Multi2Cap: Improving Automated Audio Captioning with Cross-modal Feature Distillation and Large Language Model

Anonymous ACL submission

Abstract

This study proposes VggCaps and Multi2Cap. VggCaps is a dataset designed to address the limitations of current Automated Audio Captioning (AAC) datasets, which typically feature short and simple captions. VggCaps includes longer, more complex captions and has been validated through comprehensive analysis and human evaluations. Multi2Cap is a framework that extends AAC tasks into a multi-modal domain. By leveraging Cross-modal Feature Distillation (CFD) and Large Language Model (LLM), Multi2Cap improves the quality and depth of the generated captions. Multi2Cap achieves state-of-the-art results on the Clotho and AudioCaps benchmarks, offering new challenges and directions for AAC research.

1 Introduction

Automated Audio Captioning (AAC, Drossos et al., 2017) is the task of generating natural language descriptions from audio content, and it is becoming an increasingly important task in the field of artificial intelligence. Unlike Automatic Speech Recognition (ASR, Benesty et al., 2008), which converts speech to text, AAC must comprehensively understand and describe not only linguistic elements but also non-verbal audio signals (e.g., environmental noise, animal sounds, music, etc.). Although AAC research is a relatively recent field, it is receiving growing attention due to the increasing demand for complex utilization of audio data, such as audio interaction and retrieval (Mei et al., 2022; Xu et al., 2023). Driven by this demand, AAC is actively ongoing, with new technologies continuously being developed to efficiently process and describe diverse audio information (Wu et al., 2024; Liu et al., 2024).

However, existing AAC researches (Liu et al., 2024; Kim et al., 2024a) have primarily been limited to the audio-text modality, leaving the potential of multi-modal learning underexplored. Real-world experiences are composed of multi-modalities that combine various types of information (Marian et al., 2021; Holmes et al., 2024), such as visual and auditory, and humans understand and explain the world through multiple senses. Moreover, research in artificial intelligence that reflects the real world through multi-modal approaches has been gaining attention (Li et al., 2024b), and many studies are actively exploring the integration of different modalities to complement each other and improve performance across various fields and tasks (Li et al., 2024a). In this context, AAC task require new insights from a broader multi-modal perspective. Utilizing additional modalities in AAC tasks will play a crucial role in handling more complex auditory situations and providing richer descriptions.

Furthermore, current AAC datasets tend to describe audio content in a short and concise manner (Kim et al., 2019; Mei et al., 2024). In fact, we identified that the existing datasets are represented by an average of 9 words, and the lexical diversity of them are not abundant, which is illustrated in table 1 and figure 2. This can lead to overfitting (Eldan and Li, 2023) during training and, limit the ability to provide rich and detailed descriptions of audio content, restricting the task's scalability. In the real world, audio is not described merely as sound (Heittola et al., 2021; Zhang et al., 2024); it can be explained with complex elements, including the surrounding environment, the source of the sound, and its context. Therefore, it suggests that AAC tasks need to evolve to express this complex information and provide longer descriptions.

Based on the above motivations, we construct and propose VggCaps, a multi-modal audio captioning dataset that incorporates visual elements. VggCaps is designed to include longer and more complex captions compared to existing datasets. Specifically, we design the prompt to LLMs for incorporate both audio and image data simultane-

