Fine-grained Video Dubbing Duration Alignment with Segment Supervised Preference Optimization

Chaoqun Cui[1], Liangbin Huang[1,2], Shijing Wang[3], Zhe Tong[1] Zhaolong Huang[1], Xiao Zeng[1], Xiaofeng Liu[2]*

[1]Alibaba Digital Media and Entertainment Group
[2]School of Software Engineering, Huazhong University of Science and Technology
[3]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University, Beijing 100044, China

Correspondence: cuichaoqun.ccq@alibaba-inc.com, liuxf@hust.edu.cn

Abstract

Video dubbing aims to translate original speech in visual media programs from the source language to the target language, relying on neural machine translation and text-to-speech technologies. Due to varying information densities across languages, target speech often mismatches the source speech duration, causing audio-video synchronization issues that significantly impact viewer experience. In this study, we approach duration alignment in LLM-based video dubbing machine translation as a preference optimization problem. We propose the Segment Supervised Preference Optimization (SSPO) method, which employs a segment-wise sampling strategy and fine-grained loss to mitigate duration mismatches between source and target lines. Experimental results demonstrate that SSPO achieves superior performance in duration alignment tasks.

1 Introduction

Video dubbing involves translating the original speech from a source language to a target language in visual media programs, relying on machine learning speech language processing techniques. Typically, video dubbing systems are not end-to-end but consist of three cascaded sub-tasks (Federico et al., 2020; Wu et al., 2023), namely Automatic Speech Recognition (ASR) (Yu and Deng, 2016; Chen et al., 2020), Neural Machine Translation (NMT) (Vaswani, 2017; Cheng and Cheng, 2019), and Text-to-Speech (TTS) (Wang et al., 2017; Ren et al., 2019; Li et al., 2019; Tan et al., 2021). ASR converts the original speech into text. When subtitles are available or can be obtained through Optical Character Recognition (OCR) (Memon et al., 2020; Nguyen et al., 2021), ASR can be bypassed. NMT is used to translate the source language text to the target language, after which TTS synthesizes the translated text into speech in the target language.

In video dubbing systems, maintaining strict isochronous constraints between the original source speech and the synthesized target speech in terms of speech duration is crucial for ensuring synchronization with the original video footage, which is vital for preserving an immersive experience for the audience (Wu et al., 2023). However, due to varying information densities across different languages, translating from one language to another often results in a duration mismatch between the source speech and the target speech (Tiedemann, 2012; Guzmán et al., 2014; Lakew et al., 2019). For instance, when translating from Chinese, a high information density language, to lower information density languages such as English or Thai, the resulting translations frequently exceed the timing notes of the original subtitles (see Table 7 for examples of subtitle format.), significantly impacting the audience's viewing experience. If relying solely on TTS to adjust the pause and duration of words, the vast differences in information density between languages necessitate that TTS adjusts the speaking rate of each word within a wide range to match the total speech duration. This can severely impact the fluency and naturalness of the synthesized speech, leading to a dissonance in the speaking rates between adjacent lines (Wu et al., 2023). Consequently, Duration Alignment (DA) is a significant challenge that must be addressed during NMT.

Recently, Large Language Models (LLMs) have been widely applied in NMT, bringing significant improvements to translation tasks (Hendy et al., 2023; Jiao et al., 2023; Zhu et al., 2024; Zhang et al., 2024). LLMs have also been applied to video dubbing NMT. However, relying solely on Prompt Engineering (PE) and Supervised Fine-tuning (SFT) on human-translated subtitles does not handle DA well. This is primarily because LLMs lack direct awareness of speech duration for the text, and the available human-translated subtitles used for SFT typically focus on the text itself

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4524–4546

rather than the speech duration of the lines. For DA, although generating a translation that is a few words shorter or longer may seem like a simple task, it actually requires good control over the target language. As illustrated in Table 3, LLMs must implicitly adopt strategies such as choosing more concise phrasing, using different verb tenses, avoiding redundant adverbs and adjectives (Lakew et al., 2019), while also maintaining translation accuracy and fluency to ensure that the translation quality does not deteriorate after DA.

In this study, we consider the DA task as a special preference optimization problem, termed as a localized multi-segment preference optimization problem. This framework addresses the situation where the output of LLMs consists of multiple semantically interconnected segments. The preference metric is evaluated for individual segments rather than the entire outcome (segment supervised), requiring segment-wise alignment across each segment. To tackle this issue, we propose the **S**egment **S**upervised **P**reference **O**ptimization (SSPO) method. SSPO effectively controls the duration of translations for each line at a fine-grained level while ensuring accuracy and fluency in handling DA tasks.

In summary, this study contributes as follows:

- We define duration consistency metrics and established an evaluation framework for DA.
- We propose the SSPO method, which formulates DA as preference optimization problem.
- We elucidate the theoretical foundation of SSPO and formalize the localized multi-segment preference optimization task.
- Experimental results demonstrate the effectiveness and robustness of SSPO.

2 Related Work

2.1 Duration Controllable Generation

Previous research on text length control includes: 1) using reward functions or models incorporating length information to guide decoding (Kikuchi et al., 2016; Murray and Chiang, 2018); 2) modifying model embeddings to inject length information (Lakew et al., 2019; Takase and Okazaki, 2019); and 3) interfering with training using length prediction metrics or models (Yu et al., 2021; Wu et al., 2023). However, these methods primarily target traditional sequence-to-sequence models and are unsuitable for LLMs. This is because LLMs are highly optimized through large-scale pre-training, and modifying their embeddings or interfering with training would significantly degrade overall performance (Nie et al., 2024; Sun and Dredze, 2024). Moreover, these methods aim to generate shorter texts, whereas DA seeks to produce translations with consistent durations to the original. Additionally, subtitle dialogues are short texts with strong contextual relationships (Bassnett, 2013; Cintas and Remael, 2014). In summary, unlike length control tasks aiming for overall shorter texts, DA's objective is to *generate fine-grained translations for each line in LLMs' responses, ensuring consistent (not necessarily shorter) durations for individual lines rather than the entire response*.

2.2 Language Model Preference Optimization

Reinforcement learning offers an effective solution for aligning LLMs with human values and controlling text generation (Bender et al., 2021; Bommasani et al., 2021; Thoppilan et al., 2022; Taori et al., 2023; Chiang et al., 2023; Ji et al., 2023). Reinforcement learning from human feedback (RLHF) framework has been developed, based on human feedback reward models (MacGlashan et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022b,a; Zheng et al., 2023). Despite RLHF's effectiveness, its complexity, instability, and hyperparameter sensitivity remain unresolved (Engstrom et al., 2020; Andrychowicz et al., 2021). Recently proposed DPO (Rafailov et al., 2024) simplifies the RLHF framework by eliminating the need for explicit reward modeling or reinforcement learning processes, thus avoiding dependence on reward models. Several variants have emerged, including SimPO, KTO, and IPO (Meng et al., 2024; Azar et al., 2024; Ethayarajh et al., 2024). However, these methods still face limitations such as coarse granularity and gradient dilution when addressing localized preference alignment tasks like DA.

3 Preliminaries

3.1 Notations

DA of video dubbing is essentially a task of controllable text generation (CTG) (Liang et al., 2024) task, which requires the LLM's output to: 1) strictly conform to the corresponding format, so that the translation of each line can be matched with the original; 2) maintain the duration of each line's translation as consistent as possible with the orig-

inal. Specifically, we leverage human-translated subtitles to perform SFT on an "off-the-shelf" LLM (Dubey et al., 2024; Bai et al., 2023; Du et al., 2022), and then perform DA on the SFT model. During SFT, the LLM's input prompt x includes the instruction, a terminology translation table, and a set of n source lines $s_1, s_2, \ldots, s_n$ to be translated. The LLM's response y includes the original and translated lines $s_1, t_1, s_2, t_2, \ldots, s_n, t_n$, where the original lines are output to avoid mismatches caused by model omissions or line merging. Although this generates more output tokens, it is crucial for ensuring the accuracy and correspondence of the translation. After DA, the output format of the LLM needs to be consistent with the SFT model, and it is necessary to maintain the duration consistency between s_i and t_i. In our experiments, we consistently set $n = 35$. Examples of the model's prompt and response are shown in Table 11 and Table 12 in Appendix E.1.

3.2 Preference Metric

In our experiments, we utilize Microsoft Edge's online TTS service `edge-tts`[1] to obtain the duration of dialogue lines. It can be replaced with any TTS component. We use `edge-tts` to synthesize speech for both the original text s_i and the translated text t_i of each dialogue line, and then acquire their respective durations $\text{Dur}(s_i)$ and $\text{Dur}(t_i)$. Subsequently, we employ the following metric to measure the duration consistency between s_i and t_i:

$$\begin{aligned}\mathcal{P}(s_i, t_i) =& \exp(\max(0, \text{Dur}(t_i) - \text{Dur}(s_i))) \\ &+ \max(0, \text{Dur}(s_i) - \text{Dur}(t_i)) - 1.\end{aligned} \tag{1}$$

$\mathcal{P}(s_i, t_i)$ represents the penalty imposed when $\text{Dur}(s_i)$ and $\text{Dur}(t_i)$ are inconsistent, as shown in Figure 1. When $\text{Dur}(t_i) > \text{Dur}(s_i)$, $\mathcal{P}(s_i, t_i)$ is an exponential term, and when $\text{Dur}(t_i) < \text{Dur}(s_i)$, $\mathcal{P}(s_i, t_i)$ is a linear term. This design is based on the consideration that for video dubbing, longer translation duration is less acceptable than shorter one, as they may cause the translated subtitles to exceed the timing notes range of the original subtitles. The lower $\mathcal{P}(s_i, t_i)$ is, the higher the duration consistency between s_i and t_i, and vice versa.

To evaluate the translation quality during the subsequent sampling process, we employ two widely used reference-free translation assessment models: `Unbabel/wmt23-cometkiwi-da-xxl` (denoted as KIWI-XXL) (Freitag et al., 2023) and `Unbabel/XCOMET-XXL` (denoted as XCOMET) (Guerreiro et al., 2024). Both models have 10B parameters and demonstrate high correlation with human judgments.

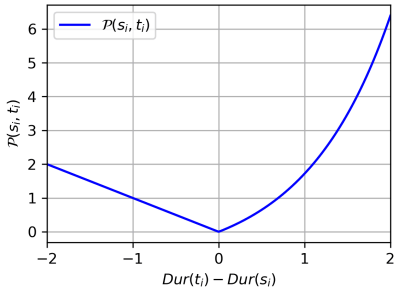


Figure 1: The function graph of $\mathcal{P}(s_i, t_i)$.

