in \mathbb{R}^{U \times 1}$ is the attention score of U adaptation prompts and $A_l^T \in \mathbb{R}^{M \times 1}$ is that of M history tokens. Since the adaptation prompts are randomly initialized, their attention scores may cast disturbance on next-token prediction at early training stage. To this end, a learnable gating factor g_l with zero initialization is introduced to adaptively control the weight of prompt in attention:

$$A_l^g = [g_l \cdot \text{softmax}(A_l^{\mathcal{P}});\ \text{softmax}(A_l^T)]^T, \tag{5}$$

Finally, the attention output of l-th Transformer layer is obtained with a linear projection:

$$O_l^{(M)} = \text{Linear}_o(A_l^g \cdot V_l) \in \mathbb{R}^{1 \times D}, \tag{6}$$

It is then employed to predict the next token $T_l^{(M+1)}$ as response. The zero-initialization mechanism yields an effective trade-off between the pre-trained knowledge of LLM and the learned instructional knowledge through adaptation prompt.

3.3 HypoTranslate Dataset

In order to support the LLM finetuning for GenTranslate, we release a HypoTranslate dataset that contains over 592K pairs of N-best hypotheses and ground-truth translation in 11 languages. In particular, we use the state-of-the-art SeamlessM4T-Large as foundation translation model to decode N-best hypotheses from input speech by beam search algorithm, where the beam size N is set to 5. Specifically, for ST task we investigate two popular pipelines in literature, *i.e.*, end-to-end ST and cascaded ASR+MT. Thanks to the universal ability of SeamlessM4T on ST, ASR and MT tasks, we only need one model to build above two pipelines.

To build HypoTranslate dataset, we select several public ST and MT corpora in both X→En and

X→En	Ar	Cy	De	El	Es	Fa	Fr	Hi	It	Ja	Pt	Ta	Uk	Vi	Zh	Avg.
End-to-end ST Methods																
Whisper-Large V2 (2023)	25.5	13.0	34.6	23.7	23.3	19.6	32.2	22.0	23.6	18.9	38.1	9.2	29.4	20.4	18.4	23.5
AudioPaLM2 (2023)*	29.0	7.2	38.7	18.8	26.9	25.7	36.5	21.7	27.8	11.1	38.4	15.0	26.9	15.6	21.3	24.0
SeamlessM4T-Large (2023a)	32.8	31.7	35.8	25.6	25.0	28.2	33.1	26.3	25.0	17.0	38.9	16.0	30.2	21.6	19.8	27.1
GenTranslate (ours)	**34.6**	**33.6**	**39.2**	**29.4**	**29.8**	**30.5**	**37.0**	**28.3**	**29.7**	**18.6**	**43.0**	**17.4**	**33.9**	**24.1**	**21.7**	**30.1**
SeamlessM4T-Large-V2 (2023b)†	34.7	34.9	37.1	27.3	25.4	30.3	33.7	28.5	26.5	19.5	38.5	22.1	33.2	25.7	23.0	29.4
GenTranslate-V2 (ours)	**37.6**	**36.8**	**40.7**	**31.5**	**29.9**	**33.4**	**37.8**	**30.4**	**31.2**	**21.0**	**43.0**	**23.4**	**36.2**	**27.2**	**25.0**	**32.3**
Cascaded ASR+MT Methods																
Whisper + NLLB-3.3b (2022)	35.5	29.6	40.5	31.1	30.9	28.2	39.7	26.7	30.0	24.7	44.3	20.0	35.3	26.4	25.4	31.2
SeamlessM4T (ASR+MT) (2023a)	38.9	37.0	39.7	29.0	27.7	34.1	37.7	33.9	28.9	21.7	42.3	23.7	34.0	24.9	24.4	31.9
GenTranslate (ours)	**39.9**	**39.4**	**41.6**	**32.8**	**31.2**	**35.9**	**40.6**	**34.9**	**32.1**	**22.8**	**45.0**	**24.1**	**36.9**	**27.4**	**25.7**	**34.0**
SeamlessM4T-V2 (ASR+MT) (2023b)†	39.2	36.8	39.1	29.4	26.7	33.9	35.7	32.9	29.3	22.5	43.2	25.4	34.8	29.7	25.9	32.3
GenTranslate-V2 (ours)	**40.0**	**39.1**	**40.9**	**33.8**	**30.0**	**35.4**	**40.0**	**33.0**	**31.6**	**23.7**	**44.2**	**26.4**	**37.1**	**30.9**	**26.9**	**34.2**

Table 1: Speech translation results on FLEURS **X→En** test sets in terms of BLEU score, where more results on chrF++ metric (Popović, 2017) are in Table 16. We use **bold** to denote surpassing SeamlessM4T baseline, and use underline to denote the state-of-the-art. The baseline methods are introduced in §B.3. * denotes reported by original paper, or else it denotes reproduced by ourselves (same for Table 2 to 5). † denotes the most latest baseline[3].

X→En	Fr	De	Ca	Es	Ru	Zh	Nl	Tr	Et	Mn	Ar	Lv	Sl	Ja	Id	Avg.
End-to-end ST Methods																
XLS-R-2b (2021)*	37.6	33.6	33.8	39.2	39.5	9.4	31.7	16.7	11.1	1.6	17.1	19.5	19.6	3.5	16.5	22.0
Whisper-Large V2 (2023)	35.5	35.0	31.0	39.6	42.3	16.9	40.2	27.5	14.0	0.2	38.5	13.0	16.3	24.7	47.3	28.1
ComSL-Large (2023)*	38.8	36.0	35.3	40.4	49.2	21.4	39.7	33.6	19.2	2.9	41.4	21.3	31.6	21.3	46.6	31.9
AudioPaLM2 (2023)*	44.8	43.4	38.4	44.2	55.6	25.5	48.3	41.0	30.0	7.6	48.7	35.0	42.6	25.9	56.2	39.1
SeamlessM4T-Large (2023a)	41.3	38.8	38.4	41.1	48.6	20.9	41.1	31.2	26.3	7.5	45.0	26.5	37.6	21.8	51.4	34.5
GenTranslate (ours)	**41.7**	**39.2**	**38.7**	**42.0**	**50.1**	**21.6**	**42.1**	**33.5**	**28.2**	**8.7**	**49.7**	**30.3**	**38.2**	**22.9**	**54.3**	**36.1**
SeamlessM4T-Large-V2 (2023b)	42.4	40.0	39.0	42.9	53.6	22.4	42.7	33.2	26.9	8.6	46.5	27.5	41.7	23.7	52.6	36.2
GenTranslate-V2 (ours)	**42.7**	**40.6**	**39.4**	**43.6**	**54.0**	**23.3**	**44.8**	**37.0**	**27.7**	**10.2**	**48.0**	**30.5**	**42.3**	**25.4**	**55.9**	**37.7**
Cascaded ASR+MT Methods																
Whisper + NLLB-3.3b (2022)	34.4	35.5	31.7	37.9	45.4	19.0	39.8	26.7	17.5	0.1	37.0	20.6	29.4	25.5	45.9	29.8
Whisper + mBART-50 (2023)*	38.8	37.0	33.0	40.7	49.0	21.5	39.9	32.7	16.3	0.4	37.0	21.4	25.0	23.0	45.5	30.7
SeamlessM4T (ASR+MT) (2023a)	41.5	39.8	37.5	41.1	53.2	21.4	42.4	29.9	26.5	8.0	45.2	28.8	38.6	22.0	50.6	35.1
GenTranslate (ours)	**41.8**	**40.2**	**38.4**	**42.1**	**53.7**	**22.9**	**43.8**	**34.3**	**29.4**	**9.5**	**49.7**	**31.2**	**39.6**	**22.3**	**54.6**	**36.9**
SeamlessM4T-V2 (ASR+MT) (2023b)	43.0	40.6	38.8	43.0	55.2	22.9	43.2	33.9	27.2	8.6	47.0	27.8	41.9	24.7	53.1	36.7
GenTranslate-V2 (ours)	**43.1**	**41.1**	**39.5**	**43.3**	**55.6**	**24.5**	**44.9**	**37.4**	**27.8**	**10.3**	**48.7**	**30.4**	**42.0**	**26.0**	**58.4**	**38.2**

