AlignCap: Aligning Speech Emotion Captioning to Human Preferences

Ziqi Liang[1*], **Haoxiang Shi**[1*], **Hanhui Chen**[2]
[1] University of Science and Technology of China, Hefei, China
[2] Southern University of Science and Technology, Shenzhen, China
{tymlzq, shihaoxiang}@mail.ustc.edu.cn, hanhuichencn@gmail.com

Abstract

Speech Emotion Captioning (SEC) has gradually become an active research task. The emotional content conveyed through human speech are often complex, and classifying them into fixed categories may not be enough to fully capture speech emotions. Describing speech emotions through natural language may be a more effective approach. However, existing SEC methods often produce hallucinations and lose generalization on unseen speech. To overcome these problems, we propose **AlignCap**, which Aligning Speech Emotion Captioning to Human Preferences based on large language model (LLM) with two properties: **1) Speech-Text Alignment**, which minimizing the divergence between the LLM's response prediction distributions for speech and text inputs using knowledge distillation (KD) Regularization. **2) Human Preference Alignment**, where we design Preference Optimization (PO) Regularization to eliminate factuality and faithfulness hallucinations. We also extract emotional clues as a prompt for enriching fine-grained information under KD-Regularization. Experiments demonstrate that AlignCap presents stronger performance to other state-of-the-art methods on Zero-shot SEC task.

1 Introduction

The identification and description of speech emotions play a crucial role in improving communication efficiency. It also aids in understanding the speaker's intentions. Previous work usually treats emotion acquisition as a classification task, such as Speech Emotion Recognition (SER) (Ye et al., 2023; Chen et al., 2023; Shi et al., 2024), which assigns speech to different emotion categories based on the emotions such as sadness, anger, and happiness contained within the speech. However, there may be a mixture of emotions within one utterance, and classifying speech into a single emotion

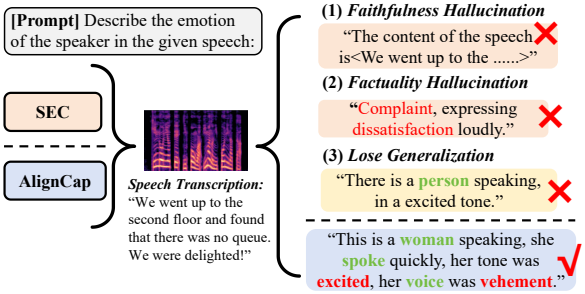


Figure 1: Hallucination and lack of generalization.

category is not enough to capture the true emotion. Moreover, different annotators may assign different emotion category labels to a piece of speech, leading to the label ambiguity problem in SER task (Li et al., 2017; Lian et al., 2023b). This can result in inaccurate emotion labels in existing SER datasets.

Given the limitations of speech emotion classification, employing natural language descriptions instead of emotion category labels is a more accurately approach. SECap (Xu et al., 2024) first proposes a speech emotion captioning framework to describe speech emotions using natural language effectively. It utilizes HuBERT (Hsu et al., 2021) as an audio encoder to extract speech features while leveraging mutual information learning to decouple content and emotion-related features. (Deshmukh et al., 2024) employs GPT-2 (Radford et al., 2019) as the decoder to generate captions based on the pre-trained CLAP (Wu et al., 2023) audio encoder. (Salewski et al., 2023) exploits OPT (Zhang et al., 2022) as the LLM to produce captions that describe the audio content. However, facing with unseen speech, these methods tend to produce hallucinations of factuality and faithfulness, resulting in false emotional descriptions and responses that are inconsistent with user instructions. In addition, the paradigm of text-only training and zero-shot inference on speech like (Kouzelis and Katsouros, 2023) will lead to training-inference mismatch, resulting in poor model generalization.

In this paper, we propose a novel SEC frame-

* Equal contribution.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3837–3846
November 12-16, 2024

work **AlignCap**, which aims to generate rich and coherent captions while maintaining high consistency with speech emotion. We design KD-Regularization, which can minimize the distribution gap between LLM's response to speech input and those to corresponding text inputs. It bridges the training-inference mismatch in zero-shot SEC and model generalization is improved. AlignCap is the first to align SEC models to human preferences via PO-Regularization, which eliminates factuality and faithfulness hallucinations of SEC models on unseen speech. We also utilize a acoustic prompt generated from emotional clues to enrich fine-grained information. To summarize, our contributions are as follows:

1. We propose KD-Regularization to achieve Speech-Text Alignment and use the KL-divergence of next-token prediction distributions as a measure of alignment.

2. We propose PO-Regularization to achieve Human Preference Alignment, which generates rich, consistency and rationality emotion descriptions.

3. We analyze the issue of distribution gaps in SEC task and explore various alignment methods to bridge the gap. Experiments demonstrate AlignCap's superiority in both zero-shot and cross-domain scenarios.

2 Background and Discussion

The section describes the speech-text distribution gap of traditional SEC methods. To explore this gap, we conducted preliminary experiments to analyze its potential impact on train-inference mismatch and performance degradation. Furthermore, we discuss the impact of modal alignment position on downstream SEC performance.