4 Method

4.1 Overall Framework

Although we have defined a quantitative duration consistency metric $\mathcal{P}(s_i, t_i)$ in Section 3.2, we are still unable to design a differentiable loss function to directly optimize the SFT model for DA. This is primarily because LLMs do not directly generate text, but rather predict the probability of token generation (Radford and Narasimhan, 2018). Consequently, utilizing a metric like $\mathcal{P}(s_i, t_i)$ to optimize the LLM directly through gradient descent is infeasible. In light of this, we approach DA as a preference optimization problem. Within the preference optimization framework, we can leverage the $\mathcal{P}(s_i, t_i)$ metric to guide the generation probabilities of the LLM, thereby optimizing the LLM's parameters in the direction of duration consistency.

However, we cannot directly apply preference alignment algorithms (such as DPO (Rafailov et al., 2024) or RLHF (Ouyang et al., 2022; Zheng et al., 2023)). This is primarily because the translation of each line of dialogue depends on its context, and the input to the SFT model needs to include multiple lines of dialogue. Consequently, fine-grained duration consistency alignment is required for each line of dialogue in the SFT model's response (see Table 12). Furthermore, further training of the SFT model must not alter the model's output format, as this could lead to issues such as translation omissions, resulting in synchronization problems with the timing notes of the original subtitles.

We present the overall framework of SSPO in

[1] https://github.com/rany2/edge-tts

Figure 2. During DA, SSPO employs a fine-grained segment-wise sampling strategy to sample multiple translation candidates for each line of dialogue from SFT model. It then selects preferred and non-preferred translations for each line based on the duration consistency metric $\mathcal{P}(s_i, t_i)$. Subsequently, it optimizes the SFT model using a segment-wise DPO loss function. Furthermore, to ensure that the DA model does not deviate significantly from the SFT model and to maintain consistency in the model's output format, we incorporate a token-level KL divergence penalty term to constrain parameter updates during the training process.

4.2 Sampling Strategy

We first utilize the demonstration dataset to obtain a basic model π_{sft} through SFT. For a sample $x \in \mathcal{D}_{\text{query}}$ from the query dataset $\mathcal{D}_{\text{query}}$ used for DA, which contains n lines of dialogue $s_1, s_2, \ldots, s_n$ (see Table 11 for examples). For each line s_i, we sample k translation results based on the prefix $x, s_1, t_1^{(\text{c})}, s_2, t_2^{(\text{c})}, \ldots, s_{i-1}, t_{i-1}^{(\text{c})}, s_i$, i.e., $\pi_{\text{sft}}(t_i|x, s_1, t_1^{(\text{c})}, \ldots, s_{i-1}, t_{i-1}^{(\text{c})}, s_i)$, obtaining $\{t_i^j | j = 1, 2, \ldots, k\}$. After deduplication and discarding the bottom 20% ranked by KIWI-XXL and XCOMET, we select the chosen translation $t_i^{(\text{c})}$ and rejected translation $t_i^{(\text{r})}$ based on the duration consistency metric $\mathcal{P}(s_i, t_i)$. Specifically, the sample with the minimum $\mathcal{P}(s_i, t_i)$ is chosen as $t_i^{(\text{c})}$, and the one with the maximum as $t_i^{(\text{r})}$ (see Table 1 for demonstration). Finally, for the sample $x \in \mathcal{D}_{\text{query}}$, we obtain the corresponding sampling result $\mathcal{S}(x) \equiv \{(s_i, t_i^{(\text{c})}, t_i^{(\text{r})}) | i = 1, 2, \ldots, n\}$. Algorithm 1 provide a detailed illustration of the entire sampling process.

The data sampling strategy in Algorithm 1 is predicated on the generation diversity of dialogue translations. Specifically, LLMs typically generate various translations t_i for most lines s_i, with each t_i having a distinct duration. However, simple lines such as "Good morning" and "How are you?" with lower translation diversity should not be utilized for model optimization. Furthermore, if the duration differences among various translations t_i are insignificant, they contribute little to model optimization. Consequently, based on the sampled data, we establish two thresholds, ε_1 and ε_2, to filter out lines with low translation diversity. Specifically, if the number of deduplicated samples from k samples of s_i is less than ε_1 or $\mathcal{P}(s_i, t_i^{(\text{r})}) - \mathcal{P}(s_i, t_i^{(\text{c})}) < \varepsilon_2$, the line s_i will not be involved in the preference optimization process. In our experiment, we set $\varepsilon_1 = 4$ and $\varepsilon_2 = 0.08$. Ultimately, we obtain the dataset $\mathcal{D}_{\text{dpo}}$ used for DPO training.

Algorithm 1 DPO Sampling Strategy.

Input: SFT model π_{sft}, query dataset $\mathcal{D}_{\text{query}}$, sampling number k.
Output: sampled sentence-level pairs set $\mathcal{S}(x)$.

// Iterate through the query dataset $\mathcal{D}_{\text{query}}$.
for any $x \in \mathcal{D}_{\text{query}}$ **do**
  // Iterate through the dialogue lines in x.
  for $i = 1$ to n **do**
    // Sample multiple candidates.
    for $j = 1$ to k **do**
      Sample $\pi_{\text{sft}}(t_i^j|\text{prefix})$
    end for
    Deduplicate $\{t_i^j | j = 1, \ldots, k\}$.
    Measure $\{t_i^j | j = 1, \ldots, k\}$ by $\mathcal{P}$.
    Select chosen $t_i^{(\text{c})}$ and rejected $t_i^{(\text{r})}$.
  end for
end for
return $\mathcal{S}(x) \equiv \{(s_i, t_i^{(\text{c})}, t_i^{(\text{r})}) | i = 1, \ldots, n\}$.

4.3 Alignment Loss Optimization

Unlike preference alignment tasks for language models, DA task requires fine-grained alignment of multiple segments within the LLM's response, rather than aligning the entire response as in preference alignment. Additionally, due to the contextual dependencies in dialogue translation, DA must ensure that the LLM output format (see Table 12) remains unchanged to prevent interference with the correspondence between the original line and its translation. SSPO utilizes DPO loss and sampled data to achieve fine-grained alignment of the duration for each line of dialogue. SSPO similarly requires the scheduling of two models: the policy π_θ and the reference π_{ref}, both of which are initialized from the SFT model π_{sft}. Specifically, for a sample $(x, \mathcal{S}(x)) \in \mathcal{D}_{\text{dpo}}$ from the sampled DPO dataset $\mathcal{D}_{\text{dpo}}$, we employ the standard DPO loss (Rafailov et al., 2024) to calculate a DPO loss term for a single line of dialogue s_i based on $(s_i, t_i^{(\text{c})}, t_i^{(\text{r})})$:

$$\mathcal{L}_{\text{dpo}}(s_i) = \log \sigma \left(\beta \log \frac{\pi_\theta(t_i^{(\text{c})}|p_i)}{\pi_{\text{ref}}(t_i^{(\text{c})}|p_i)} - \beta \log \frac{\pi_\theta(t_i^{(\text{r})}|p_i)}{\pi_{\text{ref}}(t_i^{(\text{r})}|p_i)} \right), \quad (2)$$

where p_i is $x, s_1, t_1^{(\text{c})}, s_2, t_2^{(\text{c})}, \ldots, s_{i-1}, t_{i-1}^{(\text{c})}, s_i$, and β is a hyperparameter used to control the sen-

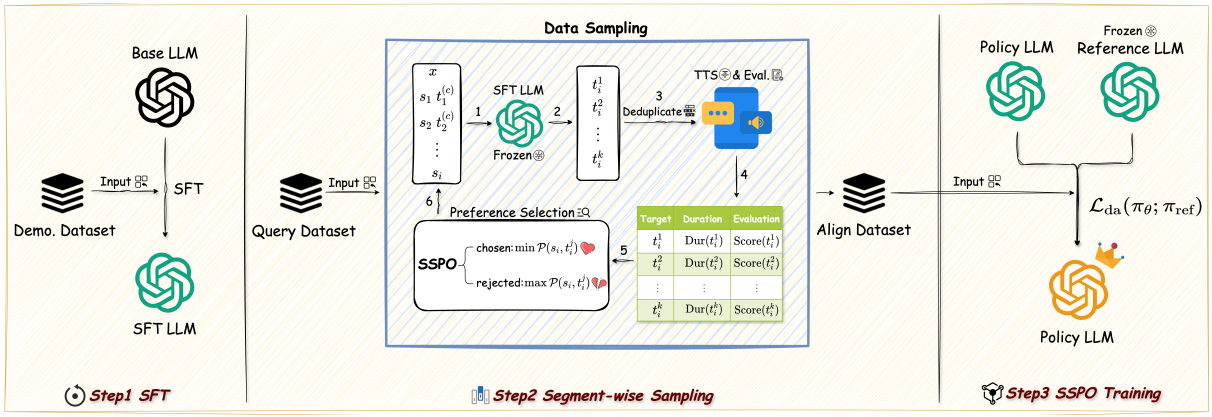


Figure 2: The overall framework of SSPO.

Line	Duration(s)	Evaluation	Operation
历史虽然会重演，但是人类是无法回到过去的。	2.89	-	-
History repeats, but we can't go back to what was.	2.66	85.6	-
History might replay, but mankind cannot go back in time.	2.73	84.2	discard
Even if history repeats, the past remains forever inaccessible to us.	**2.93**	89.3	chosen
Although history may repeat itself, humans cannot return to the past.	3.03	**91.4**	-
History often echoes, yet there's no way for us to turn back the clock.	3.19	89.8	rejected

Table 1: Chosen and rejected translation selection. The evaluation score is the average of KIWI-XXL and XCOMET.

sitivity of the optimization process to reward differences. $\mathcal{L}_{\text{dpo}}(s_i)$ only controls the duration of t_i without affecting other lines in x, thereby achieving independent customized DA for each line. We can now derive the loss function for DA as follows:

$$\mathcal{L}_{\text{da}}(\pi_\theta; \pi_{\text{ref}}) = - \mathbb{E}_{(x,\mathcal{S}(x))\sim\mathcal{D}_{\text{dpo}}} \left(\sum_{i=1}^{n} \mathcal{L}_{\text{dpo}}(s_i) \right). \quad (3)$$

4.4 Output Format Control

DA, as a CTG task, utilizes $\mathcal{L}_{\text{da}}(\pi_\theta; \pi_{\text{ref}})$ to achieve precise control over the duration of translated dialogue. However, it fails to maintain the consistent output format of LLMs. This limitation primarily stems from the fact that vanilla DPO is designed for open-ended generation tasks (Rafailov et al., 2024; Kong et al., 2025), relying on sentence-level KL divergence constraints, which are negligible for generation tasks with fixed output formats. Consequently, DA requires more stringent constraints. We adapt two methods to constrain the generation format: Token-level KL Divergence (TKLD) constraints and Low-Rank Adaptation (LoRA) training (Hu et al., 2021).