Table 2: Speech translation results on CoVoST-2 **X→En** test sets in terms of BLEU score. Remarks follow Table 1.

En→X language directions. For speech translation, we select FLEURS (Conneau et al., 2023), CoVoST-2 (Wang et al., 2020), and MuST-C (Di Gangi et al., 2019). For machine translation, we select FLORES (Costa-jussà et al., 2022), WMT'16 (Bojar et al., 2016), WMT'19 (Barrault et al., 2019), and WMT'20 (Loïc et al., 2020) corpora. As a result, we obtain over 592K hypotheses-translation pairs in 11 languages. The details of dataset statistics are presented in §A.3 and Table 15, 17.

Since the hypotheses-translation data pairs in HypoTranslate dataset are monolingual, we can also use ASR dataset to benefit GenTranslate training, especially for low-resource language pairs. Relevant studies are illustrated in §4.3.2 and Table 7. Our best result was obtained by first performing translation with SeamlessM4T and then integrating the N-best candidates using LLMs.

4 Experiments

4.1 Setup

4.1.1 Model Selection

LLMs. We select the popular LLaMA-2 (Touvron et al., 2023b) for our paradigm. Specifically, we employ LLaMA-2-7b[4] for English-target directions (X→En) and LLaMA-2-13b for non-English-target directions (En→X), as LLaMA-2 shows superior ability on English language while less-optimal on other languages. In addition, for En→X we also try some latest multilingual LLMs like BigTranslate[5] (Yang et al., 2023b) and ALMA[6] (Xu et al., 2023b) that are finetuned on LLaMA-13b.

Adapter. We follow the default settings of LLaMA-Adapter (Zhang et al., 2023b). The number of tunable Transformer layers L is set to $H-1$, which means all layers except the first one are tunable

[3] Our experiments are mainly conducted on SeamlessM4T-Large as they had already been done before Meta released the latest SeamlessM4T-Large-V2 on November 30th, 2023. For comprehensive evaluation, we rerun the main experiments on V2, which demonstrate similar effectiveness of our paradigm.

[4] https://huggingface.co/meta-llama/Llama-2-7b-hf

[5] https://huggingface.co/James-WYang/BigTranslate

[6] https://huggingface.co/haoranxu/ALMA-13B

En→X	FLEURS							CoVoST-2				MuST-C			
	Es	Fr	It	Ja	Pt	Zh	Avg.	Fa	Ja	Zh	Avg.	Es	It	Zh	Avg.
End-to-end ST Methods															
SeamlessM4T-Large (2023a)	23.8	41.6	23.9	21.0	40.8	28.6	30.0	18.3	24.0	34.1	25.5	**34.2**	**29.9**	16.2	26.8
GenTranslate (ours)	**25.4**	**43.1**	**25.5**	**28.3**	**42.4**	**34.3**	**33.2**	**21.1**	**29.1**	**42.8**	**31.0**	33.9	29.4	**18.5**	**27.3**
SeamlessM4T-Large-V2 (2023b)	23.8	42.6	24.5	21.7	43.0	29.5	30.9	16.9	23.5	34.6	25.0	32.1	**27.5**	15.6	25.1
GenTranslate-V2 (ours)	**25.5**	**44.0**	**26.3**	**28.9**	**44.5**	**34.9**	**34.0**	**19.4**	**29.0**	**43.6**	**30.7**	**32.2**	27.3	**18.1**	**25.9**
Cascaded ASR+MT Methods															
Whisper + NLLB-3.3b (2022)	25.1	41.3	25.0	19.0	41.5	23.5	29.2	13.6	19.0	32.0	21.5	35.3	29.9	13.5	26.2
SeamlessM4T-Large (ASR+MT) (2023a)	24.6	44.6	25.4	22.5	41.9	31.2	31.7	18.8	24.0	35.1	26.0	35.1	30.8	17.7	27.9
GenTranslate (ours)	**26.8**	**45.0**	**26.6**	**29.4**	**43.1**	**36.8**	**34.6**	**21.8**	**30.5**	**43.3**	**31.9**	**35.5**	**31.0**	**19.6**	**28.7**
SeamlessM4T-V2 (ASR+MT) (2023b)	24.7	44.1	25.1	20.6	43.6	30.6	31.5	17.4	23.8	35.4	25.5	33.0	27.8	14.5	25.1
GenTranslate-V2 (ours)	**27.0**	**44.3**	**26.4**	**27.8**	**44.5**	**36.1**	**34.4**	**20.8**	**29.7**	**43.5**	**31.3**	**33.2**	**28.3**	**16.9**	**26.1**

Table 3: Speech translation results on FLEURS, CoVoST-2, and MuST-C **En→X** test sets in terms of BLEU score. We use **bold** to highlight surpassing SeamlessM4T baseline, and use underline to highlight the state-of-the-art performance. The baseline methods are introduced in §B.3, and all of their results are reproduced by ourselves.

