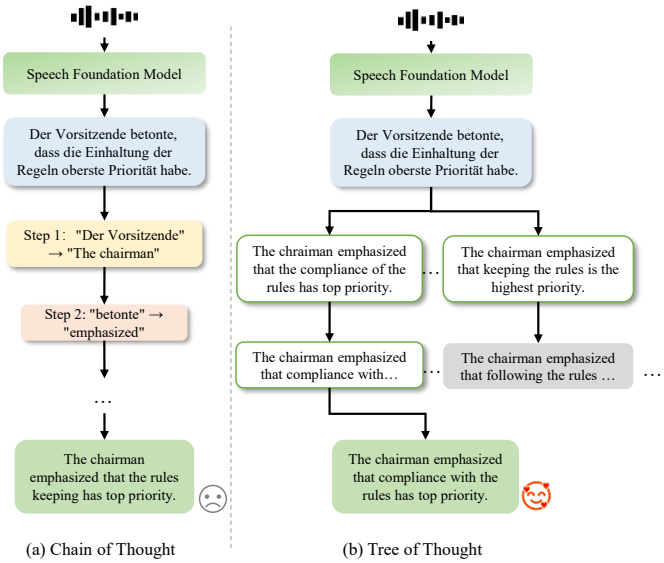
ToT-ST: Tree-of-Thought Reasoning with Large Language Models for Zero-Shot Speech Translation

*Dandan Xiao, Zhenbei Guo, Kaidi Zhu, Fei Ouyang, Yan Xiang, Zhengtao Yu**
Faculty of Information Engineering and Automation
Kunming University of Science and Technology, Kunming, China
xiaodandan1008@163.com, guozb777@163.com, 395021920@qq.com, 550058960@qq.com,
sharonxiang@126.com, ztyu@hotmail.com

***Abstract*—In recent years, combining pre-trained speech foundation models(SLM) with large language models (LLMs) has emerged as an effective paradigm for improving speech-to-text translation performance. However, most existing LLM-based speech translation methods rely on linear Chain-of-Thought (CoT) reasoning and single-path decoding strategies, which limits their ability to fully exploit diverse translation hypotheses and makes them prone to error propagation, especially under zero-shot settings. In this paper, we propose a novel Tree-of-Thought reasoning framework for speech translation task, termed ToT-ST, which reformulates speech translation as a structured reasoning problem. Instead of being restricted to a single reasoning trajectory, ToT-ST enables LLMs to generate, evaluate, and select multiple candidate translation paths in parallel, thereby explicitly modeling diverse translation strategies. This structured reasoning mechanism allows the model to better explore the translation hypothesis space effectively improving translation performance in zero-shot speech translation scenarios. We evaluate ToT-ST on multiple benchmark datasets, including FLEURS, CoVoST-2, and MuST-C, covering zero-shot and cross-lingual scenarios. Experimental results demonstrate that ToT-ST achieves competitive improvements over end-to-end models and LLM-based baselines in terms of BLEU and COMET scores.**

***Index Terms*—Speech Translation, Large Language Models, Tree-of-Thought, Zero-shot Learning, Structured Reasoning**



(a) Chain of Thought (b) Tree of Thought

Fig. 1: Translation process (e.g., De → En) of Chain-of-Thought and Tree-of Thought in LLMs.

I. Introduction

Speech Translation (ST) aims to directly map source-language speech signals into target-language text, constituting a representative multimodal and multi-task challenge that integrates automatic speech recognition, semantic understanding, and cross-lingual generation [1]. While both traditional cascade approaches and end-to-end models have achieved substantial progress, they still face significant limitations in low-resource languages, domain transfer, and robustness to acoustic noise.

In recent years, Large Language Models (LLMs) have profoundly reshaped the technical paradigm of speech translation, owing to their strong multilingual representations, deep contextual modeling, and instruction generalization capabilities. Rather than being treated merely as post-processing modules, LLMs are increasingly integrated into the encoding–reasoning–generation pipeline of ST systems, endowing them with more human-like semantic perception and error-correction abilities. A representative line of research adopts a "speech foundation model (SLM) + LLM" paradigm [2] [3] [4], in which speech inputs are first mapped to source-language transcripts or multiple candidate translations, followed by LLM-based refinement or aggregation. For example, GenTranslate [2] leverages the linguistic knowledge of LLMs to re-rank and fuse N-best translation hypotheses generated by speech foundation models, thereby producing higher-quality outputs. TGCoT [3] proposes a Translation Graph Chain-of-Thought [5] reasoning framework, enabling iterative feedback and back-translation to self-correct potential errors. LLM-based refinement strategies under diverse scenarios have also been explored [4], [6], [7], including iterative prompting and explicit incorporation of intermediate reasoning chains to encourage multi-step semantic reasoning prior to final generation.