4.4.1 Token-level KL Divergence

TKLD constraint is employed to regulate the token generation distribution output by the policy model π_θ. During the model optimization process, this constraint ensures that the output distribution of π_θ remains as consistent as possible with that of the reference model π_{ref}. This approach not only guarantees the consistency of output formats between π_θ and π_{ref}, but also prevents π_θ from deviating too far from π_{ref}, thereby ensuring that the translation quality of the model after DA does not significantly deteriorate. The loss function incorporating the TKLD constraint is as follows:

$$\mathcal{L}_{\text{tkld}}(\pi_\theta; \pi_{\text{ref}}) = \mathcal{L}_{\text{da}}(\pi_\theta; \pi_{\text{ref}}) + \lambda \cdot \sum_t \text{KL}(\pi_\theta(\cdot|x, y_t), \pi_{\text{ref}}(\cdot|x, y_t)), \quad (4)$$

where λ is a hyperparameter used to control the constraint strength, and y_t represents all tokens generated at step t.

4.4.2 Low-Rank Adaptation Training

In addition to TKLD constraint, employing LoRA training can also help maintain the output format of the policy model, preventing model collapse while significantly reducing computational resource requirements during the training process. However, the LoRA training process converges more slowly compared to full-parameter training, necessitating a greater number of training iterations.

5 Experiments

5.1 Experimental Settings

We use our custom PolySC dataset for zh⇒en and zh⇒th translation experiments. Each direction's dataset is further divided into Demonstration and Query datasets for SFT and DA training. Additionally, we reserve 4 television series for the test set, ensuring these data are not present in the training set. Details of PolySC is shown in Appendix A.1.

We compare SSPO with the following baselines:

- **AutoDubbing** (Federico et al., 2020) models isochrony by controlling verbosity of NMT.
- **VideoDubber** (Wu et al., 2023) constructs a speech-aware length-controlled NMT model.
- **GPT-3.5-Turbo** is an early chat language model released by OpenAI (`gpt-3.5-turbo-0125`).
- **GPT-4o**[2] is OpenAI's current most advanced multimodal model (`gpt-4o-2024-11-20`).
- **Claude 3.5 Sonnet**[3] is a multimodal model released by Anthropic in June 2024.
- **Llama3.1-8B-Chinese-Chat**[4], **GLM-4-9B-Chat** (GLM et al., 2024), and **Qwen2.5-14B-Instruct** (Yang et al., 2024) are open-source language models released by Meta Platforms Inc., Zhipu AI, and Alibaba Group respectively, used as foundation models for SSPO.

For detailed experimental settings, refer to Appendix A. The source code for SSPO's data sampling and training is available at https://github.com/CcQunResearch/SSPO.

5.2 Results and Discussion

In this subsection, we present SSPO evaluation results, visualizations, and case studies.

5.2.1 Main Experiments

In Table 2, we present the main evaluation experiments for SSPO, reporting six metrics: S>T Rate, S>T Dur, T>S Rate, T>S Dur, Consistency Rate (CR), and $\mathcal{P}$. These metrics respectively represent the proportion of lines where the source duration exceeds the target duration by more than 0.1s and the average excess duration (s), the proportion of lines where the target duration exceeds the source duration by more than 0.1s and the average excess duration (s), the proportion of lines where the difference between source and target durations is within 0.1s, and the average value of the duration consistency metric $\mathcal{P}$. Additionally, we compare our results with the Gold Reference human translations from the test set and the Alignment Bound of DA. It is important to note that while DA aims to minimize $\mathcal{P}$ by making the duration of translated lines as consistent as possible with the source lines, achieving perfect consistency ($\mathcal{P} = 0$) is infeasible. This is because, when maintaining translation quality, the most duration-consistent translation for each line typically does not yield a $\mathcal{P}$ of 0 with the source (see the example in Table 1). Consequently, DA has an upper limit, termed the Alignment Bound, which is inaccessible. However, we can estimate this bound by calculating the average $\mathcal{P}$ between all chosen translations and their corresponding source lines in the data sampled using Algorithm 1.

Results in Table 2 demonstrate that after SSPO training, SFT models show a significant decrease in $\mathcal{P}$ and a notable increase in the dialogue duration consistency rate, outperforming other baselines. SSPO produces consistent alignment effects across different base models, validating its universal applicability. GPT-3.5, GPT-4, and Claude 3.5, which use PE to control translation duration (see Appendix E.2 for the prompt design), showed some improvement compared to the gold reference, but failed to match the performance of traditional methods like AutoDubbing and VideoDubber. This indicates that LLMs inherently lack sufficient perception of text duration and require additional duration information to effectively complete DA tasks. For LLMs, it is challenging to design a differentiable loss function that incorporates extra duration information to directly optimize them without modifying their underlying embeddings, model architecture, or introducing additional model parameters. Therefore, SSPO approaches this as a preference optimization problem, achieving DA through fine-grained sampling and training.

In addition, we conduct experiments on Spanish-related zh⇒es and es⇒zh translations in Appendix B.1, which also demonstrate similar performance. And we explore two other alternative solutions for DA in Appendix B.2.

[2] https://platform.openai.com/docs/models
[3] https://docs.anthropic.com/en/api
[4] https://huggingface.co/shenzhi-wang/Llama3.1-8B-Chinese-Chat

Method	Train	zh⇒en						zh⇒th					
		S>T Rate	S>T Dur	T>S Rate	T>S Dur	CR	$\mathcal{P}$	S>T Rate	S>T Dur	T>S Rate	T>S Dur	CR	$\mathcal{P}$
Gold Reference	-	*18.0%*	*0.344*	*64.1%*	*0.464*	*17.9%*	*0.501*	*19.4%*	*0.369*	*60.2%*	*0.460*	*20.3%*	*0.489*
AutoDubbing	SFT	22.6%	0.355	56.4%	0.400	21.0%	0.388	20.5%	0.300	51.7%	0.388	27.8%	0.334
VideoDubber	SFT	26.5%	0.363	51.3%	0.371	22.2%	0.344	23.3%	0.305	47.8%	0.369	29.0%	0.314
GPT-3.5-Turbo	PE	**9.2%**	**0.267**	73.1%	0.465	17.7%	0.526	**14.2%**	0.302	65.2%	0.487	20.7%	0.567
GPT-4o	PE	14.7%	0.295	66.1%	0.407	19.2%	0.417	18.2%	0.299	54.9%	0.353	27.0%	0.318
Claude 3.5 Sonnet	PE	**11.4%**	**0.267**	69.3%	0.401	19.3%	0.410	**14.8%**	**0.274**	57.7%	0.342	27.5%	0.313
Llama3.1-8B-CN-Chat	SFT	22.7%	0.352	56.6%	0.398	20.7%	0.389	17.2%	0.287	56.4%	0.410	26.4%	0.370
	SSPO	31.7%	0.341	**42.4%**	**0.309**	**25.9%**	**0.263**	30.7%	**0.285**	**32.9%**	**0.279**	**36.4%**	**0.206**
GLM-4-9B-Chat	SFT	19.5%	0.342	60.5%	0.427	20.0%	0.428	18.1%	0.291	55.0%	0.391	27.0%	0.360
	SSPO	29.6%	0.350	45.9%	**0.323**	24.5%	0.283	25.9%	0.291	42.0%	0.318	32.1%	0.254
Qwen2.5-14B-Instruct	SFT	20.0%	0.341	59.8%	0.439	20.2%	0.423	18.2%	0.294	55.4%	0.397	26.4%	0.362
	SSPO	34.4%	0.366	**40.6%**	0.324	**24.9%**	**0.272**	38.6%	0.290	**25.3%**	**0.279**	**36.1%**	**0.198**
Alignment Bound	-	*16.4%*	*0.278*	*39.3%*	*0.331*	*44.3%*	*0.220*	*9.2%*	*0.232*	*40.4%*	*0.313*	*50.4%*	*0.203*

Table 2: zh⇒en and zh⇒th results on test set. The best and second best results are denoted as **blue** and **orange**.

5.2.2 Visualization and Case Study

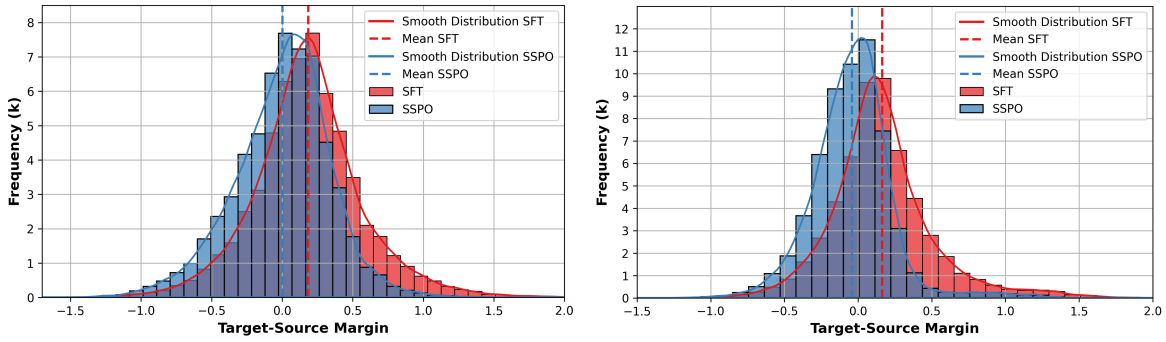
In Figure 3, we present the frequency distribution of the duration difference between the translation and the original text for both SFT and SSPO models of Qwen2.5-14B-Instruct, to observe the changes in translation duration after SSPO alignment. It is evident that after SSPO training, the duration discrepancy between the original text and the translation significantly narrows. This is reflected in Figure 3, where the histogram for SSPO is noticeably more concentrated around zero compared to that of SFT. Additionally, in Table 3, we showcase comparative case studies of translations for certain lines by the SFT and SSPO models of Qwen2.5-14B-Instruct. These examples visually demonstrate that the translations aligned by SSPO exhibit greater duration consistency with the original text compared to those generated by the SFT model.