X→En	Ar	De	El	Es	Fa	Fr	It	Ja	Uk	Zh	Avg.
ALMA-13b (Xu et al., 2023b)	10.8	27.7	12.1	18.1	10.2	27.4	19.6	14.2	22.7	16.9	18.0
BigTranslate (Yang et al., 2023b)	18.6	35.9	9.5	29.0	1.4	38.7	29.0	16.9	25.9	23.0	22.8
NLLB-3.3b (Costa-jussà et al., 2022)	43.0	44.6	37.7	32.2	38.7	46.2	34.6	28.1	40.8	29.5	37.5
SeamlessM4T-Large (Barrault et al., 2023a)	43.7	45.1	37.7	31.5	39.0	45.1	35.2	26.1	41.2	29.9	37.5
GenTranslate (ours)	**43.9**	**45.3**	**38.5**	**35.5**	**39.4**	**46.4**	**36.6**	**26.7**	**41.8**	**30.5**	**38.5**
SeamlessM4T-Large-V2 (Barrault et al., 2023b)	41.5	44.1	35.6	29.9	37.6	45.5	33.5	25.5	39.0	29.0	36.1
GenTranslate-V2 (ours)	**42.0**	**44.5**	**36.6**	**34.4**	**38.1**	**46.7**	**35.1**	**26.7**	**39.3**	**29.9**	**37.3**

Table 4: Machine translation results on FLORES **X→En** test sets in terms of BLEU score. Remarks follow Table 3.

En→X	WMT'16	WMT'19		WMT'20		Avg.
	Ro	Cs	Lt	Ja	Zh	
ALMA-13b (2023b)	6.2	6.1	0.3	3.5	11.3	5.5
BigTranslate (2023b)	21.4	19.0	8.7	7.3	29.0	17.1
NLLB-3.3b (2022)	31.0	25.3	16.0	15.2	26.9	22.9
SeamlessM4T-Large	32.7	26.0	17.2	17.0	27.2	24.0
GenTranslate (ours)	**33.5**	**27.2**	**19.4**	**21.4**	**30.7**	**26.4**
SeamlessM4T-Large-V2	32.2	25.2	16.2	15.2	28.7	23.5
GenTranslate-V2 (ours)	**33.2**	**26.6**	**18.2**	**19.3**	**31.6**	**25.8**

Table 5: Machine translation results on WMT'16,19,20 **En→X** test sets in BLEU. Remarks follow Table 3.

with inserted prompts. The prompt length U is set to 10. More details are provided in §B.1.

4.1.2 Training Details

The batch size is set to 4, with accumulation iterations set to 8 (*i.e.*, real batch size is 32). We train 2 epochs with AdamW optimizer (Loshchilov and Hutter, 2018), with learning rate initialized to $1e^{-2}$ and then linearly decrease to $1e^{-5}$ during training.

4.2 Comparison with the State-of-the-art

4.2.1 Speech Translation

X→English (En). Table 1 and 2 present the X→En speech translation performance on FLEURS and CoVoST-2 datasets. We can observe from Table 1 that all the strong baselines like Whisper, AudioPaLM2 and SeamlessM4T-Large perform well on 15 X→En directions, where SeamlessM4T-Large is the best (27.1 BLEU). With LLMs introduced for N-best integration, our GenTranslate achieves consistent improvements on various source languages X, where further analysis on language family is presented in §4.4.1. As a result, our GenTranslate shows 3.0 BLEU improvement over SeamlessM4T-Large, which verifies the effectiveness of LLMs for generative translation[7].

Following the speech translation literature, we also investigate cascaded ASR+MT methods for evaluation. We can observe from Table 1 that, with the same SeamlessM4T-Large backbone, cascaded system outperforms end-to-end system by 4.8 BLEU score, which is consistent with previous findings (Xu et al., 2023a). Latest SeamlessM4T-Large-V2 further improves V1 model, and our GenTranslate shows significant and consistent gains of performance over theses two backbones.

Table 2 presents the X→En ST results on more language directions of CoVoST-2 dataset, where we introduce more latest baselines for comprehensive comparison. In end-to-end methods, SeamlessM4T-Large achieves a good 34.5 BLEU score though underperforms the state-of-the-art AudioPaLM2[8]. In comparison, our GenTranslate achieves a promis-

[7] Latest SeamlessM4T-Large-V2 achieves significant gains over V1, based on which the proposed GenTranslate also shows similar effectiveness in our study.

[8] We speculate it could be attributed to the train-test domain mismatch because SeamlessM4T-Large outperforms AudioPaLM2 by a large margin on FLEURS dataset in Table 1.

En→X	FLEURS					CoVoST-2				WMT					
	Es	Fr	It	Pt	Avg.	Fa	Ja	Zh	Avg.	Ro	Cs	It	Ja	Zh	Avg.
SeamlessM4T-Large (2023a)	24.6	44.6	25.4	41.9	34.1	18.8	24.0	35.1	26.0	32.7	26.0	17.2	17.0	27.2	24.0
GenTranslate *with*															
BigTranslate (2023b)	25.3	44.2	25.5	40.8	34.0	5.2	23.5	42.6	23.8	31.3	24.9	15.8	13.9	27.9	22.8
ALMA-13b (2023b)	24.9	43.5	25.1	40.6	33.5	19.2	29.3	**43.9**	30.8	31.1	25.5	17.7	17.3	26.8	23.7
LLaMA-2-13b (2023b)	**26.8**	**45.0**	**26.6**	**43.1**	**35.4**	**21.8**	**30.5**	43.3	**31.9**	**33.5**	**27.2**	**19.4**	**21.4**	**30.7**	**26.4**

Table 6: Effect of different multilingual LLMs on GenTranslate, in terms of the speech translation results on FLEURS and CoVoST-2 En→X test sets, as well as the machine translation results on WMT En→X test sets.

De→En	BLEU Score
End-to-end ST Methods	
SeamlessM4T (ST) (Barrault et al., 2023a)	35.8
SeamlessM4T (ST) + GenTranslate	**39.2**
Cascaded ASR+MT Methods	
SeamlessM4T (ASR+MT) (Barrault et al., 2023a)	39.7
SeamlessM4T (ASR+MT) + GenTranslate	**41.6**
ASR+GenTranslate Method	
SeamlessM4T (ASR) + GenTranslate *with*	
LLaMA-2-7b (Touvron et al., 2023b)	36.8
BigTranslate (Yang et al., 2023b)	38.2
ALMA-7b (Xu et al., 2023b)	**40.6**

Table 7: Performance of ASR+GenTranslate system on FLEURS De→En ST test set. As shown in Fig. 4, it first uses ASR to produce German N-best hypotheses, and then leverages LLMs to generate the English translation from them. Different LLMs are investigated here.

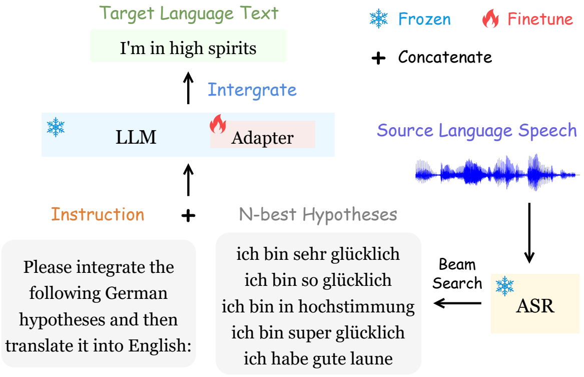


Figure 4: Illustration of the "ASR+GenTranslate" system for ST task as introduced in Table 7 and §4.3.2. This system engages LLMs into the translation process by combining it with the N-best integration process.

ing improvement over SeamlessM4T. Similar phenomenon can be observed in cascaded systems, where SeamlessM4T significantly outperforms the competitive baselines that combine state-of-the-art ASR and MT models, and our GenTranslate moves one step forward with 1.8 BLEU improvement. Similar improvements can be observed on SeamlessM4T-Large-V2 backbone.