Distribution Gap and Alignment. As the creation of speech-caption pairs is costly, traditional SEC methods usually are trained using only text, and employed zero-shot inference on speech. The distribution of speech and text embeddings do not exactly coincide, which degrades the SEC's performance. To analyze the effect of modal alignment on eliminating distribution gap, we adopt No-alignment, Contrastive Learning alignment (CL-Align) (Deshmukh et al., 2024), and Projection-based alignment (CL+Proj-Align) (Deshmukh et al., 2024), and evaluate the performance of SEC on BLEU@4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) metrics. We conducted three experiments: **1) No-Align:** Speech encoder of Pretrained CLAP model (Wu et al., 2023) is used to zero-shot inference directly. **2) CL-Align:** We fine-tuning the text encoder and speech encoder of CLAP using contrastive learning on speech-caption pairs $\mathcal{D}_s = \{(x_n, y_n)\}$ which are randomly selected from SEC datasets. **3) CL+Proj-Align:** Based on CL-Align, we add Projection-based decoding to project the speech embedding into the text embedding space through cosine similarity. As shown in Fig 2, captions generated from model with Alignment exhibit superior similarity compared to that No-alignment model. This findings proves that the distribution gap adversely affects the SEC's performance.

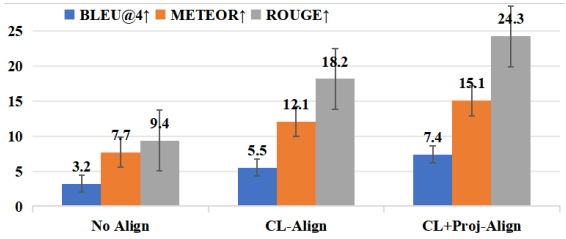
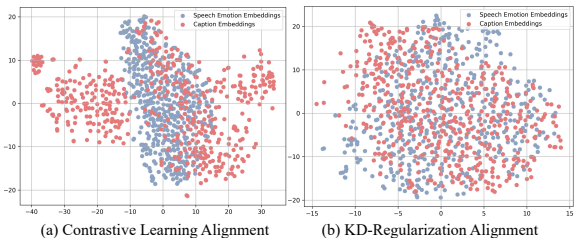


Figure 2: Results of different alignment methods.

Align before or after LLM Decoding? According to (Jiang et al., 2023), complete alignment between modalities is often not the optimal solution for downstream tasks. Such alignment may result in information loss, especially when the information provided by the two modalities differs significantly. Traditional SEC models achieve Speech-Text Alignment via fine-tuning encoder on speech-caption pairs, which bridges the distribution gap before LLM decoding. However, complete alignment of speech and text embedding may result in information loss, and it lacks a direct measure for assessing speech-text alignment quality.

To address these problems, we propose KD-Regularization which achieve Speech-Text Align-



(a) Contrastive Learning Alignment (b) KD-Regularization Alignment

Figure 3: T-SNE visualizations of LLM's output from speech and text input. (a) Align before LLM Decoding. (b) Align after LLM Decoding.

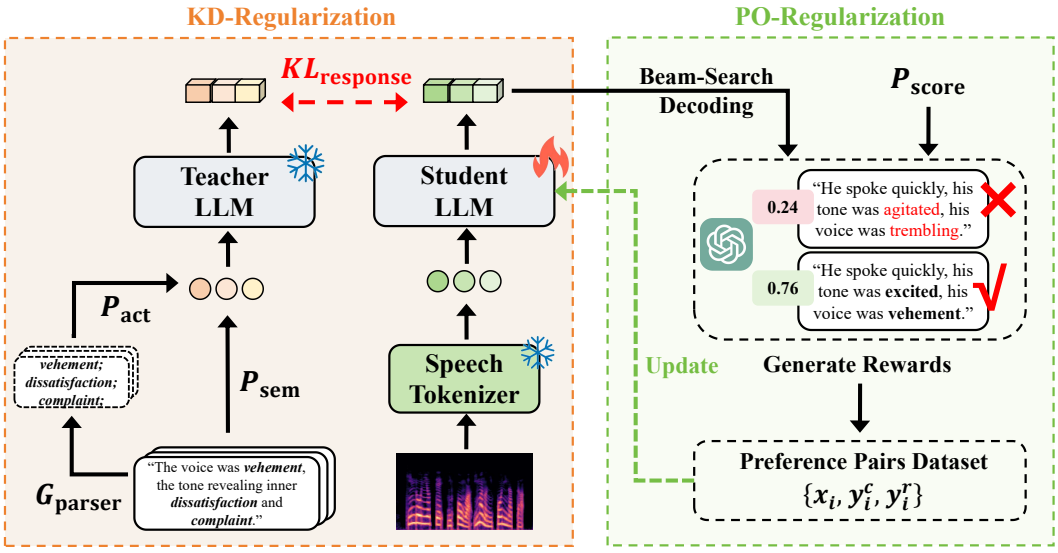


Figure 4: The framework of AlignCap. **Left:** Illustration of Knowledge Distillation Regularization. Acoustic prompt $\mathbf{P}_{\text{act}}$ is generated from emotional clues, which is extracted by an emotion grammar parser $\mathbf{G}_{\text{parser}}$. Semantic prompt $\mathbf{P}_{\text{sem}}$ is generated from LLM tokenizer. **Right:** Illustration of Preference Optimization Regularization.

ment and bridge the distribution gap after LLM decoding. It use the KL-divergence of next-token prediction distributions between LLM's response as a measure of Speech-Text Alignment. As shown in Fig 3, we observe that align after LLM decoding using knowledge distillation can more effectively improve the speech-text alignment performance.

3 AlignCap

3.1 KD-Regularization

Our goal is to generate speech emotion captions for speech clips, we design a student LLM to implement speech tokens to text generation and employ a teacher LLM's response to guide student LLM's next-token generation. LLaMA-7B (Touvron et al., 2023a) is chosen to implement this decoding process due to its exceptional language understanding and modeling capabilities. We simply choose rank value of 8 for LoRA fne-tuning (Hu et al., 2022b) conducted on Student-LLaMA, while the Teacher-LLaMA parameters are frozen.