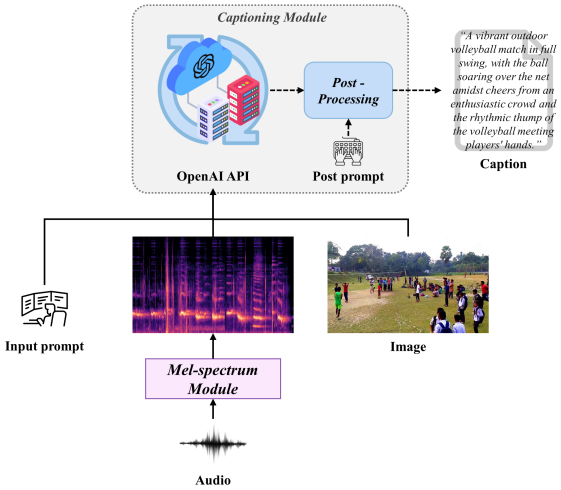


Figure 1: **Pipeline of VggCaps Data Processing**

ously, enabling the model to capture the complex relationships between the two modalities more precisely. These captions focus on using visual information complementarily to provide more accurate and enriched descriptions of the audio's meaning. Additionally, to verify the validity and robustness of the dataset, we conduct human evaluations and measured human performance on the dataset.

Additionally, we propose the Multi2Cap framework, a multi-modal approach that extends the existing AAC tasks by incorporating visual information. The goal is to improve the quality and depth of audio captions using image-audio feature distillation and LLMs. Specifically, we apply the Cross-modal Feature Distillation (CFD) technique, which uses image modality complementarily to learn the interactions between audio and images. The audio encoder learns the unique features of the image, maximizing the interaction between the two modalities. This allows the generation of richer audio descriptions based on more comprehensive information. In this study, we utilize VggCaps as pre-training data to train Multi2Cap, and report model performance across various benchmarks.

Main contributions of this study are as follows:

- **We attempt a multi-modal expansion of AAC**. By utilizing image modality complementarily, we perform the AAC task and present a new paradigm in AAC research. This multi-modal approach broadens the boundaries of the traditional AAC task and provides new motivation for future research.
- **We propose a novel multi-modal audio captioning dataset with added visual information**. Using LLMs as a guide, we generate captions by supplementing visual information based on audio content, and these captions are longer and contain more diverse vocabulary than existing datasets. This leads to the creation of a dataset that requires more intricate and high-level descriptions compared to traditional audio captioning datasets.
- **We propose the Multi2Cap framework**, which enhances the understanding of audio-visual interactions for AAC tasks through Cross-modal Feature Distillation. Additionally, through various experiments, we analyze the impact of adding extra modalities to AAC tasks.