English (En)→X. For comprehensive evaluation, we also present En→X ST results on three datasets in Table 3. SeamlessM4T (both Large and Large-V2) achieves excellent performance on En→X ST tasks under both end-to-end and cascaded systems. In comparison, our proposed GenTranslate achieves significant performance improvements (∼3 BLEU score) in various language directions. Since En→X translation tasks produce non-English N-best hypotheses for LLM integration, such performance gains indicates the excellent multilingual abilities of LLMs (*i.e.*, LLaMA-2).

4.2.2 Machine Translation

X→English (En). Table 4 presents the X→En MT results on FLORES dataset. The baseline methods ALMA-13b and BigTranslate show limited performance. NLLB-3.3b achieves an improved performance of 37.5 BLEU, which is comparable to SeamlessM4T-Large. Based on that, our GenTranslate achieves the state-of-the-art with consistent gains on all language directions except Ja→En.

English (En)→X. Table 5 presents the En→X MT results on WMT test sets. Similar to previous results, we observe much higher BLEU scores of NLLB-3.3b than ALMA-13b and BigTranslate. SeamlessM4T-Large surpasses NLLB-3.3b by large-scale multitask training. The proposed GenTranslate achieves the state-of-the-arts on all language directions with a gain of 2.4 BLEU score. Please note that SeamlessM4T-Large-V2 underperforms V1 on selected MT datasets, but our GenTranslate achieves consistent gains on both of them.

In summary, we observe consistent improvements of GenTranslate over various baselines (*i.e.*, SeamlessM4T, Whisper, etc.), various tasks (*i.e.*, ST and MT), various test data (*i.e.*, FLEURS, WMT, etc.), and various language directions (*i.e.*, X→En and En→X). Therefore, the effectiveness and generality of our approach are well verified.

4.3 Ablation Study

4.3.1 Effect of Different LLMs

According to Table 3 and 5, LLaMA-2 has shown excellent multilingual ability. To further investigate the role of this ability in GenTranslate, we select

X→En	Ar	Cy	De	El	Es	Fa	Fr	Hi	It	Ja	Pt	Ta	Uk	Vi	Zh	Avg.
SeamlessM4T (ASR+MT)	38.9	37.0	39.7	29.0	27.7	34.1	37.7	33.9	28.9	21.7	42.3	23.7	34.0	24.9	24.4	31.9
GenTranslate *with*																
LLaMA-Adapter	39.9	**39.4**	41.6	**32.8**	31.2	35.9	**40.6**	34.9	32.1	**22.8**	45.0	**24.1**	**36.9**	**27.4**	25.7	34.0
LLaMA-LoRA	**40.2**	39.3	**41.8**	**32.8**	**31.6**	**36.0**	**40.6**	**35.2**	**32.4**	22.5	**45.1**	**24.1**	36.7	27.1	**26.0**	**34.1**

Table 8: Comparison between LLaMA-Adapter and LLaMA-LoRA for efficient LLM finetuning in our GenTranslate, in terms of the speech translation results on FLEURS X→En test sets.

X→En	Indo-European											non-Indo-European					
	Fa	Hi	It	Es	Fr	Pt	Cy	De	El	Uk	Avg.	Ar	Vi	Ja	Ta	Zh	Avg.
SeamlessM4T (ASR+MT)	34.1	33.9	28.9	27.7	37.7	42.3	37.0	39.7	29.0	34.0	34.4	38.9	24.9	21.7	23.7	24.4	26.7
GenTranslate (ours)	35.9	34.9	32.1	31.2	40.6	45.0	39.4	41.6	32.8	36.9	37.0	39.9	27.4	22.8	24.1	25.7	28.0
Δ BLEU	1.8	1.0	3.2	3.5	2.9	2.7	2.4	1.9	3.8	2.9	2.6	1.0	2.5	1.1	0.4	1.4	1.3

Table 9: Effect of language family on our proposed GenTranslate. We report speech translation results on FLEURS X→En test sets in this study. For simplicity, we split all the languages into two families, *i.e.*, Indo-European (same as English) and non-Indo-European, and more detailed information are presented in Table 14.

X→En		Ar	De	Es	Fr	Pt	Zh	Avg.
SeamlessM4T-Large		32.8	35.8	25.0	33.1	38.9	19.8	30.9
GenTranslate *with*								
N =	1	31.3	35.4	26.9	35.2	41.5	19.3	31.6
	3	34.2	38.9	29.5	36.4	42.8	21.3	33.9
	5	34.6	39.2	**29.8**	**37.0**	43.0	**21.7**	34.2
	8	34.8	**39.9**	29.4	36.9	43.0	21.5	**34.3**
	10	**35.3**	39.8	29.4	36.6	**43.2**	21.6	**34.3**
	15	34.9	39.5	29.6	36.4	42.8	21.6	34.1

Table 10: Effect of N-best list size on GenTranslate (default N=5), in terms of ST results on FLEURS X→En.

two latest multilingual LLMs for comparison, *i.e.*, BigTranslate and ALMA-13b. Table 6 shows that both of them perform worse than LLaMA-2-13b for ST and MT tasks. One explanation is, BigTranslate and ALMA-13b are finetuned on MT task that requires cross-lingual ability, while the En→X GenTranslate mainly requires strong monolingual ability of language X, such mismatch may explain why MT finetuning fails to enhance GenTranslate.

4.3.2 Role of LLMs in GenTranslate

To further investigate the role of LLMs in our GenTranslate, we build an ASR+GenTranslate system for ST task as shown in Fig. 4. Take De→En as an example, we first send the German speech input into ASR to produce N-best transcriptions, which are then fed by LLMs to generate English translation. In other words, LLMs are assigned N-best integration and translation tasks at the same time. As shown in Table 7, among the three evaluated LLMs, ALMA-7b achieves the best performance thanks to its MT finetuning during development, but it still underperforms the best cascaded method (40.6 vs. 41.6). We can conclude from such observations that 1) LLaMA-2 provides reasonable translation ability and it can be further improved via MT task finetuning (*i.e.*, ALMA). 2) In this study, LLM underperforms SeamlessM4T in translation task, but it shows remarkable ability in N-best integration. Therefore, future work may focus on how to better engage LLMs into the translation part.

4.3.3 Effect of N-best List Size

GenTranslate relies on powerful LLMs and informative N-best hypotheses to generate higher-quality translation output. Therefore, the amount of information in N-best hypotheses could be a key factor of GenTranslate's performance. We can observe from Table 10 that with the increase of N, the performance of GenTranslate first improves and then drops, where the best choice ranges from 5 to 10. We believe that small N results in insufficient information for generation of ground-truth translation, while too large N leads to information redundancy and thus increases the miscorrection and hallucination. In this work, we set N to 5 for the best trade-off between efficiency and quality.