Acoustic Prompt. We first construct a vocabulary of emotional clues, adjectives such as the speaker's tone, intonation, pitch, rhythm, and volume in captions are all regraded as emotional clues. We design an emotion grammar parser (based on NLTK toolkit) to recognize these clues, which are filtered by the vocabulary. Then these clues are inserted into a prompt template $\mathrm{P_T}$: <Feeling $e_1, e_2, ...,$ and e_n>, where e_n is the n_{th} emotion entity. The acoustic prompt can capture rich and delicate emotion information in emotional clues. It can enrich fine-grained emotional description and enhance the robustness of zero-shot captioning for unseen speech, leveraging its training-agnostic nature, which is denoted as:

$$\begin{aligned} e_{1\sim n} &= G_{\text{Parser}}(y_i = \{c_i^1, ..., c_i^{|y_i|}\}) \\ \mathrm{P_{act}} &= \text{Insert}(\mathrm{P_T}, idx, e_{1\sim n}) \end{aligned} \tag{1}$$

Where y_i is a series of captions, c_i^m is the m_{th} caption of y_i. G_{Parser} and $\mathrm{P_T}$ represent emotion grammar parser and prompt template respectively. We insert the emotional clues $e_{1\sim n}$ into the index position idx of $\mathrm{P_T}$ to get acoustic prompt $\mathrm{P_{act}}$.

Text Token Generation. We denote the captions in speech-caption pairs as the semantic prompt $\mathrm{P_{sem}}$ and concat $\mathrm{P_{act}}$ and $\mathrm{P_{sem}}$ as a prefix prompt, then we provide the prefix prompt along with an instruct prompt (user's instructions) to the LLM to condition its subsequent generation of speech emotion captions using prefix language modeling. This setup leverages external knowledge and the language understanding and modeling capabilities of the teacher-LLM to guide the student-LLM to generate plausible sentences.

Given a caption c_i with token T_i, language model P_θ learns to reconstruct y_i conditioned on the $\mathrm{P_{act}}$ and $\mathrm{P_{sem}}$. The probability of generating the next token is calculated as follows:

$$p_\theta(T_t \mid \underbrace{T_{0\sim a-1}}_{\mathrm{P_{act}}}, \underbrace{T_{a\sim b-1}}_{\mathrm{P_{sem}}}, \underbrace{T_{b\sim c-1}}_{\mathrm{P_{instruct}}}, \underbrace{T_{c\sim t-1}}_{\text{autoregressive}}) \tag{2}$$

This process is iterated until the LLM generates a token containing a period and the training loss is the maximum likelihood estimate, and the next token T_t is selected according to:

$$T_t = \arg\max_{i \in 1,\ldots,k} \left\{ p_\theta \left(c_i \mid p_n, T_0, \ldots, T_{t-1} \right) \right\} \quad (3)$$

Where prefix prompt $p_n = \mathrm{P}_{\mathrm{act}} \oplus \mathrm{P}_{\mathrm{sem}} \oplus \mathrm{P}_{\mathrm{instruct}}$. Trained on limited data, simply using semantic prompt as prefix prompt may overfit the In-Domain dataset, leading to significant domain shift and performance degradation of language model using out-of-domain (OOD) speech. In contrast, the acoustic prompt based on emotion-aware clues, inherits the powerful transferability from captions.

Speech Token Generation. For each speech, we adopt the pre-trained SpeechTokenizer (Zhang et al., 2024) to extract discrete representations and denote the tokens of the first residual vector quantization (RVQ) layer as speech tokens. The first layer of RVQ can be regarded as a semantic token, which contains more content information from speech, resulting in capturing semantically accurate emotional clues. We append this speech token x_t to LLM's input and generate the next token in an autoregressive modeling manner, for each time step t, the next token T_t is selected according to:

$$T_t = \arg\max_{i \in 1,\ldots,k} \left\{ p_\theta \left(c_i \mid x_t, T_0, \ldots, T_{t-1} \right) \right\} \quad (4)$$

Modality Alignment. Modality adapters (Deshmukh et al., 2024; Hu et al., 2024) are often used to compress the speech encoder's feature representations. Similar to (Yang et al., 2023), we treat the input from speech and text modality as a token sequence and learn a joint embedding space for all modalities. Speech tokens are expanded to text token's codebook in advance so that text and speech share the same codebook. We pad the shorter token sequence to make it the same length as the longer token sequence. We use a mask to ignore the padding part, ensuring that the model only focuses on valid tokens.

Knowledge Distillation. As shown in Fig 4, given a $\mathcal{D}_s = \{(x_n, y_n)\}$, we treat the LLM's prediction distribution $p_\theta \left(y_n \,|\, p_n, y_{<n} \right)$ of the next response token, after having observed the text input p_n and generated partial response $\{y_0, \ldots, y_{n-1}\}$, as the teacher distribution. Where p_n is the concatenation of $\mathrm{P}_{\mathrm{act}}$ and $\mathrm{P}_{\mathrm{sem}}$. In contrast, we consider the corresponding distribution $p_\theta \left(y_n \,|\, x_n, y_{<n} \right)$ for the speech input x_n as the student distribution. If speech and text are well-aligned, the two distributions should be close to each other, as measured by KL-divergence, which is as followed:

$$\begin{aligned} &\min_{\mathrm{LLM}_{\mathrm{stu}}(\cdot)} \mathcal{L}_{\mathrm{KL}}(p, x, y) = \\ &- \sum_{t, y_n} p_\theta \left(y_n | p_n, y_{<n} \right) \log p_\theta \left(y_n | x_n, y_{<n} \right) \end{aligned} \quad (5)$$

$\mathcal{L}_{\mathrm{KL}}$ introduces a quantitative measure of speech-text alignment at each step of the response generation process. By minimizing this loss, we can learning a student LLM using LoRA fine-tuning (Hu et al., 2022b) for speech input that facilitates generation behaviors similar to those of text inputs when generating speech emotion captions.