5.3 Human Evaluation of Translation Quality

We did not utilize traditional translation quality evaluation metrics such as BLEU and ROUGE. These metrics overlook the semantics of the translation, lack contextual understanding, and cannot handle the diversity and flexibility of LLM translations. Therefore, we completely abandoned these metrics in favor of human evaluation. In Table 4, we present the human evaluation results for the translation quality of SSPO. We conducted evaluations in both the zh⇒en and es⇒zh directions. For each direction, we employed four evaluators, all of whom are bachelor's or master's degree professionals in English or Spanish translation, with Chinese as their native language. Due to the subjective preferences of different evaluators, we did not use scoring in the human evaluation. Instead, we performed pairwise comparisons of different translations to assess the win rate metric.

The SSPO model was evaluated against four baselines: the gold reference, vanilla base model, GPT-4o, and the SFT model, across the dimensions of accuracy, naturalness, and vividness: 1) *Accuracy*: Does the translation accurately convey the original meaning of the dialogue? 2) *Naturalness*: Is the translation fluent and does it conform to the grammar and lexical conventions of the target language? 3) *Vividness*: Is the translation expressive and does it convey the emotion and ambiance of the original dialogue? Additionally, we conducted a comprehensive evaluation, with instructions provided to evaluators as referenced in Appendix E.3, similar to those for other dimensions. The multidimensional evaluation across both directions comprised a total of 64 evaluation tasks, with each evaluator randomly assigned 8 tasks. In each evaluation task, we provided evaluators with challenger and competitor translations of 200 dialogue segments from the test set, each segment containing 20 lines of dialogue. Evaluators were required to select the superior translation or mark both as "no significant difference." The dialogue segments for different evaluation tasks were randomly selected subsets from the test set.

Translations by LLM-based methods, like SSPO and GPT-4o, often surpass human in terms of accuracy but fall short in vividness. Human translators typically reference scene and emotional cues from the video and audio, which LLMs are currently unable to incorporate. Future research should focus on enhancing translation models' ability to perceive and understand multimodal information to achieve more vivid localized translations. SSPO shows significant improvement over its vanilla base model, demonstrating the positive impact of fine-tuning LLMs on visual media data. The comparison between SSPO and SFT model further highlights the influence of SSPO on model performance. In translations from high-to-low information den-



(a) zh⇒en (b) zh⇒th

Figure 3: Frequency distribution of Qwen2.5-14B-Instruct model on zh⇒en and zh⇒th translations.

Type	Source	Target	Model
T>S	灵王的交代您还记得吧？ (1.52s)	Do you still remember what the Spirit King told you? (2.37s)	SFT
		Do you remember what the Spirit King said? (1.97s)	SSPO
T>S	你们摆平那群玄门人了？ (1.53s)	Have you dealt with those arcanists? (1.61s)	SFT
		You've dealt with those arcanists? (1.55s)	SSPO
T>S	或许你我可以尝试做一对有情人 (1.75s)	Maybe you and I can try to be a pair of lovers. (2.32s)	SFT
		Maybe you and I can be that lovebird. (1.86s)	SSPO
S>T	必须赶紧去取幽冥赋了 (1.82s)	I must go and get the Ghost Charmer. (1.7s)	SFT
		I have to retrieve the Ghost Charmer soon. (1.83s)	SSPO
S>T	这中餐合你胃口吗？ (1.44s)	Do you like Chinese food? (1.12s)	SFT
		Is Chinese cuisine palatable to you? (1.54s)	SSPO

Table 3: Case studies of Qwen2.5-14B-Instruct model on zh⇒en translation.

Challenger	Competitors	zh⇒en				es⇒zh			
		Accuracy	Naturalness	Vividness	Comprehensive	Accuracy	Naturalness	Vividness	Comprehensive
SSPO (GLM-4-9B)	Gold Reference	27:50:23	16:67:17	23:51:26	24:45:31	19:61:20	23:55:22	24:45:31	24:48:28
	Vanilla GLM-4-9B	26:50:24	21:60:19	24:55:21	27:50:23	28:51:21	26:52:22	27:51:22	28:53:19
	GPT-4o	23:50:27	19:57:24	23:51:26	20:52:28	23:55:22	19:58:23	20:56:24	22:52:26
	SFT	23:52:25	18:60:22	24:49:27	29:40:31	23:57:20	23:56:22	27:45:28	27:51:22
SSPO (Qwen2.5-14B)	Gold Reference	21:63:16	24:55:21	23:51:26	24:50:26	21:59:20	24:55:21	25:48:27	27:49:24
	Vanilla Qwen2.5-14B	24:51:25	26:50:24	26:53:21	32:41:27	31:45:24	26:51:23	27:48:25	30:43:27
	GPT-4o	25:51:24	19:58:22	23:48:29	23:49:28	25:53:22	27:51:22	24:50:26	31:43:26
	SFT	23:48:29	21:54:25	24:49:27	24:50:26	27:49:24	33:38:28	25:51:24	30:46:24

Table 4: Human translation quality evaluation, reporting win rate (win:tie:loss). The winning and losing contrasts are marked in blue and orange, respectively.

sity, SSPO usually reduces the length of generated translations, which inevitably results in some information loss as the model sacrifices some translation quality for better duration control. Conversely, in low-to-high information density translations, the model tends to generate longer (more informative) translations, thereby improving translation quality. Unlike the strict accuracy requirements in legal text translations, subtitle translation can tolerate some loss of accuracy because viewers can rely on other modalities, such as video and audio, to supplement their understanding of the program's current scene, even if some information is lost in the translation.

5.4 Ablation Study

We conducted a series of ablation studies to investigate the impact of various factors on SSPO.

5.4.1 Format Control Measures

As a CTG task, DA requires precise control over the duration of each line while also ensuring that the model's output adheres to the format shown in Table 12. In Section 4.4, we proposed two format control methods: TKLD and LoRA training, and compared the three configurations of full parameter fine-tuning, TKLD, and LoRA training in Table 5 in terms of the output conforming to the required format. We conducted experiments with two models on translation tasks in two languages, where the efficient rate represents the proportion of lines in the test set that conform to the format. The results show that full parameter fine-tuning leads to a significant drop in the efficient rate, and the model may encounter issues such as complete

output collapse, omission of certain lines, or failure to adhere to the required format. Both LoRA and TKLD are able to maintain the output format after SSPO alignment, with LoRA achieving an efficient rate close to 100%. Moreover, LoRA requires less GPU memory compared to TKLD. Therefore, we recommend using LoRA for SSPO training.

Base Model	Train	zh⇒en		zh⇒th	
		Efficient Rate	$\mathcal{P}$	Efficient Rate	$\mathcal{P}$
Llama3.1-8B	-	89.6%	**0.251**	85.1%	**0.197**
	TKLD	98.1%	0.273	97.4%	0.203
	LoRA	**99.7%**	0.263	**99.8%**	0.206
Qwen2.5-14B	-	81.2%	**0.258**	73.9%	0.202
	TKLD	96.9%	0.283	97.2%	0.209
	LoRA	**99.8%**	0.272	**99.7%**	**0.198**

Table 5: The impact of format control measures.

5.4.2 Reward Difference

The hyperparameter β in SSPO loss controls the model's sensitivity to implicit reward differences. We investigated the impact of β on DA performance using two models for translation in two languages. Figure 4 reports the changes in the duration consistency metric $\mathcal{P}$ and the format efficiency rate as β varies. The results indicate that as β increases, $\mathcal{P}$ steadily increases, suggesting that smaller values of β lead to higher duration consistency. However, when using smaller β values, occasional decreases in the format efficiency rate were observed, with the model showing instances of non-adherence to the output format. Therefore, considering these factors comprehensively, we opted for a moderate value of $\beta = 0.5$ in our experiments.

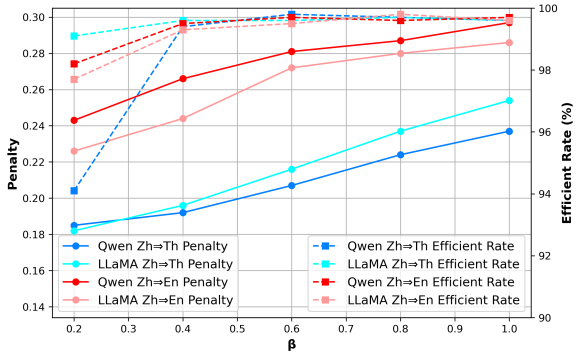


Figure 4: The impact of hyperparameter β.

5.4.3 Data Scale

Another question worth exploring is "How much data does SSPO require to achieve acceptable DA performance?" To address this, we investigated the impact of the number of dialogue lines from the Query dataset on performance using Qwen2.5-14B-Instruct for translation in two languages. Figure 5 reports the changes in the duration consistency metric $\mathcal{P}$ and the format efficiency rate as the data scale varies. The results show that as the data scale increases, $\mathcal{P}$ gradually decreases, indicating that using more data yields better alignment effects. However, employing an excessive amount of data leads to a sharp decline in the format efficiency rate. Considering these factors, we used approximately 10,000 dialogue lines in our experiments, which is equivalent to about 600 prompt-response pairs from the Query dataset, representing roughly 3% of the entire PolySC dataset. This approach achieves notable performance, demonstrating that SSPO does not require large amounts of data, and significant improvements in duration consistency can be achieved using a relatively small dataset.

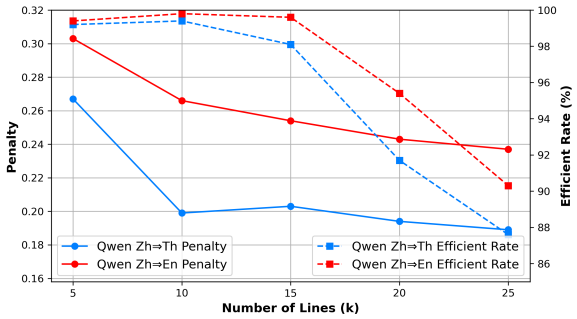


Figure 5: The impact of data scale.

6 Conclusion

In this study, we focus on the duration alignment task in video dubbing, which we consider as a preference optimization problem. To address this, we propose Segment Supervised Preference Optimization (SSPO) method. SSPO employs segment-wise sampling strategy and fine-grained preference alignment loss to mitigate the duration mismatch between source and target lines. Experiments demonstrate that SSPO achieves significant improvements in enhancing duration consistency between source and target speech compared to baseline methods, while maintaining translation quality.

Acknowledgments

The authors would like to thank all the anonymous reviewers for their help and insightful comments.

Ethical Statement

This research has been conducted with adherence to ethical guidelines and standards. The data utilized in the study, specifically the subtitles of film and television programs, were sourced from the Youku platform. All data collection and usage were performed following formal authorization and consent from Youku, ensuring that permissions were fully granted for academic research purposes.