*Corresponding author. This work was supported by the NSFC fund (U24A20334, 62376111), Yunnan Provincial Major Science and Technology Special Project (202502AD080014), Yunnan Provincial Basic Research Major Project (202401BC070021), Intergovernmental International Science, Technology and Innovation Cooperation Project (2025YFE0121700), Science and Technology Projects of Yunnan Universities Serving Key Industries (FWCY-ZD2025006), the Yunnan Province Young and Middle-aged Academic and Technical Leaders Reserve Talent Program (202305AC160063), and Yunnan Fundamental Research Projects (KKAE202603015).

Despite demonstrating the potential of LLMs as the core of linguistic reasoning and generation for speech translation, these approaches predominantly rely on linear Chain-of-Thought (CoT) reasoning paradigms. Under this framework, models generate intermediate reasoning steps and final translations along a single sequential path, making the process vulnerable to error propagation once mistakes occur in early stages [8]. As illustrated in Figure 1, the CoT wrongly optimized the translation prompt for text "Einhaltung der Regeln", resulting in a discrepancy with the original text and generating an unnatural phrase such as "rules keeping". Although backtracking mechanisms [9], [10] can alleviate certain errors, effective correction remains challenging and computationally expensive. Moreover, linear reasoning restricts the ability to explore multiple translation hypotheses in parallel, thereby limiting the search space of translation strategies and preventing the full exploitation of potentially useful information contained in alternative candidates.

To overcome these limitations, this paper introduces a Tree-of-Thought (ToT) [11]-based structured reasoning mechanism for speech translation, which enables the model to generate, evaluate, and select multiple candidate translation paths in parallel. By explicitly modeling alternative translation strategies, the proposed approach enhances robustness and stability. Transforming translation from a linear decision process into a structured tree-based search, our method more effectively exploits translation diversity and achieves more reliable performance under zero-shot speech translation settings.

The main contributions of this work are summarized as follows:

- We propose **ToT-ST**, a novel *Tree-of-Thought-based speech translation framework* that reformulates speech translation as a structured reasoning problem, enabling parallel exploration of multiple translation hypotheses instead of relying on a single linear reasoning path.
- We introduce a **tree-structured inference strategy** for LLM-based speech translation, in which candidate translations are iteratively generated, evaluated, and selected, effectively mitigating error propagation and improving robustness under zero-shot translation scenarios.
- Extensive experiments on multiple benchmarks, including **FLEURS, CoVoST-2, and MuST-C**, demonstrate that ToT-ST achieves competitive or superior performance compared to strong end-to-end and LLM-based baselines in both zero-shot and cross-lingual speech translation settings.

II. RELATED WORK

Zero-shot Speech Translation

Zero-shot Speech Translation focuses on directly translating speech into target-language text under the absence of target language pairs or speech–translation annotations, and represents a more challenging setting in speech translation research. Depending on the availability of supervision during training, this problem generally includes both zero-shot language pair scenarios and the more stringent zero-resource setting, where no speech–translation paired data is used during training. Some studies address this challenge by decomposing the general zero-resource problem into several simpler sub-tasks [12].

With the advancement of end-to-end speech translation models, researchers have attempted to alleviate data scarcity by introducing weak supervision or cross-lingual transfer learning. Certain approaches employ shared encoders [13], multi-task learning, or explicit distance constraints [14] [15] to map speech and text into a unified semantic space, enabling zero-shot translation. However, such methods typically still rely on a non-negligible amount of speech translation data for fine-tuning, and therefore fail to satisfy the strict zero-resource requirement.

More recent work has shifted toward systematic modeling of the modality gap, simultaneously addressing the mismatch between speech and text in both sequence length and representation space. Representative approaches such as ZEROSWOT [16] [17] [18] demonstrate that, without using any speech translation data, it is possible to achieve true zero-shot end-to-end speech translation by relying solely on automatic speech recognition and machine translation data. These methods employ CTC-based [19] subword compression and Optimal Transport [20] [21] to explicitly align speech and text representations, enabling the speech encoder to be directly integrated with large-scale multilingual machine translation models. Experimental results show that, when cross-modal alignment is sufficiently strong, speech translation performance can approach or even surpass that of certain supervised models.

In parallel, another research investigates LLM-based zero-resource speech translation and recognition frameworks, where speech representations are mapped into the token embedding space of LLMs and used as soft prompts [22] to directly drive text generation. These studies further indicate that, under zero-resource conditions, the performance ceiling of speech translation systems is largely determined by the reasoning and generation strategies of LLMs, rather than by acoustic modeling alone.