3.2 PO-Regularization

High-quality emotional description needs to consider not only the richness of emotions but also aspects such as consistency and rationality. The alignment of SEC's output to human preferences is often neglected. There is a problem that the LLM's response is inconsistent with the user's instructions (faithfulness hallucination) and results in false emotional descriptions (factuality hallucination). Therefore, we propose PO-Regularization to solve these problems.

Preference Pairs Creation. Inspired by (Ouyang et al., 2022; Yuan et al., 2024), we construct a preference pairs dataset by utilizing GPT-3.5 scoring prompt P_{score} on LLM's beam-search decoding output. The P_{score} to act as reward model is used to create preference pairs, which is as follow:

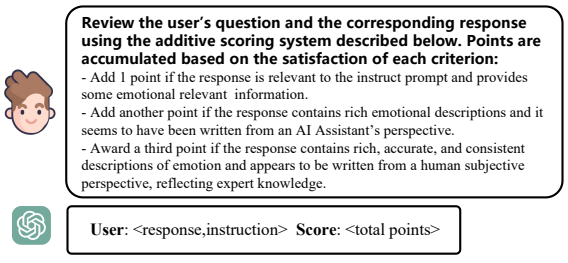


Figure 5: Scoring prompt for candidate responses.

Following above steps, we can get the preference pairs dataset $D_p = \{(x_n, y_n^c, y_n^r)\}_{n=1}^N$, which is consisted of chosen response y_n^c and rejected response y_n^r. Finally, we select the highest score as the y_n^c and the rest as y_n^r.

Preference Optimization. To solve the hallucination problem of LLMs, the prevalent RLHF methods (Ouyang et al., 2022; Touvron et al., 2023b;

Cui et al., 2023) involve fitting a reward model on the preference data, and the training the policy, value and critic models to maximize the reward without deviating too far from the reference model. However, RLHF method contains four models and has too many hyperparameters, making the training complex and high computation cost. Inspired by DPO (Rafailov et al., 2023; Yuan et al., 2024), We propose a simpler equivalent supervised approach PO-Regularization that addresses this reinforcement learning goal, the policy model can be directly optimized on the reward feedback based on preference pairs:

$$\mathcal{L}_{\mathrm{PO}} = \mathbb{E}_{(x,y_g,y_n^c)}\Big[\beta \log \sigma(\log \frac{\pi_\theta(y_g|x)}{\pi_{ref}(y_g|x)} - \log \frac{\pi_\theta(y_n^c|x)}{\pi_{ref}(y_n^c|x)})\Big] \quad (6)$$

Where β is is a hyperparameter and we only update the policy model $\pi_\theta(y|x)$ during finetuning, while reference model $\pi_{ref}(y|x)$ is the same as $\pi_\theta(y|x)$ which is frozen to prevent over-optimizing. PO-Regularization considers the likelihood of the preferred response y_n^c over dispreferred response y_n^r and optimizes the LLM towards this objective.

4 Experiments

4.1 Dataset

We select speech-caption paired samples from the large-scale video emotion reason dataset MER2023 (Lian et al., 2023a) to form the MER23SEC dataset. A Chinese interactive multimodal emotion corpus NNIME (Chou et al., 2017) is used to evaluate the transferability of our model trained on other datasets. Due to the lack of publicly available high-quality SEC task datasets, we propose a new dataset named EMOSEC[1], which is about 41 hours of Chinese-English Speech Emotion Captioning datasets. It consists of 15 male and 15 female speakers and covers 45039 sentences, with a sampling rate of 16kHz. We divide MER23SEC, EMOSEC, and NNIME datasets into training, validation and testing according to the ratio of 8:1:1.

4.2 Settings

Evaluation Metrics. We use GPT-3.5 to evaluate the degree of overlap of emotional clues and summarized states as shown in Fig 6. The automatic evaluation indicators are denoted as AES_c and AES_s respectively. The higher the score, the higher quality of generated captions.

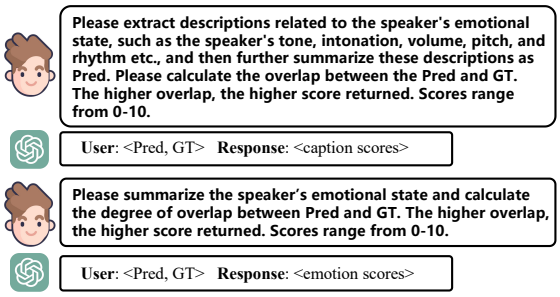


Figure 6: Prompt for Automatic Evaluation.