The study respects intellectual property rights and confidentiality agreements, complying with all terms and conditions as stipulated by Youku. No personal or sensitive information was gathered or used during this research. The focus of the study remains strictly on the linguistic and translational aspects of the subtitled content.

We maintain a commitment to transparency and ethical integrity in research, ensuring that all findings and methodologies are presented honestly and without misrepresentation. This research seeks to contribute valuable insights to the field of subtitle translation while upholding the highest ethical standards.

Limitations

Limitations of this study are listed as follows:

Emotion Induced Duration Variability. SSPO measures duration consistency by referencing the duration of synthesized speech from open-source TTS services. However, the duration of real visual media speech may vary due to factors such as character emotion, suggesting that the metric for duration consistency could be further optimized in future research.

Language Dependent Alignment Limits. There is an upper limit to optimization in DA tasks. As demonstrated in experiments involving Spanish, inherent language characteristics can prevent the complete resolution of duration inconsistency, even under optimal conditions.

References

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. 2021. What matters for on-policy deep actor-critic methods? a large-scale study. In International conference on learning representations.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In International Conference on Artificial Intelligence and Statistics, pages 4447–4455. PMLR.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.

Susan Bassnett. 2013. Translation studies. routledge.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages 610–623.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Yang Chen, Weiran Wang, and Chao Wang. 2020. Semi-supervised asr by end-to-end self-training. arXiv preprint arXiv:2001.09128.

Yong Cheng and Yong Cheng. 2019. Semi-supervised learning for neural machine translation. Joint training for neural machine translation, pages 25–40.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See https://vicuna. lmsys. org (accessed 14 April 2023), 2(3):6.

Jorge Díaz Cintas and Aline Remael. 2014. Audiovisual translation: subtitling. Routledge.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335.

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

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2020. Implementation matters in deep policy gradients: A case study on ppo and trpo. In International Conference on Learning Representations.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. arXiv preprint arXiv:2402.01306.

Marcello Federico, Robert Enyedi, Roberto Barra-Chicote, Ritwik Giri, Umut Isik, Arvindh Krishnaswamy, and Hassan Sawaf. 2020. From speech-to-speech translation to automatic dubbing. In Proceedings of the 17th International Conference on Spoken Language Translation, pages 257–264.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. 2023. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In Proceedings of the Eighth Conference on Machine Translation, pages 578–628.

Bofei Gao, Feifan Song, Yibo Miao, Zefan Cai, Zhe Yang, Liang Chen, Helan Hu, Runxin Xu, Qingxiu Dong, Ce Zheng, et al. 2024. Towards a unified view of preference learning for large language models: A survey. arXiv preprint arXiv:2409.02795.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. arXiv preprint arXiv:2406.12793.

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

Francisco Guzmán, Shafiq Joty, Lluís Arquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In 52nd Annual Meeting of the Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210.

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.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. 2023. Belle: Be everyone's large language model engine.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. arXiv preprint arXiv:2301.08745, 1(10).

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1328–1338.

Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, and Fei Huang. 2025. Sdpo: Segment-level direct preference optimization for social agents. arXiv preprint arXiv:2501.01821.

Surafel Melaku Lakew, Mattia Antonino Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In Proceedings of the 16th International Workshop on Spoken Language Translation.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. 2019. Neural speech synthesis with transformer network. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 6706–6713.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. 2024. Controllable text generation for large language models: A survey. arXiv preprint arXiv:2408.12599.

James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In International conference on machine learning, pages 2285–2294. PMLR.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). IEEE access, 8:142642–142668.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. arXiv preprint arXiv:2405.14734.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 212–223.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. ACM Computing Surveys (CSUR), 54(6):1–37.

Zhijie Nie, Zhangchi Feng, Mingxin Li, Cunwang Zhang, Yanzhao Zhang, Dingkun Long, and Richong Zhang. 2024. When text embedding meets large language model: A comprehensive survey. arXiv preprint arXiv:2412.09165.

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.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. Advances in neural information processing systems, 32.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33:3008–3021.

Kaiser Sun and Mark Dredze. 2024. Amuro & char: Analyzing the relationship between pre-training and fine-tuning of large language models. arXiv preprint arXiv:2408.06663.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. In Proceedings of the 2019 Conference of the North. Association for Computational Linguistics.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. arXiv preprint arXiv:2201.08239.

Jrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. European Language Resources Association (ELRA).

A Vaswani. 2017. Attention is all you need. Advances in Neural Information Processing Systems.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: A fully end-to-end text-to-speech synthesis model. arXiv preprint arXiv:1703.10135, 164.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. 2024. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. arXiv preprint arXiv:2407.16216.

Yihan Wu, Junliang Guo, Xu Tan, Chen Zhang, Bohan Li, Ruihua Song, Lei He, Sheng Zhao, Arul Menezes, and Jiang Bian. 2023. Videodubber: Machine translation with speech-aware length control for video dubbing. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 13772–13779.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115.

Dong Yu and Lin Deng. 2016. Automatic speech recognition, volume 1. Springer.

Zhongyi Yu, Zhenghao Wu, Hao Zheng, Zhe XuanYuan, Jefferson Fong, and Weifeng Su. 2021. Lenatten: An effective length controlling unit for text summarization. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 363–370.

Ran Zhang, Wei Zhao, and Steffen Eger. 2024. How good are llms for literary translation, really? literary translation evaluation with humans and llms. arXiv preprint arXiv:2410.18697.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. 2023. Secrets of rlhf in large language models part i: Ppo. arXiv preprint arXiv:2307.04964.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 2765–2781.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.

A Experimental Details

In this section, we primarily describe the main experimental settings adopted in this study. Unless certain experiments require specific hyperparameters, we employ consistent hyperparameters across all experiments to maintain consistency and fairness in experimental comparisons.

A.1 Data Resources

In this study, we use the Polylingual Subtitle Corpus (PolySC) from 42 films and TV series (2021-2024) on the online video platform Youku. The corpus includes original Chinese subtitles and professionally translated English, Thai, and Spanish subtitles, used for zh⇒en, zh⇒th, zh⇒es, and es⇒zh subtitle translation. Chinese is a high-information-density language, Thai is medium-density, while English and Spanish are low-density languages. We set $n = 35$. Each translation direction's dataset contains approximately 26,000 prompt-response pairs for LLM training. For each direction, 97% of the data is used as the SFT Demonstration dataset, including both prompts and responses. The remaining 3% serves as the DA Query dataset, retaining only prompts. Statistics for the PolySC dataset are presented in Table 6.

PolySC dataset encompasses a diverse range of programs, including a total of 46 live-action television programs from 2021-2024 (42 for the training set and 4 for the test set). These programs span various genres such as fantasy, period drama, romance, and comedy, and include both long-form Series and mini series. The complete program list for PolySC dataset can be found at https://github.com/CcQunResearch/SSPO/blob/main/SSPOTraining/Playlist.md.

We present examples of Chinese and English subtitles from the PolySC dataset (in .ass file format) in Table 7. The "Start" and "End" columns identify the beginning and end times of the lines in the episode. The purpose of DA is to align the duration of the LLM's translation with that of the original line. It is important to note that in our experiments, we did not use the interval between the "Start" and "End" columns as the duration for the lines. This is because in most productions, human-translated subtitles set the start and end times of the translation to match those of the original text. We use edge-tts as a unified standard to measure the duration of both the original and translated lines.

A.2 Main Setting

We primarily use PyTorch[5] and Transformers[6] library to implement our methods, while leveraging DeepSpeed[7] for multi-GPU parallel training. Due to the limitations imposed by the GLM4 series models' left-padding feature for batch texts, we employ LLaMA-Factory[8] library for GLM4-related experiments. We re-implemented the AutoDubbing and VideoDubber methods, and utilize the paid APIs provided by OpenAI and Anthropic to obtain experimental results related to GPT-3.5-Turbo, GPT-4, and Claude 3.5 Sonnet. In our main experiments, ablation studies, and extended experiments, we strive to maintain consistent non-relevant hyperparameters to ensure fairness and consistency in comparisons. All critical hyperparameter settings are presented in Table 8.

A.3 Computational Resources

We conduct all experiments on 8 A800 80GB SXM GPUs. The time required for the three stages - SFT, sampling, and SSPO Training - is approximately 2h, 3h, and 1.5h, respectively. It's worth noting that due to the non-parallelizable nature of LLM inference, the sampling stage is more time-consuming compared to the training stage. Overall, executing a complete workflow on 8 A800 GPUs can be accomplished within an acceptable time frame.

B Extended Experiments

In this section, we will present additional evaluation experiments of SSPO across various aspects.

B.1 Performance on Bidirectional Translation

In this subsection, we validate the performance of SSPO on zh⇒es and es⇒zh translations.

B.1.1 Experiments

We conducted experiments on zh⇒es and es⇒zh translation tasks, with the results presented in Table 9. The findings demonstrate that SSPO consistently improves the duration consistency of the translated content. When translating from Chinese, a language with high information density, to Spanish, which has lower information density, SSPO reduces the proportion of translated lines exceeding the duration of the original lines. Conversely, when translating from Spanish to Chinese, SSPO

[5] https://github.com/pytorch/pytorch
[6] https://github.com/huggingface/transformers
[7] https://github.com/microsoft/DeepSpeed
[8] https://github.com/hiyouga/LLaMA-Factory

Statistic	zh⇒en	zh⇒th	zh⇔es
period		2021-2024	
# plays		42	
# lines		684625	
total duration (h)		471.4	
# avg source token		6.19 (zh)	
# avg target token	8.17	21.11	8.66 (es)
avg source duration (s)		1.096 (zh)	
avg target duration (s)	1.314	1.336	1.581 (es)

Table 6: Statistics of the datasets.

Start	End	Text
0:02:21.96	0:02:24.87	你去城北的铺子买一些安神的香料
0:02:25.32	0:02:27.39	大帅最近睡的不踏实
0:02:27.55	0:02:29.27	他可以焚香办公
0:02:29.27	0:02:31.48	这样也可以安心养神一些
0:02:32.80	0:02:33.83	好嘞少夫人
0:02:36.52	0:02:37.52	小雨
0:02:37.52	0:02:38.92	你去城南的花坊
0:02:38.92	0:02:40.67	购置些新鲜的鲜花回来
0:02:41.67	0:02:43.24	咱们这聂府啊
0:02:43.43	0:02:44.87	整日沉闷得很
0:02:22.00	0:02:24.91	Go to the shop in the north and buy some calming incense ingredients
0:02:25.36	0:02:27.43	for the Grand Marshal. He has had trouble sleeping well lately.
0:02:27.59	0:02:29.31	He can burn incense while working
0:02:29.31	0:02:31.52	to calm his nerves.
0:02:32.84	0:02:33.87	OK, Young Madam.
0:02:36.56	0:02:37.56	Xiaoyu,
0:02:37.56	0:02:38.96	go to the flower shop south of the city
0:02:38.96	0:02:40.71	and buy some fresh flowers.
0:02:41.71	0:02:43.28	Our Nie manor
0:02:43.47	0:02:44.91	is dreary every day

Table 7: Examples of Polylingual subtitles.

decreases the proportion of original lines exceeding the duration of the translated lines. Simultaneously, the average duration overrun is reduced. These experiments on Spanish translation further validate the universal effectiveness of SSPO in addressing duration inconsistencies arising from disparities in language information density.