III. METHODS

A. Overall Framework

In this section, we introduce the design overview of ToT-ST as shown in Figure 2,including automatic speech recognition and Tree-of-Thought reasoning.

1) Speech Encoder: We employ Whisper-Large-V3 [23] as the speech foundation model to transcribe the source speech. The transcription is then processed sentence by sentence through LLM to generate translations in the target language.

2) Tree-of-Thought Reasoning: Given an input sentence, the proposed ToT framework performs structured multi-path reasoning. First, the input sentence is translated into an initial translation. Second, the ToT further refines the translation by breaking it down into multiple independent sentences or clauses. From the perspective of fluency and contextual coherence, an improved translation is generated for each

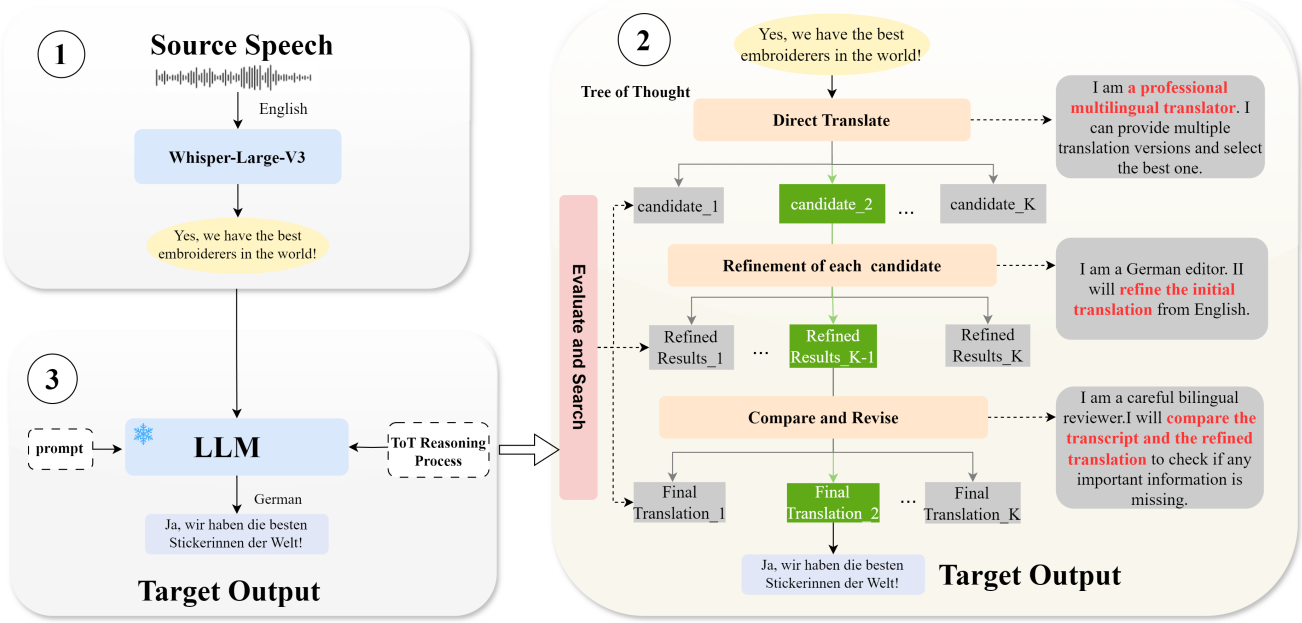


Fig. 2: Overall framework of ToT-ST (e.g., En → De). (1) Input layer: raw speech is encoded by a pre-trained Whisper-Large-V3 speech encoder to extract text. (2) Tree-of-Thought (ToT) reasoning process, which enables parallel generation, evaluation, and selection of multiple translation paths. (3) LLM backbone: Qwen3-8B serves as the core language model for reasoning and translation generation.

segment, resulting in a more accurate, natural, and coherent overall output. Finally, during the optimization process, we continuously perform comparative analyzes and supplement any missing information until the final translation is produced.

B. Tree-of-Thought Reasoning

In the section, we introduce the implementation detail of Tree-of-Thought Reasoning.

1) Generate: For any ToT node T_i and text x, it generates K candidate thought strategies for next node T_{i+1} according to Eq. (1):

$$k^j|(j=1,\ldots,K) \sim p_\theta(T_{i+1}|x,T_i) \tag{1}$$

where $p_\theta(T_{i+1}|x,T_i)$ denotes the probability of generating the next thought strategy T_{i+1} based on the current state (x,T_i), and θ denotes the current parameters of LLMs. Each candidate thought strategy $\{k^j|(j=1,\ldots,K)\}$ is independently generated from different thought spaces, ensuring the diversity of translation results and enhancing the translation performance.