4.3.4 LLaMA-Adapter vs. LLaMA-LoRA

Apart from LLaMA-Adapter, low-rank adaptation (LoRA) (Hu et al., 2021; Yu et al., 2023) is another popular efficient LLM finetuning strategy. Table 8 compares the performance between LLaMA-Adapter and LLaMA-LoRA for proposed GenTranslate, in terms of the BLEU results of ST task on FLEURS X→En test sets. We can observe similar BLEU performance of these two strategies on GenTranslate (34.0 vs. 34.1), indicating that the efficient LLM finetuning strategy is not a key factor in GenTranslate paradigm.

Method	Utterance	BLEU Score
N-best Candidates	TV reports show that white smoke is escaping from the plant.	28.6
	TV reports show that white smoke is escaping from the facility.	12.2
	Television reports show that white smoke is escaping from the plant.	34.2
	Television reports show that white smoke is escaping from the facility.	19.2
	TV reports show that white smoke escapes from the plant.	31.7
GenTranslate (ours)	Television reports show white smoke coming out of the plant.	**58.8**
Ground-truth Translation	Television reports show white smoke coming from the plant.	-

Table 11: Case study of GenTranslate. The test sample is selected from the FLEURS De→En ST test set.

4.4 Analysis

4.4.1 Effect of Language Family

Table 9 analyzes the effect of language family using the X→En ST results. The source language X is grouped into two categories depending on whether it belongs to Indo-European family (English is also Indo-European language). First, we observe better results of SeamlessM4T when X belongs to Indo-European family, indicating that translation within same family is easier than across different families. Then, we also observe larger BLEU improvement of GenTranslate over baseline when X is Indo-European language (2.6 vs. 1.3). The reason could be, within-family translation produces N-best hypotheses with higher quality and richer information, which is beneficial to GenTranslate.

4.4.2 Case Study

Table 11 shows a case study where GenTranslate outperforms the 1-best hypothesis by a large margin. We may speculate two key points about its working mechanism, where it first extract the word "Television" from $3^{rd}/4^{th}$ hypotheses to replace "TV" and then reason out the word "coming" that does not exist in N-best list. Therefore, our paradigm may not only integrate the N-best sentences for better result, but also improve the translation quality by itself. Another non-English case study is in Appendix C.1.

4.4.3 Visualizations of GenTranslate Output

Fig. 5 visualizes the n-gram tokens in GenTranslate output, which contains sufficient semantic information to match the ground-truth translation. In comparison, the 1-best hypothesis lacks such information to produce high-quality translation output, which highlights the contribution of N-best hypotheses in GenTranslate paradigm (see Fig. 2).

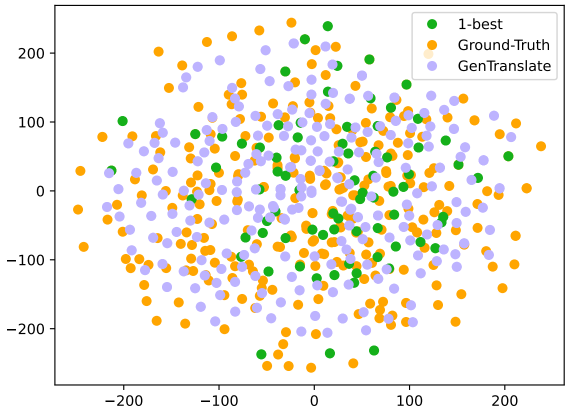


Figure 5: t-SNE visualization of n-grams in 1-best hypothesis (green), ground-truth translation (orange) and GenTranslate output (purple). It's an extension of Fig. 2.

5 Conclusion

In this paper, we propose a generative paradigm for translation tasks, namely GenTranslate, which leverages LLMs to integrate the diverse candidates in the decoded N-best list and generate a higher-quality translation result. Furthermore, we release a HypoTranslate dataset to support LLM finetuning, which contains over 592K hypotheses-translation pairs in 11 languages. Experimental evidence on various speech and machine translation benchmarks shows that our GenTranslate significantly outperforms the state-of-the-art model.

Limitations

There are two limitations existed in this work. First, the contribution of LLMs in our GenTranslate paradigm focuses on N-best hypotheses integration, while the translation part is actually done by SeamlessM4T model. Experiment results in Table 7 also indicate that LLMs are good at N-best hypotheses integration and SeamlessM4T is good at translation. Therefore, our future work could focus on how to better engage LLMs into the translation part to further improve the translation quality. Another limitation is about the latest second version of SeamlessM4T released by Meta, which indicates a stronger baseline for GenTranslate. In fact, our experiments had already been done on SeamlessM4T-Large before Meta released the latest SeamlessM4T-Large-V2 on November 30th, 2023. For comprehensive evaluation, we also rerun our main experiments on this latest V2 backbone, and our GenTranslate has shown similar effectiveness on it (highlighted in gray in Table 1 to 5). For brevity, we prefer to leave the ablation study and analyses on SeamlessM4T-Large backbone only, as our GenTranslate paradigm has shown similar effectiveness and patterns on V1 and V2 backbones.

Ethics Statement

This work does not pose any ethical issues. All the data used in this paper are publicly available and are used under following licenses: Creative Commons BY 4.0 License, Creative Commons CC0 License, Creative Commons BY-NC-ND 4.0 License, and Creative Commons BY-SA 4.0 License.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-GC-2022-005). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, et al. 2020. Paracrawl: Web-scale acquisition of parallel corpora. Association for Computational Linguistics (ACL).

Loıc Barrault, Ondrej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation. *Proceedings of WMT*.

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

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, et al. 2023b. Seamless: Multilingual expressive and streaming speech translation. *arXiv 2023*.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Ensiong Chng. 2023. Hyporadise: An open baseline for generative speech recognition with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Alexis Conneau, Min Ma, et al. 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, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *Proc. NAACL*, pages 2012–2017. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, et al. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ke Hu, Tara N Sainath, et al. 2023a. Improving multilingual and code-switching asr using large language model generated text. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Zhiqiang Hu, Yihuai Lan, et al. 2023b. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Chenyang Le, Yao Qian, Long Zhou, Shujie Liu, Michael Zeng, and Xuedong Huang. 2023. Comsl: A composite speech-language model for end-to-end speech-to-text translation. *arXiv preprint arXiv:2305.14838*.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiangyang Li, Edward Lin, and Tie-Yan Liu. 2021. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. *Advances in Neural Information Processing Systems*, 34:21708–21719.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Dan Liu, Mengge Du, Xiaoxi Li, Yuchen Hu, and Lirong Dai. 2021. The ustc-nelslip systems for simultaneous speech translation task at iwslt 2021. *arXiv preprint arXiv:2107.00279*.