B.1.2 Visualization

In Figure 6, we present the frequency distribution of the duration differences between the translated and original content for both the SFT and SSPO models of Qwen2.5-14B-Instruct. Similar to our previous experiments, it is evident that after SSPO training, the duration discrepancies between the original and translated content are significantly reduced. The histogram for the SSPO model is noticeably more concentrated around zero compared to that of the SFT model. This observation indicates that regardless of the information density disparities between the source and target languages, SSPO consistently improves duration consistency, resulting in a shift of the histogram towards zero.

B.2 Further Exploration

We explore two alternative solutions here.

B.2.1 Vanilla DPO Training

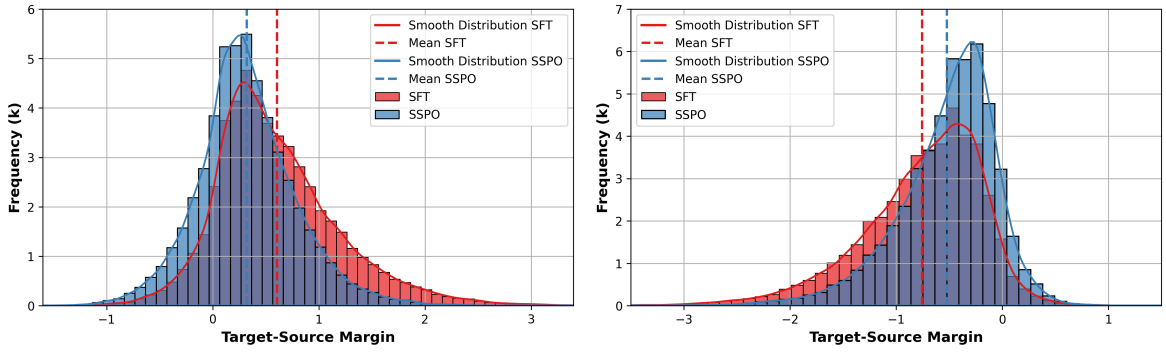
In considering DA as a preference optimization problem, we sought to investigate the question: "Can the vanilla DPO algorithm directly solve the DA problem?" Based on this inquiry, we conducted relevant designs and experiments. SSPO achieves good control over the duration of translated lines through a segment-wise sampling strategy and fine-grained DPO loss. We aimed to validate the impact

Type	Hyperparameter	Value	Remark
Sampling	ε_1	4	segment-level sampling indicator function threshold
	ε_2	0.08	
	k	20	sampling number
	temperature	1.4	sampling text generation hyperparameters
	top k	60	
	top p	0.95	
Training	**optimizer**	AdamW	-
	learning rate	4e-6	-
	epoch	4	-
	batch size	64	# lines
	total data size	1e4	# training lines
LoRA	**lora** r	16	-
	lora α	32	-
	lora targets	Q&K&V	-
TKLD	λ	1e-4	weight of the TKLD divergence constraint
SSPO	β	0.5	hyperparameter in SSPO loss

Table 8: Hyperparameter configuration.

Method	Train	zh⇒es						es⇒zh					
		S>T Rate	S>T Dur	T>S Rate	T>S Dur	CR	$\mathcal{P}$	S>T Rate	S>T Dur	T>S Rate	T>S Dur	CR	$\mathcal{P}$
Gold Reference	-	*6.2%*	*0.356*	*86.0%*	*0.801*	*7.8%*	*1.490*	*86.2%*	*0.808*	*6.2%*	*0.355*	*7.6%*	*0.590*
AutoDubbing	SFT	4.3%	**0.239**	91.3%	0.744	4.4%	1.304	92.1%	0.757	**1.9%**	0.237	6.0%	0.564
VideoDubber	SFT	5.3%	0.254	88.3%	0.732	6.4%	1.273	87.4%	0.731	5.7%	**0.223**	6.9%	0.535
GPT-3.5-Turbo	PE	**1.6%**	**0.223**	94.9%	0.937	3.5%	2.124	94.9%	0.941	**1.8%**	**0.222**	3.3%	0.720
GPT-4o	PE	3.0%	0.252	90.6%	0.779	6.3%	1.429	91.2%	0.791	2.8%	0.257	5.9%	0.589
Claude 3.5 Sonnet	PE	**2.3%**	0.242	92.3%	0.756	5.4%	1.290	92.6%	0.760	**1.9%**	0.246	5.5%	0.570
Llama3.1-8B-CN-Chat	SFT	5.8%	0.336	86.3%	0.793	7.9%	1.502	91.8%	0.826	2.6%	1.223	5.7%	0.620
	SSPO	8.2%	0.315	80.3%	0.614	11.5%	0.858	**82.2%**	**0.640**	6.9%	0.483	**10.9%**	**0.450**
GLM-4-9B-Chat	SFT	5.9%	0.346	85.8%	0.785	8.3%	1.492	91.9%	0.834	2.5%	0.601	5.6%	0.624
	SSPO	15.9%	0.362	**70.2%**	**0.548**	**13.9%**	**0.665**	**81.5%**	0.654	7.5%	0.542	**11.0%**	0.460
Qwen2.5-14B-Instruct	SFT	7.0%	0.335	84.0%	0.758	8.9%	1.320	91.8%	0.830	2.6%	0.297	5.6%	0.622
	SSPO	15.8%	0.359	**69.6%**	**0.538**	**14.6%**	**0.644**	84.8%	**0.629**	4.8%	0.308	10.3%	**0.447**
Alignment Bound	-	*4.3%*	*0.291*	*84.3%*	*0.566*	*11.3%*	*0.654*	*86.2%*	*0.660*	*1.4%*	*0.245*	*12.4%*	*0.466*

Table 9: zh⇒es and es⇒zh results on test set. The best and second best results are denoted as **blue** and **orange**.



(a) zh⇒es (b) es⇒zh

Figure 6: Frequency distribution of Qwen2.5-14B-Instruct model on zh⇒es and es⇒zh translations.

of these two components. We employed the standard DPO training process (Rafailov et al., 2024) to perform DA on the SFT model. Specifically, we utilized either coarse-grained or fine-grained sampling to sample a chosen response $y^{(c)}$ and a rejected response $y^{(r)}$ for each sample $x \in \mathcal{D}_{\text{query}}$ in the Query dataset. Subsequently, we optimized the policy model using the standard DPO loss (Rafailov et al., 2024).

The adopted coarse-grained and fine-grained

sampling procedures are illustrated in Algorithm 2 and Algorithm 3, respectively. In coarse-grained sampling, k complete responses are directly sampled for a given prompt x. The consistency penalty $\mathcal{P}(s_i, t_i)$ is calculated for each of the n lines in each response, and the sum of these penalties is computed. The response with the minimum sum is selected as the chosen response $y^{(c)}$, while the one with the maximum sum becomes the rejected response $y^{(r)}$. Fine-grained sampling, on the other hand, requires two segment-wise sampling cycles similar to those in Algorithm 1 for a single prompt x. In these two cycles, the lines with the minimum and maximum $\mathcal{P}$ are used as the prefix for sampling the next line, respectively. This process ultimately yields the segment-wise sampled chosen response $y^{(c)}$ and rejected response $y^{(r)}$.

B.2.2 Advantage-based PPO Training

In the literature, RLHF (primarily based on Proximal Policy Optimization (PPO)) is often considered to outperform DPO(Wang et al., 2024; Gao et al., 2024) despite its complexity. Thus, we want to explore the question, "Can PPO techniques achieve better results than DPO-based SSPO in DA?" Based on this hypothesis, we conducted related experiments. We employed an advantage-based PPO training process(Zheng et al., 2023). Specifically, we implement the PPO training process through the following steps:

1. *Rollout* - For each sample $x \in \mathcal{D}_{\text{query}}$ in query dataset, sample a trajectory τ using Algorithm 4.

2. *Compute Rewards and Advantages* - Perform Generalized Advantage Estimation (GAE) on τ using the value network V_ϕ (with Qwen2.5-7B-Instruct (Yang et al., 2024) as the backbone):

$$\delta_i = -\mathcal{P}(s_i, t_i) + \gamma V_\phi(p_{i+1}) - V_\phi(p_i), \quad (5)$$

$$A_i = \sum_{l=0}^{n-i-1} (\gamma\lambda)^l \delta_{i+l}. \quad (6)$$

3. *Update Policy Network* - Let π_{old} be the fixed old policy during sampling, and minimize the loss function (the KL divergence constraint between π_θ and π_{old} is omitted here):

$$\mathcal{L}_{\text{clip}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}_{\text{query}}}\left[\sum_{i=1}^{n} \min\left(\frac{\pi_\theta(t_i \mid p_i)}{\pi_{\text{old}}(t_i \mid p_i)} A_i, \text{clip}\left(\frac{\pi_\theta(t_i \mid p_i)}{\pi_{\text{old}}(t_i \mid p_i)}, 1-\epsilon, 1+\epsilon\right) A_i\right)\right]. \quad (7)$$

4. *Update Value Network* - Fit $V_\phi(p_i)$ to the GAE-based estimated value $\hat{V}_i = A_i + V_\phi(p_i)$ using mean squared error:

$$\mathcal{L}_V(\phi) = \mathbb{E}_{x\sim\mathcal{D}_{\text{query}}}\left[\sum_{i=1}^{n} (V_\phi(p_i) - \hat{V}_i)^2\right]. \quad (8)$$

5. *Iterate Multiple Rounds* - Continuously collect new data and update the value network and policy model until convergence.