2) Evaluate: For the branch nodes generated at each ToT layer, we use the proposed SATS to evaluate their current quality in the translation task and guide the generation of the thinking path, as shown in Eq. (2a) to Eq. (2c).

$$BLEU = BP \times exp\left(\sum_{n=1}^{N}(W_n)\,logP_n\right) \tag{2a}$$

$$Sim(u,v) = \frac{u\cdot v}{\|u\|\cdot\|v\|} = \frac{\sum_{i=1}^{d}u_i v_i}{\sqrt{\sum_{i=1}^{d}u_i^2}\cdot\sqrt{\sum_{i=1}^{d}v_i^2}} \tag{2b}$$

$$S = \alpha\cdot BLEU + (1-\alpha)\cdot Sim \tag{2c}$$

where Eq. (2a) and Eq. (2b) denotes the calculation equation for BLEU and semantic similarity, respectively. Both BLEU and semantic similar scores are normalized to $[0,1]$. Eq. (2a) uses N-Gram to to match the reference translation and the LLM translation results, calculating the BLEU score. The text of speech recognition is represented as a vector $u \in \mathbb{R}^d$, and the translated text is represented as a vector $v \in \mathbb{R}^d$. Eq. (2b) calculates the semantic similarity between v and u through the cosine similarity. Finally, Eq. (2c) performs weighted fusion. During the experiment, when $\alpha = 0.7$, the average score reaches the highest value.

3) Search: We use the breadth-first search strategy to fully explore the search space of ToT, as shown in Algorithm 1. In addition, during the exploration process, when the search depth of ToT is 3 and the branch width is 5, the overall performance tends to be stable. In Algorithm 1, x denotes the transcribed text by ASR and is regarded the initial state (line 1). For each step t, we calculate all possible extended states $S_t^{'}$. We use generator $G(p_\theta, s, k)$ to generate k candidate thoughts and attach them to state s to obtain a new state $S_t^{'} = (s, z_t)$ (lines 2-10). Then, we evaluate each new state to obtain the score V_t, choosing the top b states with the highest evaluation value V_t as the state set for the next step (lines 11-13). Finally, we select the state s^* with the highest evaluation value in S_T and use $G(p_\theta, s^*, 1)$ to generate the final result (lines 13-15).

IV. Experiment

A. Datasets

We evaluate our method on three multilingual speech translation datasets:

Algorithm 1 Breadth-First Tree-of-Thought Inference.

Input: Input x (text from Whisper-Large-V3), total steps T, generation width k, beam width b
Output: Final translation output

Initialize state set $S_0 = \{x\}$
for $t = 1$ to T **do**
  Initialize candidate state set $S'_t = \emptyset$
  for all $s \in S_{t-1}$ **do**
    Generate k candidate thoughts $\{z_t^{(1)}, \ldots, z_t^{(k)}\}$ using generator $G(p_\theta, s, k)$
    for $i = 1$ to k **do**
      Append $z_t^{(i)}$ to s to form new state $(s, z_t^{(i)})$
      Add the new state to S'_t
    end for
  end for
  Evaluate each state $s' \in S'_t$ using evaluation function $V_t(s')$
  Select the top b states with highest V_t scores to form S_t
end for
Let $s^* = \arg\max_{s \in S_T} V_T(s)$
Generate final output using $G(p_\theta, s^*, 1)$

- **MuST-C** [24]: It covers English to 15 target languages collected from TED talks.
- **FLEURS** [25]: It is released by Google for low-resource language research.
- **CoVoST-2** [26]: It is extended from CommonVoice [27] and is suitable for evaluating the generalization of model.

B. Metrics

- **SacreBLEU**: A standardized version of BLEU [28] that employs consistent tokenization and preprocessing methods, and is compatible with multilingual settings.
- **COMET** [29]: A semantic evaluation metric based on pre-trained language models, which measures the semantic similarity between generated translations and reference translations.

C. Baselines

We compare the proposed method with a diverse set of strong baselines covering different speech translation paradigms, including end-to-end models, LLM-based speech models, and SLM + LLM methods.