2 Related works

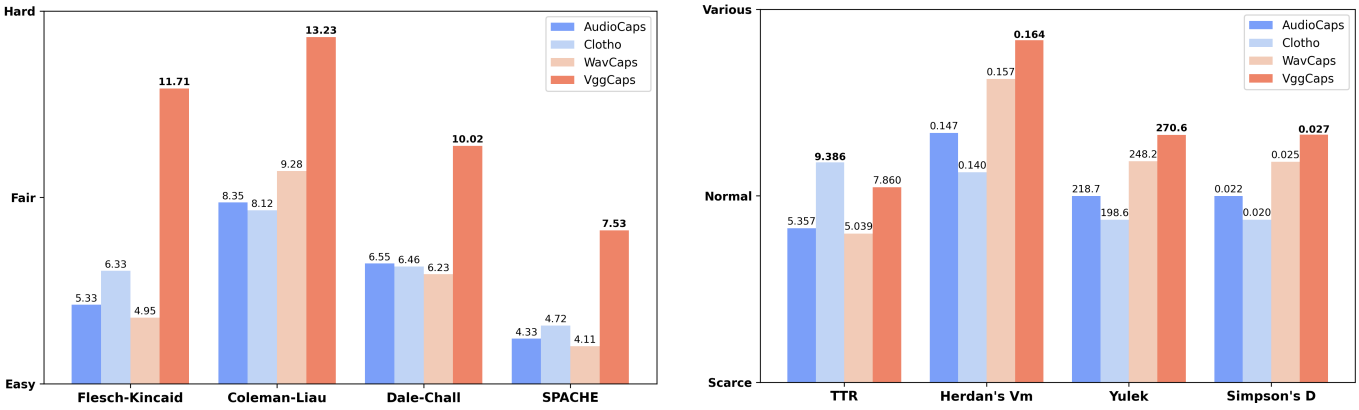
2.1 Automated Audio Captioning

Automated Audio Captioning (AAC) is a task that generates natural language descriptions from audio content, with serious research beginning around 2017. AAC can often be confused with Automatic Speech Recognition(ASR), but the two tasks are fundamentally different (Benesty et al., 2008). ASR focuses on converting speech from audio into text, aiming to transcribe the content of the audio literally. In contrast, AAC describes not only speech but also surrounding non-verbal sounds. In other words, AAC requires a comprehensive audio understanding ability to interpret the context of the sounds and describe them (Narisetty et al., 2022). Therefore, AAC demands a high-level ability to understand various sounds and their backgrounds, requiring a more in-depth processing of audio data. Due to these characteristics, AAC presents new challenges and unique motivations.

AAC research has primarily developed around neural network models with an encoder-decoder structure. Initially, the approach of combining Convolutional Neural Networks(CNN) and Recurrent Neural Networks(RNN) (Drossos et al., 2017; Mei et al., 2022) was commonly used for ACC tasks. More recently, research utilizing pre-trained Transformer models for both the encoder and decoder has gained attention (Xu et al., 2023). This developmental process has contributed to generating more sophisticated audio descriptions. Furthermore, the emergence of Large Language Models (LLMs) has brought about a significant breakthrough in the field of artificial intelligence, and has also opened new avenues in AAC research (Bommasani et al., 2021). Leveraging the capabilities of LLMs, AAC research has seen substantial performance improvements (Liu et al., 2024). However, existing AAC

Dataset	num. of row	num. of audio	avg(std). audio legnth(s)	num. of caption	avg(std). caption legnth	additional modal
AudioCaps (2019)	57,188	51,308	10.0 (0.6)	57,188	9.0 (4.3)	Image(Potentially)
Clotho (2020)	29,645	5,929	22.5 (4.3)	29,645	11.3 (2.8)	X
WavCaps (2024)	403,050	403,050	67.6 (-)	403,050	7.8 (-)	X
VggCaps (ours)	**173,494**	**173,494**	**10.0 (0.1)**	**173,494**	**21.1 (5.3)**	**Image**

Table 1: **Statistics of Dataset**: We statistically compare the existing AAC dataset with VggCaps. VggCaps includes longer captions and additional modalities compared to the existing datasets.



a) Readability Level

b) Lexical Diversity

Figure 2: **Readability Level and Lexical Diversity Comparison by Datasets**

research (Kim et al., 2024a) has been limited to the modality of Audio-Text. To address this issue, this study proposes a new approach to the multi-modal expansion of AAC tasks.

2.2 Feature Distillation

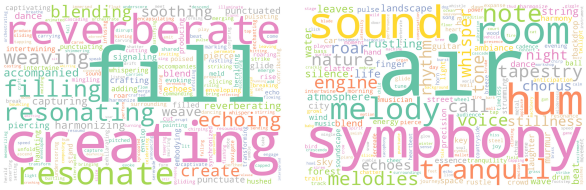
Knowledge Distillation (KD, Hinton, 2015) is a deep learning technique that distills knowledge from a complex and high-performance model to a relatively simpler and less complex model. The goal of traditional knowledge distillation is model compression, aiming to reduce computation and memory usage while maintaining the performance of the teacher model. On the other hand, recent studies have focused on effectively distilling the unique knowledge of the teacher model (Niu et al., 2024; Liu et al., 2022). Particularly, research has been actively conducted on improving the richness and precision of knowledge transfer between different modalities from a multi-modal perspective (Ienco and Dantas, 2024; Huo et al., 2024), not limited to a specific task or domain. Specifically, in the context of Contrastive Learning (Sang et al., 2024; Yu et al., 2023), KD is performed by comparing features across modalities in the same feature space. This technique (Schroff et al., 2015) trains the model so that similar inputs are positioned closer in the feature space, while dissimilar inputs are pushed farther apart. This approach enhances the interaction between modalities and helps align learned representations more consistently and effectively.