To evaluate the accuracy of the generated caption, we initially adopt traditional supervised metrics for the Automated audio captioning (AAC) task, containing standard natural language generation metrics BLEU(B@4), METEOR(M), ROUGE-L(R), CIDEr(C) (Vedantam et al., 2015), and SPICE(S) (Liu et al., 2017). B@4 focuses on the appearance frequency of emotional clues and is used to evaluate the emotional consistency and fine-grainedness of generated captions. Compared with B@4, M considers synonyms more, and R pays more attention to the sufficiency and faithfulness of output. C and S Compute accuracy of emotion captions using human consensus. Therefore, M can be used to evaluate factuality hallucinations, while R, C, and S is used to evaluate faithfulness hallucinations.

Baseline Systems. We compare our model with other systems. 1) HTSAT-BART (Mei et al., 2023): a three-stage processing framework, which performs exceptionally well in the AAC task. 2) NoAudioCap (Deshmukh et al., 2024): a weakly-supervised audio captioning model which requires a pre-trained CLAP (Wu et al., 2023). 3) SECap (Xu et al., 2024): the fisrt SEC model to generate high-quality speech emotion captions.

Training. For KD-Regularization, we optimize the student-LLM with the AdamW optimizer and the learning rate of 1e-5 on 4*V100 GPUs over 50k iterations, the batch size is 16. We employ DeepSpeed (Rajbhandari et al., 2020) and LoRA (Hu et al., 2022a) of rank 8 to implement model parallelism and parameter equivalence, applying warmup with 400 steps and gradient accumulation with 8 steps. For PO-Regularization, the learning rate is set to 5e-7 and train for 1000 steps.

[1]The EMOSEC dataset is accessible through: https://zenodo.org/records/10948423

Dataset	Methods	BLEU@4↑	METEOR↑	ROUGE↑	CIDEr↑	SPICE↑	AES_c↑	AES_s↑
NNIME	HTSAT-BART (Mei et al., 2023)	3.2±0.5	8.6±0.3	15.7±0.4	2.7±0.2	3.0±0.4	2.5±0.4	3.6±0.3
	NoAudioCap (Deshmukh et al., 2024)	4.9±0.4	10.4±0.2	17.6±0.3	4.9±0.5	5.1±0.3	3.6±0.2	4.2±0.1
	SECap (Xu et al., 2024)	5.8±0.4	11.4±0.3	17.9±0.2	8.6±0.4	5.3±0.3	4.9±0.5	4.5±0.2
	SECap-PO	6.0±0.3	12.1±0.3	18.6±0.3	8.9±0.2	5.4±0.4	5.1±0.1	4.7±0.4
	AlignCap-KD-RLHF	6.6±0.5	14.6±0.5	20.9±0.2	9.3±0.4	5.6±0.2	5.8±0.3	5.0±0.2
	AlignCap-KD-PO	**7.7±0.3**	**17.3±0.4**	**24.3±0.4**	**12.8±0.5**	**6.4±0.3**	**7.3±0.3**	**5.6±0.4**
EMOSEC	HTSAT-BART (Mei et al., 2023)	4.5±0.3	11.6±0.2	20.4±0.5	5.1±0.4	3.7±0.5	3.6±0.2	4.8±0.3
	NoAudioCap (Deshmukh et al., 2024)	6.7±0.3	14.5±0.4	21.8±0.6	10.3±0.6	5.7±0.3	4.7±0.3	5.4±0.1
	SECap (Xu et al., 2024)	7.4±0.3	16.6±0.2	25.9±0.3	11.2±0.3	5.8±0.3	5.9±0.3	5.6±0.2
	SECap-PO	7.5±0.4	17.0±0.4	26.2±0.2	11.8±0.2	6.0±0.3	6.1±0.4	5.8±0.2
	AlignCap-KD-RLHF	7.8±0.3	18.3±0.1	27.9±0.4	14.2±0.4	6.3±0.5	7.0±0.2	6.1±0.3
	AlignCap-KD-PO	**9.8±0.2**	**20.9±0.3**	**29.8±0.5**	**18.7±0.3**	**7.6±0.4**	**8.8±0.2**	**7.6±0.1**

Table 1: Zero-shot evaluation result of different SEC methods on NNIME and EMOSEC.

4.3 Main Results

For Zero-shot scenario, we conduct our model with baselines on NNIME (Chou et al., 2017) and EMOSEC dataset. Moreover, we evaluated the effects of two different Human preference alignments RLHF-PPO and DPO, on eliminating the hallucinations.

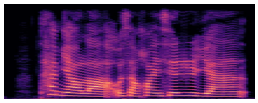
Quantitative Evaluation. The objective and automatic evaluation about zero-shot SEC methods are shown in Table 1, and we randomly select 25 sentences from test set to calculate scores. Our proposed preference-optimized models, AlignCap-KD-RLHF and AlignCap-KD-PO, outperform the baseline model in all metrics. The B@4 and M of AlignCap-KD-PO is higher than that of SECap-PO, which suggests that KD-Regularization can enhance the accuracy of emotional clues modeling. The highest R, C, and S scores demonstrate that AlignCap's output exhibits greater sufficiency and faithfulness compared to other baselines. The metrics of SECap-PO is higher than that of SECap, it is attributed to the PO-Regularization, which eliminates the faithfulness hallucinations where the output is inconsistent with user instructions. AlignCap-KD-PO achieves the highest B@4 score, demonstrating that emotional clues as P_{act} can generate more fine-grained emotion captions. It outperforms AlignCap-KD-RLHF, indicating superior performance in quantitative evaluation. This confirms that DPO-based PO-Regularization can enhance the quality of the caption generated by the model than RLHF-PPO. It also demonstrates that human preference alignment is an effective method for the SEC model to undergo self-improvement.