Barrault Loïc, Biesialska Magdalena, Bojar Ondřej, Federmann Christian, Graham Yvette, Grundkiewicz Roman, Haddow Barry, Huck Matthias, et al. 2020. Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55. Association for Computational Linguistics,.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Rao Ma, Mark JF Gales, Kate Knill, and Mengjie Qian. 2023a. N-best t5: Robust asr error correction using multiple input hypotheses and constrained decoding space. *arXiv preprint arXiv:2303.00456*.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. 2023b. Can generative large language models perform asr error correction? *arXiv preprint arXiv:2307.04172*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

OpenAI. 2022. Introducing chatgpt. *OpenAI Blog*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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.

Srijith Radhakrishnan, Chao-Han Yang, Sumeer Khan, et al. 2023. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 10007–10016.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Emiru Tsunoo, Yosuke Kashiwagi, and Shinji Watanabe. 2021. Streaming transformer asr with blockwise synchronous beam search. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 22–29. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Siyin Wang, Chao-Han Huck Yang, Ji Wu, and Chao Zhang. 2023. Can whisper perform speech-based in-context learning. *arXiv preprint arXiv:2309.07081*.

Thomas Wang, Adam Roberts, et al. 2022. What language model architecture and pretraining objective works best for zero-shot generalization? In *International Conference on Machine Learning*, pages 22964–22984. PMLR.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yanling Xiao, Lemao Liu, et al. 2022. Bitiimt: A bilingual text-infilling method for interactive machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1958–1969.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023a. Recent advances in direct speech-to-text translation. *arXiv preprint arXiv:2306.11646*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023b. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023a. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Chao-Han Huck Yang, Linda Liu, et al. 2021a. Multi-task language modeling for improving speech recognition of rare words. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1087–1093. IEEE.

Chao-Han Huck Yang, Yun-Yun Tsai, and Pin-Yu Chen. 2021b. Voice2series: Reprogramming acoustic models for time series classification. In *International Conference on Machine Learning*, pages 11808–11819. PMLR.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023b. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, et al. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Renrui Zhang, Jiaming Han, et al. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

A HypoTranslate Dataset Details

In this section, we introduce the details of our proposed HypoTranslate dataset. We first introduce the speech and machine translation corpora that we utilize to build HypoTranslate in §A.1 and §A.2. Then, we present the dataset statistics in §A.3.

A.1 Speech Translation Corpus Selection

For speech translation task, we select three popular and public datasets that cover multiple languages:

FLEURS[9] (Conneau et al., 2023): Few-shot Learning Evaluation of Universal Representations of Speech (FLEURS) benchmark provides an n-way parallel speech dataset in 102 languages built on top of the machine translation FLORES-101 benchmark (Goyal et al., 2022), with approximately 12 hours of speech supervision per language. In this work, we select 15 X→En and 6 En→X language directions of speech translation data for evaluation.

CoVoST-2[10] (Wang et al., 2020): CoVoST-2 is a popular multilingual speech translation corpus based on Common Voice (Ardila et al., 2019) that consists of 2,880 hours speech data recorded from 78K speakers. In this work, we select 15 X→En and 3 En→X language directions for evaluation. Specifically, for En→X language directions, we randomly select 1,000 testing samples from the original test split for higher evaluation efficiency.

MuST-C[11] (Di Gangi et al., 2019): MuST-C is a multilingual speech translation corpus whose size and quality facilitate the training of end-to-end systems for spoken language translation from English into 15 languages. In this work, we select 3 En→X language directions for evaluation.

A.2 Machine Translation Corpus Selection

For machine translation task, we select two popular and public datasets that cover multiple languages:

FLORES[12] (Costa-jussà et al., 2022): FLORES consists of 3001 sentences sampled from English-language Wikimedia projects for 204 total languages. Approximately one third of sentences are collected from each of these sources: Wikinews, Wikijunior, and Wikivoyage. The content is professionally translated into 200+ languages to create FLORES dataset. In this work, we select 10 X→En language directions for evaluation.

WMT: The Conference on Machine Translation (WMT) is a popular evaluation benchmark for MT task. In this work, we select the newstest data of Ro→En language direction from WMT'16[13] (Bojar et al., 2016), Cs→En and It→En directions from WMT'19[14] (Barrault et al., 2019), Ja→En and Zh→En directions from WMT'20[15] (Loïc et al., 2020) for evaluation, and corresponding newdev data is used for validation. The training data is obtained from ParaCrawl-V9[16] (Bañón et al., 2020) and JParaCrawl[17] (Morishita et al., 2020) datasets.

A.3 Statistics

After performing beam search decoding on the selected speech and machine translation corpora introduced above, we collect over 592K pairs of N-best hypotheses and ground-truth translation to build the HypoTranslate dataset. The statistics are illustrated in Table 15 and 17, which present the number of hypotheses-translation pairs and the average utterance length. We plan to release the HypoTranslate dataset to public upon publication and open the development venue for more data.

B Experimental Setup Details

B.1 Model Setups

We select two latest foundation LLMs for evaluation, including LLaMA-2-7b (Touvron et al., 2023b) and LLaMA-2-13b (Touvron et al., 2023b). In addition, in order to evaluate the multilingual ability of LLMs for GenTranslate with non-English-target directions, we also select two latest finetuned LLMs on MT task, including BigTranslate (Yang et al., 2023b) and ALMA-13b (Xu et al., 2023b). Table 12 compares their main configurations. For efficient LLM finetuning, we follow the default settings of LLaMA-Adapter[18] (Zhang et al., 2023b).

[9] https://huggingface.co/datasets/google/fleurs

[10] https://github.com/facebookresearch/covost

[11] https://mt.fbk.eu/must-c-releases/

[12] https://huggingface.co/datasets/facebook/flores

[13] https://www.statmt.org/wmt16/translation-task.html

[14] https://www.statmt.org/wmt19/translation-task.html

[15] https://www.statmt.org/wmt20/translation-task.html

[16] https://paracrawl.eu/

[17] https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

[18] https://github.com/Lightning-AI/lit-gpt/blob/main/lit_gpt/adapter.py

LLM	LLaMA-2-7b	LLaMA-2-13b	BigTranslate	ALMA-13b
Number of Transformer Layers H	32	40	40	40
Number of Attention Heads N_{head}	32	40	40	40
Embedding Size D	4,096	5,120	5,120	5,120
Block Size B	4,096	4,096	4,096	4,096
Vocabulary Size V	32,000	32,000	53,613	32,000

Table 12: Comparison between main configurations of different popular LLMs.

B.2 Inference Setups

In the response generation during inference stage, we set a temperature of 0.2 and top-1 sampling, *i.e.*, greedy search. We have observed over-confidence phenomenon in our experiments (*i.e.*, output probability distribution for decision is close to one-hot), which results in similar performance with different k for top-k sampling. Therefore, we select top-1 sampling for higher decoding efficiency.