In light of this, this study proposes Multi2Cap, which applies the Cross-modal Feature Distillation (CFD) technique, focusing on aligning the interaction between audio and image within the feature space by utilizing image modality as a supplementary component.

3 Multi-modal Audio Captioning Dataset

In this section, we propose and describe the VggCaps dataset for AAC tasks. We explain the pipeline for its construction and post-processing, as well as the multi-faceted analysis of the dataset and the human evaluation procedure. Additionally, this dataset should not be distributed in the real world for any purpose other than research.

We utilize VggSound(Chen et al., 2020) data to construct the LLM-Guided Audio Captioning Dataset. VggSound is a large-scale audio-visual dataset, designed to enable learning from the sounds occurring in videos. VggSound includes various types of sounds along with the video scenes associated with them, and it has been widely used in a variety of research (Senocak et al., 2021; Wang et al., 2023). However, since VggSound is an audio-visual dataset, it does not contain captions (descriptions) mapped to each audio. Therefore, we utilized a Large Language Model (LLM) to generate appropriate captions.



a) VERB b) NOUN

Figure 3: **Wordcloud in VggCaps**

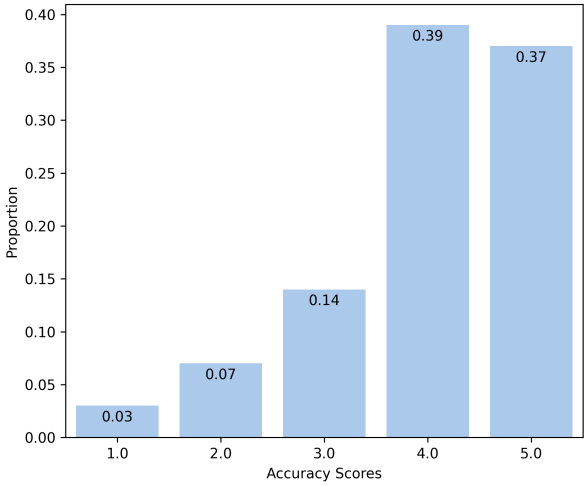


Figure 4: **Mean Opinion Score (MOS)**: This table shows the distribution of MOS for VggCaps calculated through human evaluation. It indicates that the vast majority of samples have appropriate captions.

3.1 Data Processing

The VggSound dataset contains only category information for audio and video, but no captions. To generate appropriate captions from such a dataset, we utilized a Large Language Model(LLM) to create captions mapped to the audio-visual data. The processing is conducted through the pipeline shown in figure 1.

As input for generating captions, we extract an appropriate snapshot (image) from the video. Then, we extract an audio clip of 10 seconds, consisting of 5 seconds before and after the snapshot. We convert the extracted audio into a Mel-spectrum(Hannun et al., 2014) and input it into the LLM along with the video snapshot. Additionally, we provide the input with a prompt to generate suitable captions, and finally feed it into the LLM. The raw captions generated this way undergo a Post-Processing step. This step refines the captions by removing exaggerated or unnecessary expressions and polishing them into cleaner sentences. The final captions used in this study are those generated after Post-Processing. The Captioning Module in Figure 1 utilizes GPT-4o(Shahriar et al., 2024).

In this study, we obtained a total of 173,494 VggCaps data samples. This dataset is used as the pretraining dataset for the Multi2Cap framework discussed later. Additionally, 1% of the total dataset, randomly selected, was used as a test subset for validation and performance reporting. Further details and analysis of the dataset are provided in section 3.2.

3.2 Dataset Analysis

The analysis of the constructed data focuses on comparing the generated captions with existing AAC datasets. Table 1 provides a statistical comparison between our dataset and the existing datasets. There are two notable differences. First, the caption length is about twice as long as that of the existing datasets, which is a result of using LLM to describe the content with richer vocabulary. Second, Snapshot images extracted from the video were included as an additional modality. This allows the VggCaps dataset to enable multi-modal research in AAC tasks.