- **End-to-end Speech Translation Models**: We include **SeamlessM4T-Large** [30] and SeamlessM4T-Large-V2 as representative end-to-end speech translation systems trained on large-scale multilingual data. In addition, we include **ZEROSWOT** [16] as a representative zero-shot end-to-end speech translation model. These models serve as strong supervised baselines and reflect the performance of modern end-to-end ST approaches.
- **LLM-based Speech Models**: **Qwen2-Audio** [31] is used as a representative large language model that directly performs speech translation without explicit cascaded refinement. This baseline allows us to assess the effectiveness of structured reasoning beyond the raw generative capability of LLMs.
- **SLM + LLM Methods**: We compare against recent state-of-the-art cascaded approaches that combine ASR–MT pipelines with LLM-based refinement or reasoning, including **GenTranslate** [2], **GenTranslate-V2** [2], **TG-CoT** [3], **Refine$_{ST}$** [4] and **ZERO-ST** [22], a recent zero-resource speech translation framework. These methods leverage LLMs for hypothesis aggregation, iterative refinement, or Chain-of-Thought reasoning, and represent the strongest prior art in zero-shot speech translation.

D. Experiment Results Analysis

1) X→English(En): Table I reports speech translation BLEU scores from multiple source languages into English(X→En) on the FLEURS and CoVoST-2 test sets. First, compared with end-to-end speech translation models SeamlessM4T-Large and SeamlessM4T-Large-V2, our method achieves substantially higher BLEU scores on the FLEURS dataset across all evaluated languages, and attains the highest average BLEU score of 36.1. A similar trend is observed on the CoVoST-2 dataset, where our method consistently delivers superior performance on German(De-En), French(Fr-En), and Spanish(Es-En). Second, compared with Qwen2Audio, our method yields significant improvements, indicating that structured reasoning during translation contributes more to performance gains than the raw language modeling capability of LLMs alone.

More importantly, relative to state-of-the-art SLM + LLM approaches for zero-shot speech translation, our method establishes a new benchmark. Although GenTranslate, GenTranslate-V2, and TGCoT already outperform end-to-end models by leveraging LLM-based refinement, our approach further improves performance on both datasets. Specifically, ToT-ST achieves competitive BLEU scores on nearly all language pairs and attains the highest average BLEU on the FLEURS dataset, demonstrating superior robustness and consistency under zero-shot conditions.

2) English(En)→X: Table II reports speech translation BLEU scores from English to multiple target language (En→X) on the CoVoST-2 and MuST-C test sets. ToT-ST achieves competitive improvements across all evaluated datasets. On CoVoST-2, ToT-ST attains a BLEU score of 24.9 on En–Fa language pair, outperforming the strongest SLM + LLM baseline, TGCoT, by 1.8 BLEU points. For the En–Ja ToT-ST also improves BLEU scores by 0.2, demonstrating robust performance even when compared with other strong baselines. Although zero-shot models such as ZEROSWOT-MEDIUM and ZEROSWOT-LARGE achieve strong results, our approach remains competitive within LLM-based reasoning framework. On the MuST-C dataset, ToT-ST continues to exhibit superior performance. For example, on the En–Es translation task, ToT-ST achieves a BLEU score of 36.4, outperforming the ZEROSWOT-LARGE baseline by 0.6 BLEU points. On the En–It task, ToT-ST improves BLEU by 1.5

TABLE I: Speech translation results on FLEURS and CoVoST-2 X-En test sets in terms of BLEU score. We use bold to denote the best results and underline the second-best results to emphasize competitive performance.

X-En	Fleurs						CoVoST-2					
	Cy	El	Fa	Ta	Uk	Avg	De	Fr	Es	Nl	Ca	Avg
End-to-end Methods												
SeamlessM4T-Large [30]	31.7	25.9	32.2	19.8	30.5	28.0	38.8	41.3	41.1	41.1	38.4	40.3
SeamlessM4T-Large-V2 [30]	34.5	28.5	30.3	24.2	31.5	29.8	40.0	42.4	42.9	42.7	39.0	41.4
LLM												
Qwen2-Audio [31]	32.5	29.6	32.7	22.5	33.6	30.2	35.2	38.5	40.0	22.7	25.5	32.4
SLM + LLM Methods												
ZERO-ST [22]	—	—	—	—	—	—	32.4	35.6	34.6	23.3	25.8	30.3
GenTranslate [2]	39.4	32.8	35.9	24.1	36.9	33.8	40.2	41.8	42.1	43.8	38.4	41.3
GenTranslate-V2 [2]	39.1	33.8	35.4	26.4	37.1	34.4	41.1	43.1	43.3	**44.9**	**39.5**	**42.4**
TGCoT [3]	42.7	34.5	35.1	26.4	**38.5**	35.4	—	—	—	—	—	—
Ours	**43.0**	**36.1**	**36.5**	**28.1**	36.9	**36.1**	**41.2**	**45.8**	**44.5**	42.6	36.5	42.1

TABLE II: Speech translation results on CoVoST-2, MuST-C En-X test sets in terms of BLEU score.Remarks follow Table I.