Compared with NoAudioCap (Deshmukh et al., 2024) which also utilizes a similar text-only training method, both AlignCap-KD-RLHF and AlignCap-KD-PO comprehensively surpass NoAudioCap, attributed to our proposed KD-Regularization in alleviating speech-text distribution gap after LLM decoding. Zero-shot inference used by NoAudioCap has a training-inference mismatch, which loses generalization on unseen speech. The KL-divergence in KD-Regularization is used to bridge the mismatch, it suggests that AlignCap can be generalized to unseen speech.

Qualitative Evaluation. Figure 7 supports the findings of Table 1 by presenting the output of AlignCap and HTSAT-BART (Mei et al., 2023), NoAudioCap (Deshmukh et al., 2024), SECap (Xu et al., 2024). Our method can produce richer emotional clues and more coherent emotion captions. In the "Neutral" example of Figure 7, although SECap (Xu et al., 2024) can produce rich speech emotion captions, its incorrect emotional cues are inconsistent with the real emotion. In the "Surprise" example, the output of NoAudioCap lacks fine-grained captions of the speaker's gender, tone, and intonation. AlignCap-KD-PO not only makes up for this shortcoming but also outputs the speaker's content consistent with the transcribed text, which enhances the understanding of the speech content.

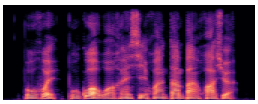
In the "Angry" example in Figure 7, AlignCap-KD-RLHF simply refers to gender as "a person", AlignCap-KD-PO can correctly identify its gender by adopting preference optimization, it is also attributed to P_{act} for enriching fine-grained information about the speaker.

What's more, NoAudioCap (Deshmukh et al., 2024) suffers from LLM's output inconsistent with user instructions, and both AlignCap-KD-



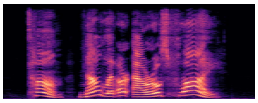
Emotion: "Happy"
Transcription: "Great ! I want to exchange some yen."
HTSAT-BART: A woman is speaking, and her tone sounds positive.
NoAudioCap: The audio is of a woman speaking, in a fast pace, with a neutral tone, and no apparent emotion.
SECap: Based on the voice, it sounds like this person is fearful in the audio, and possibly sad. They might be expressing a need or desire for something, possibly related to money or finances.

AlignCap-KD-RLHF: The person in the audio speaks quickly, with a rising intonation and a positive tone.
AlignCap-KD-PO: The woman in the audio is happy and cheerful. She speaks quickly, with a rising intonation and a positive tone.



Emotion: "Neutral"
Transcription: "No, but I love snowboarding very much."
HTSAT-BART: The tone of voice indicates that the person in the audio is emotionally calm.
NoAudioCap: The audio is of a man speaking, in a neutral tone, at a normal speed, with no voice changes or emotions.
SECap: Based on the voice, it sounds like this person is happy in the audio, and they are expressing excitement or enthusiasm.

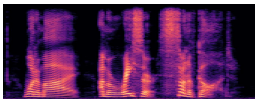
AlignCap-KD-RLHF: The emotion of the person in the audio is neutral, the tone is normal, and the speaking speed is medium.
AlignCap-KD-PO: The man in the audio is neutral. His speaking tone is normal, the intonation is calm, and the speaking speed is medium, saying, \"No, but I love snowboarding very much.\".



Emotion: "Surprise"
Transcription: "Tom now let our arrows fly!"

HTSAT-BART: The emotion of the person in the audio is angry. His tone is negative and impatient, and his speaking speed is fast.
NoAudioCap: The tone of voice indicates that the person in the audio is emotionally angry.
SECap: Based on the voice, it sounds like this person is happy in the audio, and their emotions are: happy.

AlignCap-KD-RLHF: The person speaks in a happy tone, with a fast speaking speed. The tone of voice indicates the emotion of the person in the audio is surprise.
AlignCap-KD-PO: This is a man speaking, in a positive and happy tone. He speaks English with a fast speaking speed, and the content was: "Tom now let our arrows fly!".



Emotion: "Angry"
Transcription: "What am I, I'm a racer, son of god!"

HTSAT-BART: This is a man speaking, in an excited tone, with a fast speaking speed.
NoAudioCap: The content of voice is: "What am I? I'm a racer, son of God!".
SECap: Based on the voice, it sounds like this person is angry in the audio, possibly expressing disbelief or frustration.

AlignCap-KD-RLHF: The person in the audio is in an angry mood, has an angry tone, and speaks at a normal speed.
AlignCap-KD-PO: The man in the audio is angry. The voice is vehement and the tone is excited. He speaks at a normal speed. The content of the speech is: "What am I? I'm a racer, son of God!".

Figure 7: Qualitative Results of Zero-shot SEC with different methods. Incorrect emotional clues in captions are highlighted in red, while correct emotional clues in captions are in blue.

PO and AlignCap-KD-RLHF eliminate this faithfulness hallucination, owing to our proposed PO-Regularization.

4.4 Ablation Studies

As shown in Table 2, we train AlignCap with specific components selectively removed to evaluate the effect of the proposed components to eliminating hallucinations and enrich fine-grained information.

The decrease in all objective evaluation scores shows the significance of acoustic prompt (P_{act}), KD-Regularization ($\mathcal{L}_{KL}$), and PO-Regularization ($\mathcal{L}_{PO}$). The significant decrease of P_{act} on B@4 proves the positive effect of the emotional clues extracted by P_{act} on the emotional consistency of generated captions. The lack of $\mathcal{L}_{KL}$ and $\mathcal{L}_{PO}$ leads to a significant drop in M and R, indicating that they play a crucial role in guiding Zero-shot SEC model to eliminate factuality and faithfulness hallucinations. The model without P_{act} also exhibits a decrease in R, indicating its effectiveness in generating fine-grained emotional descriptions. The M score of AlignCap-KD-PO is higher than AlignCap-KD-RLHF without adopting explicit reward modeling, allowing it to learn more human-like generated captions.