The second analysis evaluates how the constructed VggCaps dataset uses more diverse vocabulary and describes the content in a more complex manner compared to the existing datasets. The analysis focuses on readability level and lexical diversity. The results of the analysis are provided in figure 2. First, readability level (figure 2a) is evaluated using four metrics: Flesch-Kincaid(Flesch, 1948), Coleman-Liau(Coleman and Liau, 1975), Dale-Chall(Dale and Chall, 1948), and SPACHE(Spache, 1953). These metrics indicate that the lower the score, the easier the text is to read, whereas higher scores indicate the need for deeper understanding. In all metrics, the VggCaps shows a higher level compared to the existing datasets. Lexical diversity (figure 2b) is analyzed using four metrics: Type-Token Ratio(TTR)(Templin, 1957), Herdan's VM(Herdan, 1960), Yulek(Yule, 2014), and Simpson's D(Simpson, 1949). Higher values for these metrics indicate the use of more diverse vocabulary. In the figure, each metric is normalized to a scale from 0 to 10, with the actual values before normalization displayed. The analysis confirms that the constructed dataset uses a more diverse vocabulary compared to the existing datasets.

Finally, figure 3 shows the word cloud of the captions in the constructed dataset. For verbs, it can be observed that more linguistically sophisticated expressions such as "fill" and "reverberate" are used, rather than simple expressions like "hear" and "sound." Additionally, for nouns, not only words with auditory meanings but also words with spatial

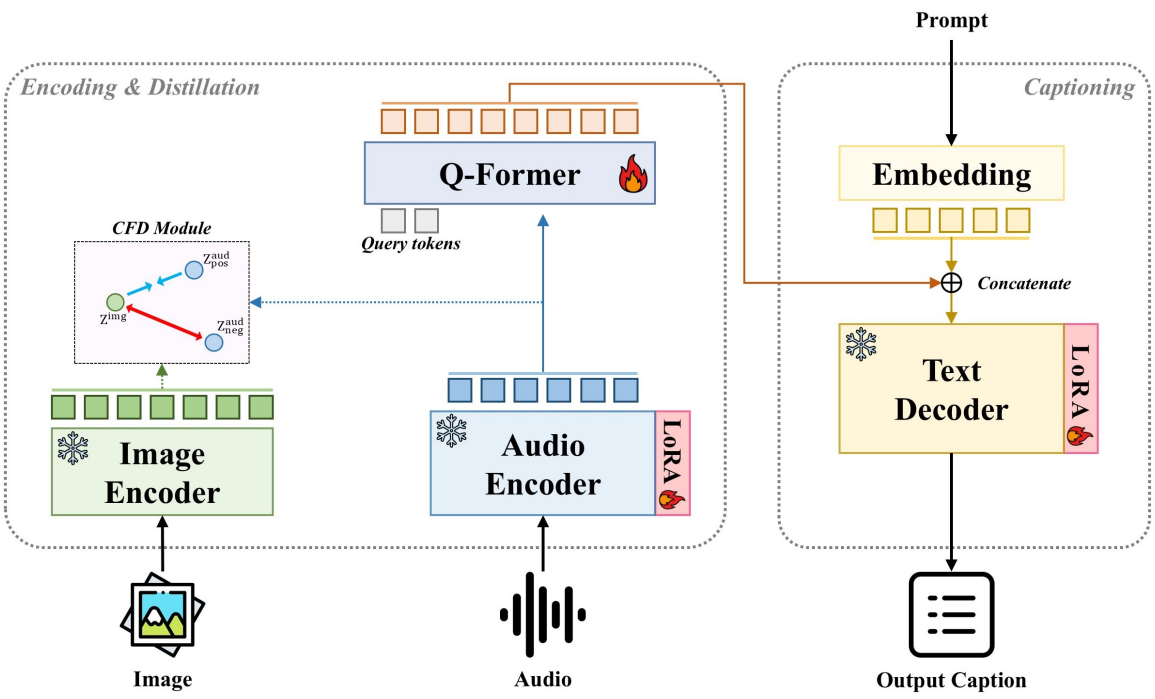


Figure 5: **Overview of Multi2Cap Pre-training Flow & Cross-modal Feature Distillation (CFD) Module**

or visual meanings are included.