En-X	CoVoST-2				MuST-C		
	Fa	Ja	Zh	Avg	Es	It	Avg
End-to-end Methods							
SeamlessM4T-Large [30]	18.3	24.0	34.1	25.5	34.2	29.9	32.1
SeamlessM4T-Large-V2 [30]	16.9	23.5	34.6	25.0	32.1	27.5	29.8
ZEROSWOT-MEDIUM [16]	26.0	46.0	39.0	37.0	34.9	30.6	32.8
ZEROSWOT-LARGE [16]	**26.6**	**47.1**	40.5	**38.1**	35.8	31.4	33.6
LLM							
Qwen2-Audio [31]	21.5	28.6	45.2	31.8	35.2	29.6	32.4
SLM + LLM Methods							
GenTranslate [2]	21.8	30.5	43.3	31.9	33.2	28.3	30.8
GenTranslate-V2 [2]	20.8	29.7	43.5	31.3	33.2	28.3	30.8
TGCoT [3])	23.1	32.3	**46.6**	34.0	35.5	30.3	32.9
Ours	24.9	32.5	44.8	34.1	**36.4**	**32.9**	**34.7**

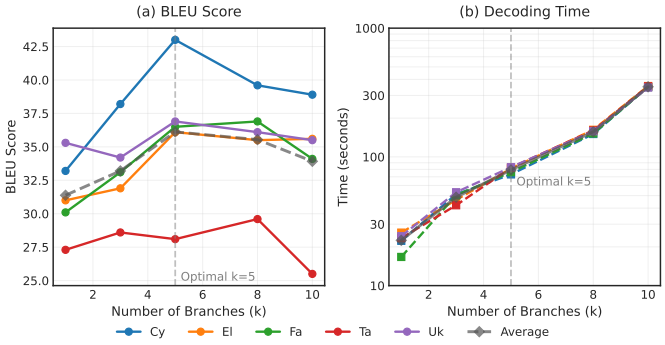


Fig. 3: Impact of tree branch number k on BLEU score and decoding time on FLEURS X-En. Subfigure (a) shows the effect on BLEU scores; subfigure (b) shows the decoding time. Colors denote target languages: blue (Cy), orange (El), green (Fa), red (Ta), purple (Uk), and gray (average).

over ZEROSWOT-LARGE, resulting in an average BLEU improvement of 1.1. These results indicate that ToT-ST remains effective even in the reverse translation direction, where target-side linguistic diversity introduces additional ambiguity and uncertainty. The observed gains highlight the effectiveness of Tree-of-Thought reasoning in handling such challenges under zero-shot conditions. To further assess translation quality, Table III reports results on CoVoST-2 En→X using both BLEU and COMET metrics. Across all evaluated languages, our method achieves the highest BLEU and COMET scores in almost all cases, as well as the best overall average performance. Compared with LLM-based refinement baselines such as Refine$_{Both}$ [4], Refine$_{ST}$ [4], and Paraphrase$_{ST}$ [4], our approach consistently improves performance on both metrics. The simultaneous gains in BLEU and COMET indicate that the proposed Tree-of-Thought structured reasoning enables more reliable translation decisions, leading to improvements not only in surface-level fluency but also in semantic adequacy and faithfulness.

E. Ablation Study

To investigate the impact of ToT reasoning on translation performance, we conduct ablation experiments on the CoVoST-2 X→En dataset, as shown in Table IV.

When only the initial translation step is used, BLEU drops significantly from 43.8 to 37.2, indicating that a single-pass generation is insufficient under zero-shot conditions. Removing either step 2 or step 3 leads to BLEU scores of 40.4 and 41.3, respectively, demonstrating the importance of iterative refinement and hypothesis comparison in the reasoning process.

Furthermore, removing the evaluation and scoring step results in an additional performance degradation, confirming that explicit selection based on quality estimation is a critical component of ToT. Overall, these results highlight that the performance gains of ToT arise from the joint effect of refinement, comparison, and evaluation, rather than from prompt design alone. Figure 3 illustrates the effect of translation candidate k on decoding performance. Since ToT relies on the power of LLMs and the number of translation candidates,

TABLE III: Speech translation results on CoVoST-2 test sets are reported in terms of BLEU and COMET scores. For each language, the left column shows BLEU results, and the right column shows COMET results.Remarks follow Table I.