B.2.3 Evaluation

We utilized the GLM-4-9B-Chat and Qwen2.5-14B-Instruct backbone models to validate the DA performance in zh⇒en and zh⇒th translations under the "Vanilla DPO Training" and "Advantage-based PPO Training" configurations. The experimental results are presented in Table 10. The results indicate that while PPO training achieved a performance improvement over the SFT model, none of the other configurations showed significant enhancement, and their performance is notably different from that of SSPO. With vanilla DPO training, neither coarse-grained nor fine-grained sampling improved duration consistency, and the fine-grained sampling it employed consumed twice the time and computational resources compared to SSPO. This validates that the effectiveness of SSPO arises from the dual factors of sentence-level sampling strategy and fine-grained DPO loss. This means that each sentence in the prompt must be sampled and optimized independently, as sampling and loss calculation on the complete response are ineffective. Additionally, although RLHF methods often outperform DPO methods in LLM preference alignment tasks, using PPO methods in DA tasks resulted in performance lower than SSPO. This is because, for DA tasks, duration consistency has a clear metric (i.e., $\mathcal{P}$), and thus SSPO optimizes towards the optimal solution by increasing the generation probability of the most consistent translations for each sentence and reducing the generation probability of the least consistent translations, while

Algorithm 2 Coarse-grained Sampling for Vanilla DPO.

Input: SFT model π_{sft}, query dataset $\mathcal{D}_{\text{query}}$, sampling number k.
Output: sampled response pairs set $\mathcal{S}(x)$.

for any $x \in \mathcal{D}_{\text{query}}$ **do**
  // Sample multiple candidate responses.
  for $i = 1$ to k **do**
    Sample $\pi_{\text{sft}}(y|x)$.
  end for
  Measure the sum of $\mathcal{P}$ for each line of y^i in the candidate set $\{y^i|i = 1, 2, \ldots, k\}$.
  Select chosen $y^{(\text{c})}$ and rejected $y^{(\text{r})}$.
end for
return $\mathcal{S}(x) \equiv \{y^{(\text{c})}, y^{(\text{r})}\}$.

Algorithm 3 Fine-grained Sampling for Vanilla DPO.

Input: SFT model π_{sft}, query dataset $\mathcal{D}_{\text{query}}$, sampling number k.
Output: sampled response pairs set $\mathcal{S}(x)$.

for any $x \in \mathcal{D}_{\text{query}}$ **do**
  // The first sampling cycle used to obtain $y^{(\text{c})}$.
  for $i = 1$ to n **do**
    for $j = 1$ to k **do**
      Sample $\pi_{\text{sft}}(t_i^j|x, s_1, t_1^{(\text{c})}, \ldots, s_{i-1}, t_{i-1}^{(\text{c})}, s_i)$.
    end for
    Deduplicate and measure $\{t_i^j|j = 1, 2, \ldots, k\}$ by $\mathcal{P}(s_i, t_i^j)$, and select chosen $t_i^{(\text{c})}$.
  end for
  Concatenate $\{(s_i, t_i^{(\text{c})})|i = 1, 2, \ldots, n\}$ yields $y^{(\text{c})}$.
  // The second sampling cycle used to obtain $y^{(\text{r})}$.
  for $i = 1$ to n **do**
    for $j = 1$ to k **do**
      Sample $\pi_{\text{sft}}(t_i^j|x, s_1, t_1^{(\text{r})}, \ldots, s_{i-1}, t_{i-1}^{(\text{r})}, s_i)$.
    end for
    Deduplicate and measure $\{t_i^j|j = 1, 2, \ldots, k\}$ by $\mathcal{P}(s_i, t_i^j)$, and select rejected $t_i^{(\text{r})}$.
  end for
  Concatenate $\{(s_i, t_i^{(\text{r})})|i = 1, 2, \ldots, n\}$ yields $y^{(\text{r})}$.
end for
return $\mathcal{S}(x) \equiv \{y^{(\text{c})}, y^{(\text{r})}\}$.

Algorithm 4 Sampling for PPO solution.

Input: SFT model π_{sft}, DA dataset $\mathcal{D}_{\text{query}}$.
Output: sampled trajectory set $\mathcal{T} \equiv \{\tau\}$.

for any $x \in \mathcal{D}_{\text{query}}$ **do**
  // Iterate through the dialogue lines in x.
  for $i = 1$ to n **do**
    Sample $\pi_{\text{sft}}(t_i|x, s_1, t_1, \ldots, s_{i-1}, t_{i-1}, s_i)$ and measure t_i by $\mathcal{P}(s_i, t_i)$.
  end for
  Obtain trajectory $\tau = \{(p_i, t_i, \mathcal{P}(s_i, t_i))|i = 1, 2, \cdots, n\}$.
end for
return $\mathcal{T} \equiv \{\tau\}$.

Method	Training	Sampling	zh⇒en						zh⇒th					
			S>T Rate	S>T Dur	T>S Rate	T>S Dur	CR	$\mathcal{P}$	S>T Rate	S>T Dur	T>S Rate	T>S Dur	CR	$\mathcal{P}$
Gold Reference	-	-	18.0%	0.344	64.1%	0.464	17.9%	0.501	19.4%	0.369	60.2%	0.460	20.3%	0.489
GLM-4-9B-Chat	SFT	-	**19.5%**	**0.342**	60.5%	0.427	20.0%	0.428	**18.1%**	**0.291**	55.0%	0.391	27.0%	0.360
	DPO	C	**19.4%**	**0.340**	61.1%	0.432	19.4%	0.432	18.7%	0.293	54.4%	0.391	26.9%	0.358
		F	**19.5%**	0.343	60.7%	0.424	19.8%	0.431	**18.4%**	0.289	55.1%	0.392	26.6%	0.360
	PPO	F	24.5%	**0.342**	**53.4%**	**0.364**	**22.1%**	**0.323**	23.5%	0.290	**47.5%**	**0.318**	**29.0%**	**0.301**
	SSPO	F	29.6%	0.350	**45.9%**	**0.323**	**24.5%**	**0.283**	25.9%	**0.291**	**42.0%**	**0.318**	**32.1%**	**0.254**
Qwen2.5-14B-Instruct	SFT	-	**20.0%**	**0.341**	59.8%	0.439	20.2%	0.423	18.2%	0.294	55.4%	0.397	26.4%	0.362
	DPO	C	**25.4%**	0.376	54.0%	0.508	20.6%	0.408	**18.1%**	0.297	55.9%	0.411	25.9%	0.369
		F	25.6%	0.378	54.2%	0.492	20.2%	0.408	**17.7%**	0.295	56.4%	0.411	26.0%	0.373
	PPO	F	30.2%	0.372	**47.4%**	**0.366**	**22.4%**	**0.315**	30.8%	**0.293**	**38.3%**	**0.314**	**30.9%**	**0.237**
	SSPO	F	34.4%	**0.366**	**40.6%**	**0.324**	**24.9%**	**0.272**	38.6%	**0.290**	**25.3%**	**0.279**	**36.1%**	**0.198**
Alignment Bound	-	-	*16.4%*	*0.278*	*39.3%*	*0.331*	*44.3%*	*0.220*	*9.2%*	*0.232*	*40.4%*	*0.313*	*50.4%*	*0.203*

Table 10: Experimental evaluation results of alternative solutions. C for Coarse-grained, while F for Fine-grained. The best and second best results are denoted as **blue** and **orange**.

PPO training follows a gradual optimization process. This contrasts with the situation in preference alignment (Wang et al., 2024; Gao et al., 2024).

C Theory: Segment Supervised Preference Optimization

In this section, we formalize the localized multi-segment preference optimization problem and validate the effectiveness of SSPO, while also highlighting the limitations of general preference optimization methods. Note that the notation used in this section may have slightly different meanings from those in previous sections.

C.1 Localized Multi-Segment Preference Optimization Problem

For a language model π_θ, given an input $x \in \mathcal{X}$ (where $\mathcal{X}$ is the input space), assume it consists of n interrelated segments, i.e., $x = (x_1, x_2, \ldots, x_n)$. Correspondingly, the output $y \in \mathcal{Y}$ (where $\mathcal{Y}$ is the output space) also comprises n interrelated segments, i.e., $y = (y_1, y_2, \ldots, y_n)$, with each x_i corresponding to y_i. Additionally, the generation of y_i is influenced by $y_1, y_2, \ldots, y_{i-1}$ (note that this influence may not only stem from the autoregressive property of the language model but also from semantic dependencies among output segments), expressed as:

$$\pi_\theta(y \mid x) = \prod_{i=1}^{n} \pi_\theta(y_i \mid x, y_1, \cdots, y_{i-1}). \quad (9)$$

The general preference optimization task involves an outcome-supervised reward function $r(x, y)$ for the output y given input x. In contrast, the localized multi-segment preference optimization problem employs segment-supervised preference metrics, denoted as $r(x_i, y_i)$, which quantify the alignment of segment output y_i with the predefined optimization preference for its corresponding segment input x_i. Our objective is to adjust model parameters θ during training such that the policy π_θ is optimized to maximize $r(x_i, y_i)$ across all segments.

C.2 Ineffectiveness of General Preference Optimization Methods

Let's take DPO (Rafailov et al., 2024) as an example to illustrate the limitations of general preference optimization methods when dealing with localized multi-segment preference optimization problems. In conducting localized preference optimization, DPO first labels the complete preferred response $y^w = (y_1^w, y_2^w, \ldots, y_n^w)$ and the less preferred response $y^l = (y_1^l, y_2^l, \ldots, y_n^l)$ corresponding to $x = (x_1, x_2, \ldots, x_n)$ using $r(x_i, y_i)$. y^w and y^l necessarily satisfy:

$$\pi_\theta(y^w \mid x) = \prod_{i=1}^{n} \pi_\theta(y_i^w \mid x, y_1^w, \cdots, y_{i-1}^w),$$
$$\pi_\theta(y^l \mid x) = \prod_{i=1}^{n} \pi_\theta(y_i^l \mid x, y_1^l, \cdots, y_{i-1}^l). \quad (10)$$

Then, DPO optimizes π_θ by increasing the log probability of the preferred response relative to the less preferred response. For the contrastive term in DPO loss, we calculate:

$$\log \frac{\pi_\theta(y^w \mid x)}{\pi_{\text{ref}}(y^w \mid x)} - \log \frac{\pi_\theta(y^l \mid x)}{\pi_{\text{ref}}(y^l \mid x)} = \sum_{i=1}^{n} \left[\log \frac{\pi_\theta(y_i^w \mid x, y_{1:i-1}^w)}{\pi_{\text{ref}}(y_i^w \mid x, y_{1:i-1}^w)} - \log \frac{\pi_\theta(y_i^l \mid x, y_{1:i-1}^l)}{\pi_{\text{ref}}(y_i^l \mid x, y_{1:i-1}^l)} \right], \quad (11)$$

where $y_{1:i-1}^w = (y_1^w, \ldots, y_{i-1}^w)$ and $y_{1:i-1}^l = (y_1^l, \ldots, y_{i-1}^l)$.