B.3 Translation Baselines

To comprehensively evaluate our GenTranslate model, we selected some of the latest and most advanced baselines in speech and machine translation for comparison. We will introduce these in the following subsections.

B.3.1 Speech Translation

XLS-R[19] (Babu et al., 2021): XLS-R is a large-scale model for cross-lingual speech representation learning based on Wav2vec 2.0 (Baevski et al., 2020). They train models with up to 2B parameters on 500K hours of publicly available speech audio in 128 languages, which achieves superior performance on a wide range of multilingual speech processing tasks, including speech translation, speech recognition and language identification.

Whisper[20] (Radford et al., 2023): Whisper is a large-scale automatic speech recognition (ASR) system trained on 680K hours of multilingual and multitask supervised data collected from the web, which shows excellent robustness to accents, background noise and technical language. Moreover, it enables transcription in multiple languages, as well as translation from those languages into English.

AudioPaLM2 (Rubenstein et al., 2023): AudioPaLM2 fuses text-based and speech-based language models, PaLM-2 (Anil et al., 2023) and AudioLM (Borsos et al., 2023), into a unified multimodal architecture that can process and generate text and speech with applications including speech recognition and speech-to-speech translation. AudioPaLM2 inherits the capability to preserve paralinguistic information such as speaker identity and intonation from AudioLM and the linguistic knowledge present only in text large language models such as PaLM-2. The resulting model significantly outperforms existing systems for speech translation and has the ability to perform zero-shot speech-to-text translation for many unseen languages.

ComSL[21] (Le et al., 2023): ComSL is a speech-language model built atop a composite architecture of public pre-trained speech-only and language-only models and optimized data-efficiently for spoken language tasks. Particularly, they propose to incorporate cross-modality learning into transfer learning and conduct them simultaneously for downstream tasks in a multi-task learning manner, which has demonstrated effectiveness in end-to-end speech-to-text translation tasks.

B.3.2 Machine Translation

NLLB[22] (Costa-jussà et al., 2022): No Language Left Behind (NLLB) is a first-of-its-kind, AI breakthrough project that open-sources models capable of delivering evaluated, high-quality translations directly between 200 languages.

BigTranslate[5] (Yang et al., 2023b): BigTranslate adapts LLaMA-13b (Touvron et al., 2023a) that covers only 20 languages and enhances it with multilingual translation capability on up to 102 languages by instruction-following finetuning, which achieves comparable MT performance to ChatGPT (OpenAI, 2022) and Google Translate.

ALMA[6] (Xu et al., 2023b): ALMA proposes a novel finetuning approach for LLMs that is specifically designed for MT task, eliminating the need for the abundant parallel data that traditional translation models usually depend on, which includes two stages: initial finetuning on monolingual data

[19] https://huggingface.co/models?other=xls_r

[20] https://github.com/openai/whisper

[21] https://github.com/nethermanpro/ComSL

[22] https://huggingface.co/facebook/nllb-200-3.3B

Method	Utterance	BLEU Score
N-best Candidates	地球河流流入海洋的20%的水来自亚马逊.	15.0
	地球河流流入海洋的20%的水源来自亚马逊.	15.0
	地球河流流入海洋的全部20%的水来自亚马逊.	12.3
	地球河流流入海洋的20%的水来自亚马逊	15.0
	地球河流流入海洋的全部20%的水来自亚马逊	12.3
GenTranslate (ours)	地球上的河流汇入大洋的 20% 的水来自亚马逊河。	**18.7**
Ground-truth Translation	亚马逊河占全世界所有河流的入海流量的 20%。	-

Table 13: Supplementary case study. The test sample is selected from the FLEURS En→Zh ST test set.

Language	Family	Sub-grouping
Persian (Fa)	Indo-European	Indo-Iranian
Hindi (Hi)	Indo-European	Indo-Iranian
Italian (It)	Indo-European	Indo-Iranian
Spanish (Es)	Indo-European	Italic
French (Fr)	Indo-European	Italic
Portuguese (Pt)	Indo-European	Italic
Welsh (Cy)	Indo-European	Celtic
English (En)	Indo-European	Germantic
German (De)	Indo-European	Germantic
Greek (El)	Indo-European	Greek
Ukranian (Uk)	Indo-European	Balto-Slavic
Arabic (Ar)	Afro-Asiatic	Semitic
Vietnamese (Vi)	Austro-Asiatic	Mon-Khmer
Japanese (Ja)	Japonic	-
Tamil (Ta)	Dravidian	Dravidian
Chinese (Zh)	Sino-Tibetan	Chinese

Table 14: Detailed language family and sub-grouping information (Babu et al., 2021) of FLEURS datasets.

followed by subsequent finetuning on a small set of high-quality parallel data. Built based on LLaMA-2, it has achieved significant improvement over prior works across multiple translation directions.

C Supplementary Experiment Results

C.1 Supplementary Case Study

Table 13 supplies a case study from FLEURS En→Zh ST test set. We can observe that the N-best candidate are semantically similar to each other and only varies in sentence structure. In our GenTranslate paradigm, LLMs integrates the different patterns of N-best hypotheses to generate a new translation result with 3.7 BLEU improvement over 1-best hypothesis. Such observation verifies the effectiveness of LLMs in our GenTranslate paradigm to generate better translation output.

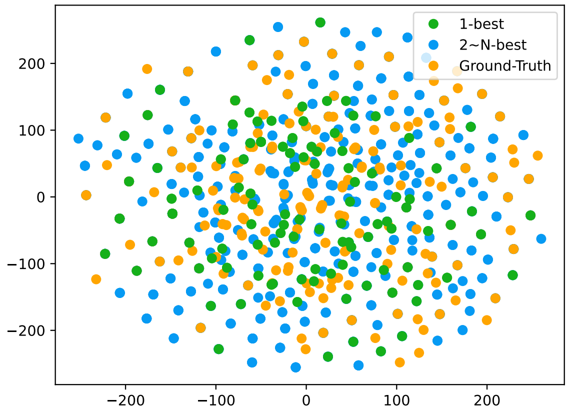


Figure 6: t-SNE visualization of n-gram tokens in ASR 1-best hypothesis (green), 2 to N-best hypotheses (blue), and the ground-truth transcription (orange). Different from the ST hypotheses in Fig. 2, ASR 1-best hypothesis aligns well with the ground-truth transcription, where the role of 2∼N-best hypotheses is to provide diverse candidate tokens for correcting errors.

C.2 BLEU vs. chrF++

We report translation performance in terms of the BLEU score (Papineni et al., 2002) in most experiments of this work. For more comprehensive evaluation, Table 16 presents both BLEU and chrF++ scores (Popović, 2017; Barrault et al., 2023a) on FLEURS X→En test sets, where we can observe consistent improvements of BLEU and chrF++ scores (2.1 Δ BLEU and 0.9 Δ chrF++) in GenTranslate. It indicates that both metrics are applicable for the evaluation of translation tasks.