Methods	P_{act}	$\mathcal{L}_{KL}$	$\mathcal{L}_{PO}$	EMOSEC B@4↑	M↑	R↑	C↑
AlignCap-KD-RLHF	✔	✔	✔	/	/	/	/
	-	✔	✔	-3.5	-2.3	-2.7	-1.2
	✔	-	✔	-1.9	-5.4	-4.3	-1.9
	✔	✔	-	-1.5	-2.7	-3.2	-1.6
AlignCap-KD-PO	✔	✔	✔	/	/	/	/
	-	✔	✔	-1.4	-1.3	-1.8	-0.7
	✔	-	✔	-0.8	-4.7	-3.0	-1.6
	✔	✔	-	-0.5	-2.6	-2.2	-0.9

Table 2: Ablation studies on EMOSEC dataset.

Methods	B@4↑	M↑	R↑
	EMOSEC→NNIME		
HTSAT-BART	1.9±0.4	3.4±0.5	6.1±0.3
NoAudioCap	4.2±0.2	8.7±0.4	10.3±0.6
SECap	3.4±0.3	8.2±0.3	13.8±0.5
AlignCap-KD-RLHF	5.2±0.4	9.6±0.3	14.4±0.4
AlignCap-KD-PO	**5.9±0.3**	**10.1±0.5**	**15.6±0.2**
	EMOSEC→MER23SEC		
HTSAT-BART	4.4±0.3	11.1±0.4	12.3±0.4
NoAudioCap	11.3±0.3	13.2±0.3	18.7±0.3
SECap	9.8±0.3	14.6±0.3	16.1±0.3
AlignCap-KD-RLHF	14.5±0.2	19.8±0.4	21.0±0.5
AlignCap-KD-PO	**16.3±0.3**	**21.6±0.3**	**22.7±0.3**

Table 3: Cross-domain SEC results on NNIME and MER23SEC dataset.

5 Analysis

5.1 Transferability on Cross-Domain Speech

As shown in Table 3, we evaluate the AlignCap in a cross-domain scenarios where the training data and testing data are from different datasets. We conduct experiments on NNIME and MER23SEC's testing set, and we only use the training set of EMOSEC's captions to fine-tuning AlignCap.

In EMOSEC→NNIME cross-domain scenarios, the results show that AlignCap outperforms all baselines. The B@4 and M metrics of SECap and HTSAT-BART are lower than NoAudioCap on the NNIME dataset. This is because they all have encoder-decoder structures and are trained on well-paired data, lacking components to enhance generalization capabilities for cross-domain data.

AlignCap outperforms NoAudioCap, demonstrating the superiority of the proposed KD-based speech-text alignment over the CLAP-based (con-

trastive learning) speech-text alignment used in NoAudioCap for cross-modal mapping. It not only bridges the audio-text distribution gap, but also improves the generalization ability in cross-domain scenarios.

Additionally, there is a domain offset between the predicted emotional description generated by LLMs and the real description of the target domain, leading to performance degradation. Equipped with the PO-Regularization, AlignCap-KD-PO outperforms the other baselines including AlignCap-KD-RLHF version on most metrics, demonstrating the effectiveness of the proposed compoents.

5.2 Effect of Different Speech-Text Alignment on Downstream SEC task

Previous alignment methods, such as Gaussian Noise Injection (CL+NI-Align) (Deshmukh et al., 2024) and Project-based Decoding (CL+Proj-Align) (Kouzelis and Katsouros, 2023), achieve alignment by adding Gaussian noise variance or mapping based on contrastive learning between speech and text embeddings before LLM decoding. The KD-Regularization (KD-Align) we proposed achieves speech-text alignment after LLM decoding and alleviates the information loss in modality alignment. Fig 9 shows that our method outperforms other alignment methods in all indicators, attributing to we treat speech-text alignment as a knowledge distillaiton problem. It can ensure that the LLM's responses to speech inputs closely mirror those to corresponding text inputs.

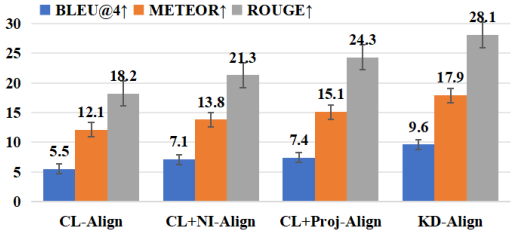


Figure 8: The impact of different Alignment.

5.3 Performance on Different Preference Pair Sizes and Steps.

As shown in Fig 9, we examine the effect of different preference pair sizes and fine-tuning steps for PO-Regularization on the performance of AlignCap. We set the preference pair sizes to be $\{0, 25k, 50k, 75k, 100k\}$. After 500 steps of fine-tuning with DPO for each of these sizes, we assess their performance in zero-shot SEC. We can observe notable improvement with increasing sizes from 0 to 50k, which indicates that an increase in preference pair data can improve zero-shot SEC. However, using more than 50k preference data for DPO does not lead to significant performance improvements, indicating a threshold beyond which additional data does not enhance learning outcomes.