It can be observed that the generation probability of the i-th segment is not conditioned on the same prefix, but rather on its own preferred or dispreferred prefix (i.e., $y_{1:i-1}^w$ or $y_{1:i-1}^l$). As a

result, the comparison of the i-th segment is not a fair "apples-to-apples" comparison. DPO only performs a single holistic preference judgment on the complete sequences (y^w, y^l), leaving the model unaware of how to adjust each individual segment. In other words, while the model knows that the full sequence y^w is superior to y^l, it lacks guidance on how to make locally optimal choices for the i-th segment. Thus, as demonstrated by the experimental results in Appendix B.2, vanilla DPO fails to effectively address the task of localized multi-segment preference optimization.

C.3 Segment Supervised Preference Optimization

Unlike DPO, SSPO labels the preferred response y_i^w and dispreferred response y_i^l for the next segment individually, using $y_{1:i-1}^w$ as the prefix, instead of obtaining the entire response sequence. Specifically, for the i-th segment, the prefix is the fixed prefix $p_i = (x, y_1^w, \cdots, y_{i-1}^w)$. This ensures that each segment is compared on the same and preferred prefix, eliminating unfair competition between 'preferred prefix vs. dispreferred prefix'. Under the same prefix, preference alignment loss (such as that used in DPO) is applied to compare $\pi_\theta(y_i^w \mid p_i)$ and $\pi_\theta(y_i^l \mid p_i)$ conditioned on the same p_i. Ultimately, the cumulative loss of all segments enforces preference constraints on each segment:

$$\mathcal{L}_{\text{SSPO}}(\pi_\theta; \pi_{\text{ref}}) = - \mathbb{E}_{(x, y_{1:n}^w, y_{1:n}^l) \sim \mathcal{D}} \left[\sum_{i=1}^{n} \mathcal{L}_{\text{DPO}}(y_i^w, y_i^l, p_i) \right]. \quad (12)$$

D Discussion

In this section, we will present further discussions on the SSPO method.

D.1 Recommendations

We offer the following development suggestions for technicians using SSPO:

- Generally, larger models will consume more time and computational resources during the sampling phase, necessitating a trade-off between performance and cost.
- Due to the varying token encoding densities for different languages in LLM, a larger number of sampling tokens should be set for Thai (e.g., 80).
- For languages with similar information density (e.g., English and German), where duration consistency is not critical or the TTS stage is not required, DA may be omitted.
- When applying SSPO to other localized preference optimization tasks, it is essential to determine the task's preference metrics and criteria for optimization exemptions.

D.2 Future Research

Translation quality evaluation experiments indicate that LLMs' translations are inferior to human translations in terms of vividness. The subtitle texts of visual media programs are typically deeply integrated with their associated video and audio. Compared to the translation of legal and religious texts, subtitle translation may not require strict accuracy; instead, it should focus more on the vividness of the translation. In future research, we aim to apply SSPO to improve subtitle translation quality, exploring methods and techniques to enhance the vividness of translations.

E Prompt and Instructions

In this section, we present the input and output formats of the LLMs we employed, as well as the evaluation instructions used for manual assessment of translation quality.

E.1 Input and Output of LLM

We present the prompt and response formats for the zh⇒en SFT translation model in Table 11 and Table 12 ((similar for other languages)). We process the original subtitle text and its translation of the television programs into this format for training the SFT model. The prompt is structured as follows:

1. *Preamble* - An introduction and instructions describing the task at hand
2. *Terminology* - A translation glossary for terminology in the dialogue
3. *Lines to Translate* - Multiple lines of dialogue requiring translation
4. *Ending* - Ending text to prompt the LLM (e.g., "Translation results:")

Maintaining consistency in terminology translation is crucial for subtitle translation, necessitating the specific designation of terminology translations. Our general process for obtaining the terminology

translation glossary used in the prompt is as follows: 1) Utilize an off-the-shelf LLM to identify and filter terminology and its translations from all dialogue; 2) Employ the identification results and an off-the-shelf LLM to train a terminology identification model, and use this model to identify and translate terminology in the test set (without ground-truth translations); 3) Retrieve the terminology appearing in the current SFT prompt's dialogue from the identification results. As terminology identification is not the primary focus of this study, we will not elaborate further on this aspect.

E.2 Prompt Engineering Template

We showcase the prompt used for the GPT-3.5, GPT-4o, and Claude 3.5 Sonnet models for the zh⇒en translation task (similar for other languages) in Table 13. This prompt differ from the input to our SFT model by the inclusion of an additional translation example for LLM context learning. The prompt is structured as follows:

1. *Preamble* - An introduction and instructions describing the task at hand
2. *Example* - A few-shot example for context learning
3. *Task* - Terminology information and the original dialogue to be translated, provided to the model
4. *Ending* - Ending text to prompt the LLM

E.3 Instruction and Prompt for Quality Evaluation

When conducting human evaluations of translation quality, it is crucial to provide evaluators with instructions that specify the evaluation perspectives, criteria, and format. This will directly influence the focus and emphasis of evaluators during the quality assessment. The provided instructions in Table 14 are structured as follows:

1. *Criteria* - Describe the professional standards that need to be followed for the translation evaluation task at hand.
2. *Task* - Outline the content provided to the evaluators and the output format that must be adhered to.

Preamble	Please translate a series of Chinese movie/TV subtitles into English according to the following requirements: 1. The translation should be colloquial, easy to understand, and consistent with the language style of the Chinese subtitles. 2. Ensure that the length of the translated English subtitles matches that of the original Chinese subtitles. 3. Proper nouns should be accurately translated according to the specified translations. 4. When outputting, ensure that the number of translated lines matches the original text, avoid merging subtitles, and both the original text and translation must be outputted together.
Terminology	Translation of proper nouns: 仙君 - Celestial Deity 帝尊 - Your Supreme Majesty 应渊君 - Sovereign Lord Yingyuan ... 玄夜 - Xuanye 永夜功 - Eternal Darkness 修罗尊主 - Asura King
Lines to Translate	According to the requirements previously stated, complete the following subtitle translation. Original text: 这两页为什么是空白的 有两位仙君从未更换过腰带 所以就没有记录 一位是帝尊 还有一位就是应渊君 ... 自你出现以后 从未更换过 应渊白日所施 竟是玄夜所创的永夜功 此功唯有修罗尊主的血脉可练
Ending	Translation results:

Table 11: zh⇒en translation prompt demonstration of SFT model.

Response	这两页为什么是空白的(Why are these two pages blank?) 有两位仙君从未更换过腰带(Two Celestial Deities never changed their belts.) 所以就没有记录(So, there is no record.) 一位是帝尊(His Supreme Majesty) 还有一位就是应渊君(and Sovereign Lord Yingyuan.) ... 自你出现以后(Since you appeared,) 从未更换过(he had never replaced it.) 应渊白日所施(What Yingyuan executed in the daytime) 竟是玄夜所创的永夜功(was the Eternal Darkness created by Xuanye.) 此功唯有修罗尊主的血脉可练(This skill was only practiced by the bloodline of the Lord of Asura Clan.)

Table 12: zh⇒en translation response demonstration of SFT model.

Preamble	[Requirements] Please translate multiple lines of Chinese subtitles into English, adhering to the following guidelines: 1. The translation should be colloquial and easily understood, maintaining consistency with the language style of the Chinese subtitles. 2. Proper nouns should be translated according to the specified translations provided. 3. Output the original text and translation together in the format of "Chinese original (English translation)". Ensure that subtitles are not merged, and the number of lines in the translated output matches that of the Chinese original. 4. Critical requirement: The reading duration of each translated line should be consistent with the Chinese original. Ensure that the duration of the translated text is neither longer nor shorter than the original.
Example	[Example] Proper noun translations: 萤灯 - Yingdeng 帝君 - Your Majesty ... 颜淡 - Yandan 妙法阁 - Magical Pavilion Original text: 萤灯姐姐 姐姐升官大喜 ... 甚是铺张 本君消受不起 Translation results: 萤灯姐姐(Sister Yingdeng,) 姐姐升官大喜(Congrats on your promotion,) ... 甚是铺张(It's quite lavish,) 本君消受不起(I can't accept such extravagance.)
Task	[Task] Now, following the requirements mentioned above and referring to the examples provided, translate the following Chinese dialogue into English. Proper noun translations: 有限合伙人 - limited partner 智慧社区 - smart community ... 宁檬 - Ning Meng 南林股份 - Nanlin Securities Original text: 但是他们对于 回报的要求也非常地高 ... 等了这么久 终于露出马脚了
Ending	Please directly output the translation result, ensuring to follow the format of "Chinese original (English translation)". Do not output any additional text.

Table 13: zh⇒en translation prompt for OpenAI and Anthropic models.

Criteria

[Evaluation Criteria]

1. Accuracy
When assessing the accuracy of translated audiovisual dialogues, take into account the following aspects:

- **Semantic Fidelity**: Check if the original dialogue's meaning is faithfully represented in the translation and if the semantic content of the source is clearly communicated in the target language.
- **Grammatical Precision**: Evaluate the grammatical correctness of the translation, including sentence structure, verb tense, voice, and other grammatical elements.
- **Terminology Translation**: Ensure that proper nouns and specialized terms are accurately translated, preserving the original terms' semantics and context.

2. Naturalness
When assessing the naturalness of translated audiovisual dialogues, consider the following dimensions:

- **Coherence**: Determine if the translation reads naturally as if authored by a native speaker of the target language, and check the logical connections between sentences.
- **Readability**: Consider whether the translation is easy to read and comprehend, and if the word choice and expressions adhere to the target language conventions.
- **Fluency**: Assess whether the translation flows smoothly, has well-constructed sentences, and is free of glaring grammatical mistakes or awkward expressions.

3. Vividness
When assessing the vividness of translated audiovisual dialogues, review the following dimensions:

- **Stylistic Consistency**: Verify that the translation preserves the style and character traits of the original, including tonal consistency and emotional subtleties.
- **Expressiveness**: Determine if the translation captures the original dialogue's spirit and atmosphere, avoiding stiff literal translations to engage and resonate with the audience.
- **Emotion**: Check if the translation accurately reflects the characters' emotions, aligns with scene and character contexts, and emotionally connects with the target language audience.

Task

[Task]

For each set of original dialogues, you are provided with two distinct translations (A and B). Utilize the multiple evaluation dimensions outlined in the [Evaluation Criteria] to assess the two translations for every set of original dialogues. Record your assessment results for translations A and B by marking [A is better], [B is better], or [No significant difference between A and B]. Note that your evaluation should focus on the overall quality of each set of dialogues as a whole, rather than on individual lines.

Table 14: Guidelines for human assessment of translation quality.