Data Source	Source / Target	Train		Dev.		Test	
	Language X	# Pairs	Length	# Pairs	Length	# Pairs	Length
FLEURS (Conneau et al., 2023) (X→En)	Arabic (Ar)	2,062	20.4	295	19.8	428	21.4
	Welsh (Cy)	3,349	21.1	447	20.6	1,021	22.1
	German (De)	2,926	20.7	363	20.1	862	21.9
	Greek (El)	3,148	20.9	271	20.5	650	21.7
	Spanish (Es)	2,732	20.8	408	20.5	908	21.8
	Persian (Fa)	3,032	20.9	369	20.1	871	21.8
	French (Fr)	3,119	20.8	289	19.9	676	21.8
	Hindi (Hi)	2,072	20.6	239	19.2	418	21.4
	Italian (It)	2,970	20.6	391	20.2	865	21.7
	Japanese (Ja)	2,241	20.2	266	19.6	650	21.3
	Portuguese (Pt)	2,731	20.7	386	20.2	919	21.9
	Tamil (Ta)	2,317	20.7	377	20.0	591	22.0
	Ukrainian (Uk)	2,741	20.8	325	20.3	750	22.0
	Vietnamese (Vi)	2,927	20.7	361	20.2	857	21.8
	Chinese (Zh)	3,178	21.0	409	20.6	945	22.1
CoVoST-2 (Wang et al., 2020) (X→En)	French (Fr)	30,000	8.9	1,000	8.9	14,760	9.4
	German (De)	30,000	9.8	1,000	10.2	13,511	9.8
	Catalan (Ca)	30,000	10.3	1,000	10.3	12,730	10.5
	Spanish (Es)	30,000	9.7	1,000	9.6	13,221	9.9
	Russian (Ru)	12,112	11.9	1,000	11.9	6,300	11.8
	Chinese (Zh)	7,085	12.0	1,000	11.9	4,898	11.6
	Dutch (Nl)	7,108	8.2	1,000	8.5	1,699	8.5
	Turkish (Tr)	3,966	8.3	1,000	8.1	1,629	8.3
	Estonian (Et)	1,782	17.8	1,000	15.5	1,571	16.1
	Mongolian (Mn)	2,067	11.2	1,000	11.2	1,759	11.3
	Arabic (Ar)	2,283	5.8	1,000	5.7	1,695	5.7
	Latvian (Lv)	2,337	6.1	1,000	6.3	1,629	6.2
	Slovenian (Sl)	1,843	7.2	509	7.0	360	6.3
	Japanese (Ja)	1,119	8.3	635	8.5	684	8.4
	Indonesian (Id)	1,243	6.6	792	6.6	844	6.7
FLEURS (Conneau et al., 2023) (En→X)	Spanish (Es)	2,502	25.0	394	25.1	643	26.1
	French (Fr)	2,592	24.4	363	24.1	612	25.5
	Italian (It)	2,564	23.2	386	22.8	640	24.4
	Japanese (Ja)	2,290	53.6	351	53.1	592	55.6
	Portuguese (Pt)	2,503	22.4	387	21.9	645	23.4
	Chinese (Zh)	2,592	42.3	394	40.7	646	42.7
CoVoST-2 (Wang et al., 2020) (En→X)	Persian (Fa)	30,000	10.8	1,000	9.3	1,000	9.5
	Japanese (Ja)	30,000	28.5	1,000	26.6	1,000	23.3
	Chinese (Zh)	30,000	19.7	1,000	19.7	1,000	16.0
MuST-C (Di Gangi et al., 2019) (En→X)	Spanish (Es)	6,000	19.4	1,316	20.1	2,502	17.1
	Italian (It)	6,000	18.2	1,309	18.8	2,574	16.4
	Chinese (Zh)	6,000	49.6	888	63.7	1,823	46.3
Total		327,533	15.9	27,920	16.9	102,378	13.3

Table 15: HypoTranslate dataset **(ST part)** statistics in terms of the number of hypotheses-translation pairs and average length of ground-truth utterance in different language directions.

X→En	Ar	Cy	De	El	Es	Fa	Fr	Hi	It	Ja	Pt	Ta	Uk	Vi	Zh	Avg.
BLEU score																
SeamlessM4T (ASR+MT)	38.9	37.0	39.7	29.0	27.7	34.1	37.7	33.9	28.9	21.7	42.3	23.7	34.0	24.9	24.4	31.9
GenTranslate (ours)	**39.9**	**39.4**	**41.6**	**32.8**	**31.2**	**35.9**	**40.6**	**34.9**	**32.1**	**22.8**	**45.0**	**24.1**	**36.9**	**27.4**	**25.7**	**34.0**
chrF++ score																
SeamlessM4T (ASR+MT)	62.7	60.0	63.8	55.0	56.0	58.7	62.4	58.8	57.0	**47.9**	65.9	**49.8**	59.2	50.5	51.5	57.3
GenTranslate (ours)	**63.1**	**61.2**	**64.9**	**57.0**	**57.1**	**59.7**	**64.0**	**59.1**	**58.0**	47.6	**67.2**	49.7	**60.8**	**51.6**	**52.0**	**58.2**

Table 16: Speech translation results on FLEURS X→En test sets in terms of chrF++ score.

Data Source	Source / Target	Train		Dev.		Test	
	Language X	# Pairs	Length	# Pairs	Length	# Pairs	Length
FLORES (Costa-jussà et al., 2022) (X→En)	Arabic (Ar)	2,062	20.4	295	19.8	1,012	21.6
	German (De)	2,926	20.7	363	20.1	1,012	21.6
	Greek (El)	3,148	20.9	271	20.5	1,012	21.6
	Spanish (Es)	2,732	20.8	408	20.5	1,012	21.6
	Persian (Fa)	3,032	20.9	369	20.1	1,012	21.6
	French (Fr)	3,119	20.8	289	19.9	1,012	21.6
	Italian (It)	2,970	20.6	391	20.2	1,012	21.6
	Japanese (Ja)	2,241	20.2	266	19.6	1,012	21.6
	Ukrainian (Uk)	2,741	20.8	325	20.3	1,012	21.6
	Chinese (Zh)	3,178	21.0	409	20.6	1,012	21.6
WMT'{16,19,20} (En→X)	Czech (Cs)	15,000	12.3	2,983	15.8	1,997	18.8
	Japanese (Ja)	15,000	49.8	1,998	53.1	1,000	59.8
	Lithuanian (Lt)	15,000	12.0	2,000	16.5	998	16.6
	Romanian (Ro)	15,000	16.7	1,999	22.6	1,999	21.7
	Chinese (Zh)	15,000	35.6	1,997	47.8	1,418	60.7
Total		103,149	24.0	14,363	27.5	17,532	26.3

Table 17: HypoTranslate dataset **(MT part)** statistics in terms of the number of hypotheses-translation pairs and average length of ground-truth utterance in different language directions.