CoVoST-2										
En-X	**De**		**Ca**		**Ar**		**Tr**		**Avg**	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Refine$_{Both}$ [4]	31.24	0.8230	39.62	0.8163	23.62	**0.8189**	22.28	0.8366	29.19	0.8237
Refine$_{ST}$ [4]	30.02	0.8021	38.21	0.8005	22.82	0.8103	21.92	0.8305	28.24	0.8108
Paraphrase$_{ST}$ [4]	28.43	0.7859	36.72	0.7823	21.95	0.7981	20.19	0.8067	26.82	0.7932
Ours	**32.11**	**0.8344**	**40.15**	**0.8202**	**23.70**	0.8079	**23.40**	**0.8536**	**29.84**	**0.8290**

TABLE IV: Translation performance of ToT on CoVoST-2 X-En datasets. The standard setting refers to the LLM directly generating a translation. "w/o step2" and "w/o step3" indicate the removal of the second and third reasoning steps, respectively. "w/o evaluate" removes the evaluation step, allowing the model to select the best translation autonomously.

X-En	**De**	**Fr**	**Es**	**Avg**
ToT	**41.2**	**45.8**	**44.5**	**43.8**
standard	35.6	36.9	39.1	37.2
w/o step2	39.2	40.3	41.8	40.4
w/o step3	38.7	41.4	43.8	41.3
w/o evaluate	36.8	35.2	39.9	37.3

TABLE V: Decoding time comparison (seconds per utterance) between ToT-ST and representative baselines on CoVoST-2 X→En. All methods are evaluated under the same hardware settings.

Method	**De**	**Fr**	**Es**	**Nl**	**Ca**	**Avg**
Qwen2-Audio [31]	46.8	44.8	42.0	44.1	46.2	44.8
Ours (Whisper + Qwen3-8B)	49.0	49.0	46.2	45.3	46.2	47.2
Ours (Whisper + Qwen2.5-7B)	2.9	4.1	3.5	2.6	3.6	3.3
Ours (Whisper + LLaMA3.1-8B)	3.8	3.7	4.5	3.4	4.1	3.9

we test translation candidate k in $\{1, 3, 5, 8, 10\}$. BLEU scores increase steadily as n grows up to 5, after which performance saturates or slightly declines. Smaller values of k limit the search space and hinder hypothesis diversity, while larger values introduce redundancy and increase the risk of hallucinations.In parallel, decoding time grows monotonically with increasing k, reflecting the computational cost of expanded search. Considering both translation quality and efficiency, we identify $k = 5$ as an optimal trade-off, which is used in all subsequent experiments.

Table V compares the decoding time of ToT-ST with representative baselines on the CoVoST-2 X→En test set. Results are reported in seconds per utterance and averaged over five source languages.Compared with Qwen2-Audio [31], ToT-ST incurs additional decoding cost due to multi-branch reasoning and hypothesis evaluation. However, the overhead remains moderate and scales linearly with the number of translation candidates.Among different LLM backbones, ToT-ST with Qwen3-8B achieves the best balance between translation quality and decoding efficiency, while smaller models such as Qwen2.5-7B offer reduced latency at the cost of translation performance. Overall, these results indicate that Tree-of-Thought reasoning improves translation robustness without introducing prohibitive computational overhead.

V. Conclusion

In this paper, we introduced ToT-ST, a Tree-of-Thought based structured reasoning framework for speech translation with large language models. Unlike conventional approaches that rely on linear Chain-of-Thought reasoning and single-path decoding, ToT-ST enables parallel generation and evaluation of multiple translation hypotheses, allowing the model to explicitly explore alternative reasoning paths and mitigate error propagation.Extensive experiments on FLEURS, CoVoST-2, and MuST-C demonstrate that the proposed approach consistently improves translation quality under zero-shot and cross-lingual settings, outperforming both end-to-end speech translation systems and recent LLM-based refinement methods. Further ablation studies validate the contribution of each reasoning component and show that an appropriate number of tree branches can achieve a favorable balance between translation accuracy and decoding efficiency.Our findings suggest that structured reasoning paradigms such as Tree-of-Thought provide a promising direction for advancing speech translation with large language models, particularly in low-resource and zero-supervision scenarios.

References

[1] C. Xu, R. Ye, Q. Dong, C. Zhao, T. Ko, and M. W. et al., "Recent advances in direct speech-to-text translation," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 2023, pp. 6796–6804.

[2] Y. Hu, C. Chen, C. H. Yang, R. Li, D. Zhang, and Z. C. et al., "Gentranslate: Large language models are generative multilingual speech and machine translators," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 74–90.

[3] C. Liu, L. Chen, P. Tang, W. Zhang, X. Li, and S. G. et al., "Large language models are efficient learners as zero-shot speech translators," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. IEEE, 2025, pp. 1–5.

[4] H. Dou, X. Tian, X. Lyu, J. Zhu, J. Li, and L. Guo, "Speech translation refinement using large language models," *CoRR*, vol. abs/2501.15090, 2025.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022.