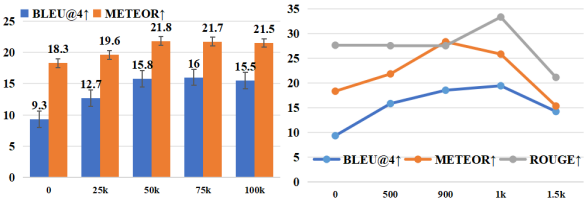


Figure 9: **Left:** Performance of AlignCap across different preference pair sizes. **Right:** Performance of AlignCap of different fine-tuning steps.

Moreover, we set the fine-tuning steps to be 0→1.5k for AlignCap on zero-shot SEC evaluation. As shown in Fig 9, all metrics demonstrate significant performance improvement when the number of fine-tuning steps is less than 1k. However, when the number of iterations exceeds 1k steps, the model suffers from overfitting, resulting in performance degradation, indicating that 1k steps are the optimal iteration steps for PO-Regularization.

5.4 Can PO-Regularization Works with Small Models?

We investigate whether PO-Regularization can bring improvements for smaller language models. The preference pair size is 50k and we fine-tuning the models on EMOSEC dataset for 1k steps. We evaluate the zero-shot SEC performance on EMOSEC test set. Tab 4 shows that after 1k iterations, PO-Regularization significantly boosts OPT's scores but decreases the GPT2-base's scores on M and R, while improving GPT-2-large very little. This indicates that PO-Regularization can improve caption generation in small language models, although the improvement is not significant for models with very small parameters.

Models	Parameters	B@4↑	M↑	R↑
GPT2-base	124M	3.3±0.3	8.4±0.5	16.1±0.3
GPT2-base-PO		3.3±0.5	8.3±0.4	16.0±0.5
GPT2-large	774M	3.7±0.2	8.9±0.3	17.3±0.4
GPT2-large-PO		3.8±0.3	9.1±0.4	17.6±0.2
OPT	1.3B	4.3±0.2	10.2±0.4	18.8±0.5
OPT-PO		5.4±0.3	12.8±0.2	20.5±0.4

Table 4: Performance on Small Models.

6 Conclusion

We proposed AlignCap, achieving speech-text alignment and human preference alignment. To minimize the distribution gap between LLM's response to speech input and those to corresponding text inputs, we design KD-Regularization to achieve speech-text alignment. Additionally, we align emotion captions to human preference by PO-Regularization. This process eliminates the factuality and faithfulness hallucinations of AlignCap on unseen speech. Experiments demonstrate AlignCap's superiority in both zero-shot and cross-domain scenarios.

Limitations

Well-paired speech-caption datasets are difficult to obtain in real-world scenarios. Captions containing emotional descriptions are easy to obtain, but high-quality speech-caption paired data is difficult to collect, how to solve this mismatch problem will be left to our future work. In addition, enhancing the robustness of alignment between speech and text inputs remains an urgent issue that needs to be addressed in the future.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*, pages 65–72.

Shuaiqi Chen, Xiaofen Xing, Weibin Zhang, Weidong Chen, and Xiangmin Xu. 2023. Dwformer: Dynamic window transformer for speech emotion recognition. In *ICASSP*, pages 1–5.

Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. 2017. NNIME: the NTHU-NTUA chinese interactive multimodal emotion corpus. In *ACII*, pages 292–298.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *CoRR*, abs/2310.01377.

Soham Deshmukh, Benjamin Elizalde, Dimitra Emmanouilidou, Bhiksha Raj, Rita Singh, and Huaming Wang. 2024. Training audio captioning models without audio. In *ICASSP*, pages 371–375.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *TASLP*, 29:3451–3460.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022a. LoRA: Low-rank adaptation of large language models. In *ICLR*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022b. LoRA: Low-rank adaptation of large language models.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. Wavllm: Towards robust and adaptive speech large language model. *CoRR*, abs/2404.00656.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *CVPR*, pages 7661–7671.

Theodoros Kouzelis and Vassilis Katsouros. 2023. Weakly-supervised automated audio captioning via text only training. In *DCASE*.

Shan Li, Weihong Deng, and Junping Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2584–2593.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. 2023a. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *ACM MM*, page 9610–9614.

Zheng Lian, Haiyang Sun, and Licai et al. Sun. 2023b. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *ACM MM*, page 9610–9614.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *ICCV*, pages 873–881.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *CoRR*, abs/2303.17395.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,

Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, volume 35, pages 27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, volume 36, pages 53728–53741.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, page 20.

Leonard Salewski, Stefan Fauth, A. Sophia Koepke, and Zeynep Akata. 2023. Zero-shot audio captioning with audio-language model guidance and audio context keywords. *CoRR*, abs/2311.08396.

Haoxiang Shi, Ziqi Liang, and Jun Yu. 2024. Emotional cues extraction and fusion for multi-modal emotion prediction and recognition in conversation. *CoRR*, abs/2408.04547.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. Secap: Speech emotion captioning with large language model. In *AAAI*, pages 19323–19331.

Zhen Yang, Yingxue Zhang, Fandong Meng, and Jie Zhou. 2023. TEAL: tokenize and embed ALL for multi-modal large language models. *CoRR*, abs/2311.04589.

Jiaxin Ye, Yujie Wei, Xin-Cheng Wen, Chenglong Ma, Zhizhong Huang, Kunhong Liu, and Hongming Shan. 2023. Emo-dna: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition. In *ACM MM*, pages 5956–5965.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *CoRR*, abs/2401.10020.

Susan Zhang, Stephen Roller, and Naman Goyal et al. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *ICLR*.