3.3 Human Evaluation/Performance

To verify the validity and robustness of the constructed dataset, we conducted an experiment to perform human evaluation on a subset of the VggCaps dataset and derive human performance. For this purpose, 100 samples were randomly selected from the test subset of the VggCaps, and we recruited 18 evaluators who volunteered to participate. Among them, 10 evaluators were responsible for the human evaluation of the captions, while the remaining 8 were tasked with human performance.

The purpose of the human evaluation was to assess how accurately the captions of the VggCaps describe the audio content. The evaluators were provided with audio samples and asked to evaluate how accurately each caption described the corresponding audio. For this, they were instructed to assign a score between 1 and 5 based on the Mean Opinion Score(MOS, Streijl et al., 2016) method. A score of 1 indicates that the caption does not describe the audio accurately at all, while a score of 5 indicates that the caption perfectly describes the audio. The evaluation results measured an MOS score of 4.1 ± 0.09. This suggests that the captions of VggCaps are generally accurate and reliable, with a high level of agreement among the evaluators. The distribution of MOS scores is visually presented in Figure 4. Additionally, the evaluators checked whether the evaluation data contained any sensitive information and agreed that it did not.

Human performance experiment was conducted in two stages. In the first stage, the evaluators were asked to generate captions for the audio content provided to them. In the second stage, aligned video snapshots were provided as supplementary material along with the audio, and the evaluators were asked to generate captions based on this information. This aimed to evaluate how humans perform in single-modality versus multi-modality situations. In other words, it allowed us to assess the impact of providing multi-modal information on caption generation and compare how performance changes when evaluators utilize multi-modal information. The specific results and analysis are further detailed in the section 5.2 and table 2.

4 Multi2Cap

In this section, we elaborate our proposed Multi2Cap. The operation of the Multi2Cap framework is illustrated in figure 5.

4.1 Creating Caption

Multi2Cap is structured similarly to previous mainstream methods that combine audio encoding and text decoding. The audio encoder and text decoder are composed of CED(Dinkel et al., 2024) and LLaMA(Touvron et al., 2023), respectively, based on previous research, and each is fine-tuned using LoRA(Hu et al., 2021). Q-Former(Li et al., 2023) is constructed between the audio encoder

		BL-1	BL-2	BL-3	BL-4	RG-L	ME	CD	SP	SD	SD-F	SB	FS
w/o image	Human	46.1	22.9	**14.6**	**8.4**	29.6	13.5	15.4	7.5	11.4	11.4	48.6	48.6
	LOAE (2024)	33.0	18.3	10.9	7.0	29.1	12.4	47.6	14.2	30.9	27.6	61.8	55.3
w/ image	Human	**49.2**	**27.0**	13.6	7.5	**32.9**	**14.8**	19.4	7.5	13.4	13.4	48.9	48.9
	Multi2Cap$_{\text{small}}$	33.8	18.9	11.5	7.5	29.5	13.7	52.1	15.4	33.7	30.8	62.6	57.7
	Multi2Cap$_{\text{base}}$	34.1	19.2	11.6	7.6	29.9	13.9	**55.0**	**15.5**	**35.2**	**33.7**	**63.5**	**59.7**

Table 2: **Performance comparisons on VggCaps**: This table shows the performance of Multi2Cap on the VggCaps dataset. Each column represents an evaluation metric, and the abbreviations for the metrics are mentioned in section 5.1. Multi2Cap is compared with LOAE(2024), which utilized the same LLM. The performance shows superior results across all metrics. Best performance for each metric is in **Bold**, and the second-best is Underlined.

and text decoder, and both the audio embedding and prompt are input together into the text decoder. The audio encoder, Q-Former, and text decoder are trained to optimize the cross-entropy loss $\mathcal{L}_{cap}$ as shown in the equation below. $\mathcal{L}_{cap}$ is expressed as follows:

$$\mathcal{L}_{\text{cap}} = \texttt{CrossEntropy}(\text{origin}, \text{output}) \quad (1)$$

where, origin refers to the ground truth caption, and output refers to the caption generated by the model.