[6] P. Chen, Z. Guo, B. Haddow, and K. Heafield, "Iterative translation refinement with large language models," in *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), EAMT 2024, Sheffield, UK, June 24-27, 2024*, C. Scarton, C. Prescott, C. Bayliss, C. Oakley, J. Wright, S. Wrigley, X. Song, E. Gow-Smith, R. Bawden, V. M. Sánchez-Cartagena, P. Cadwell, E. Lapshinova-Koltunski, V. Cabarrão, K. Chatzitheodorou, M. Nurminen, D. Kanojia, and H. Moniz, Eds. European Association for Machine Translation (EAMT), 2024, pp. 181–190.

[7] V. Raunak, A. Sharaf, Y. Wang, H. H. Awadalla, and A. Menezes, "Leveraging GPT-4 for automatic translation post-editing," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 12 009–12 024.

[8] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, and S. N. et al., "Self-consistency improves chain of thought reasoning in language models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[9] T. Qin, D. Alvarez-Melis, S. Jelassi, and E. Malach, "To backtrack or not to backtrack: When sequential search limits model reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2504.07052

[10] Y. Xu, Y. Zheng, S. Sun, S. Huang, B. Dong, and H. Z. et al., "Reason from future: Reverse thought chain enhances LLM reasoning," *CoRR*, vol. abs/2506.03673, 2025.

[11] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, and Y. C. et al., "Tree of thoughts: Deliberate problem solving with large language models," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.

[12] E. Dunbar, N. Hamilakis, and E. Dupoux, "Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1211–1226, 2022.

[13] R. Ye, M. Wang, and L. Li, "End-to-end speech translation via cross-modal progressive training," in *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 2267–2271.

[14] Q. Fang, R. Ye, L. Li, Y. Feng, and M. Wang, "STEMM: self-learning with speech-text manifold mixup for speech translation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 7050–7062.

[15] R. Ye, M. Wang, and L. Li, "Cross-modal contrastive learning for speech translation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022, pp. 5099–5113.

[16] I. Tsiamas, G. I. Gállego, J. A. R. Fonollosa, and M. R. Costa-jussà, "Pushing the limits of zero-shot end-to-end speech translation," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 14 245–14 267.

[17] P. Le, H. Gong, C. Wang, J. Pino, B. Lecouteux, and D. Schwab, "Pre-training for speech translation: CTC meets optimal transport," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 18 667–18 685.

[18] I. Tsiamas, G. I. Gállego, J. A. R. Fonollosa, and M. R. Costa-jussà, "Speech translation with foundation models and optimal transport: UPC at IWSLT23," in *Proceedings of the 20th International Conference on Spoken Language Translation, IWSLT@ACL 2023, Toronto, Canada (in-person and online), 13-14 July, 2023*, E. Salesky, M. Federico, and M. Carpuat, Eds. Association for Computational Linguistics, 2023, pp. 397–410.

[19] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, ser. ACM International Conference Proceeding Series, W. W. Cohen and A. W. Moore, Eds., vol. 148. ACM, 2006, pp. 369–376.

[20] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. A. Poggio, "Learning with a wasserstein loss," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 2053–2061.

[21] Y. Zhou, Q. Fang, and Y. Feng, "CMOT: cross-modal mixup via optimal transport for speech translation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 7873–7887.

[22] K. Mundnich, X. Niu, P. Mathur, S. Ronanki, B. Houston, and V. R. E. et al., "Zero-resource speech translation and recognition with llms," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. IEEE, 2025, pp. 1–5.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 28 492–28 518.

[24] M. A. D. Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: a multilingual speech translation corpus," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 2012–2017.

[25] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, and S. D. et al., "FLEURS: few-shot learning evaluation of universal representations of speech," *CoRR*, vol. abs/2205.12446, 2022.

[26] C. Wang, A. Wu, and J. M. Pino, "Covost 2: A massively multilingual speech-to-text translation corpus," *CoRR*, vol. abs/2007.10310, 2020.

[27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, and J. M. et al., "Common voice: A massively-multilingual speech corpus," *CoRR*, vol. abs/1912.06670, 2019.

[28] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pp. 311–318.

[29] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 2685–2702.

[30] S. Communication, L. Barrault, Y. Chung, M. C. Meglioli, D. Dale, N. Dong, and P. D. et al., "Seamlessm4t-massively multilingual & multimodal machine translation," *CoRR*, vol. abs/2308.11596, 2023.

[31] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, and Z. G. et al., "Qwen2-audio technical report," *CoRR*, vol. abs/2407.10759, 2024.