4.2 Cross-modal Feature Distillation

An image encoder is additionally constructed as the content-teacher of the audio encoder to learn image representations. For the image encoder, ResNet50(He et al., 2016) and ViT-Large(Dosovitskiy, 2020) are used, and the weights are fixed. In the proposed framework, the image encoder, as the content-teacher of the audio encoder, performs Cross-modal Feature Distillation (CFD) with the audio encoder. CFD is optimized through contrastive learning(Schroff et al., 2015) between the image representation Z^{img} extracted from the image encoder and the audio representations Z^{aud}_{pos} and Z^{aud}_{neg} extracted from the audio encoder. The image feature (Z^{img}) is trained to be closer to Z^{aud}_{pos} in the feature space ($\in \mathbb{R}^{dim}$), while it is trained to be farther from Z^{aud}_{neg} in the feature space. The objective function of the CFD module, $\mathcal{L}_{cfd}$, is expressed as follows:

$$\mathcal{L}_{\text{cfd}} = \texttt{TripletMarginLoss}(\text{anc}, \text{pos}, \text{neg}) \quad (2)$$

where anc refers to Z^{img}, and pos and neg refer to Z^{aud}_{pos} and Z^{aud}_{neg}, respectively. Additionally, the margin is set to a default value of 1.0.

4.3 Objective of Multi2Cap

The ultimate goal of Multi2Cap is to effectively combine audio encoding, Cross-modal Feature Distillation(CFD), and text decoding to generate high-quality captions about audio content. In the pre-training phase, Multi2Cap optimizes the following objective function:

$$\mathcal{L}_{total} = \mathcal{L}_{cap} + \lambda \mathcal{L}_{cfd} \quad (3)$$

where, $\mathcal{L}_{cap}$ refers to the cross-entropy loss between the ground truth caption and the model-generated caption, ensuring that the generated captions are accurate and contextually appropriate. On the other hand, $\mathcal{L}_{cfd}$ ensures that the CFD module effectively learns the interaction between the image encoder and the audio encoder. where, λ is set to a default value of 0.5. In Section 5, we analyze the effect of λ on model performance and compare different types of loss functions for $\mathcal{L}_{cfd}$ in the CFD module.

5 Experiments

5.1 Experimental Setup

Datasets. the AAC model is evaluated using two datasets. First, Clotho(Drossos et al., 2020) includes approximately 6,000 audio clips, each ranging from 15 to 30 seconds in length, and five captions are provided for each clip. AudioCaps(Kim et al., 2019) consists of approximately 50,000 audio clips, each 10 seconds long, with one caption provided in the training set and five captions in the validation and test sets. The primary evaluation dataset is the Clotho dataset, while the AudioCaps dataset is additionally included in the experiments to enhance reliability.

Performance metrics. In this study, we use various metrics, including BLEU(BL-1–4, Papineni et al., 2002), ROUGE-L(RG-L, Lin, 2004), METEOR(ME, Denkowski and Lavie, 2014), CIDEr(CD, Vedantam et al., 2015), SPICE(SP, Anderson et al., 2016), SPIDEr(SD, Liu et al., 2017), SPIDEr-FL(SD-F, Labbe et al., 2022), Sentence-BERT(SB, Reimers, 2019), and FENSE(FS, Zhou et al., 2022) to evaluate model performance. BLEU, ROUGE-L, and METEOR assess lexical similarity based on N-grams, while CIDEr measures lexical

	ME	CD	SP	SD	SD-F
ASR Whisper (2023)	17.2	41.4	12.3	26.9	26.7
ConvNeXt (2023)	19.3	48.6	14.2	31.4	31.4
BEATs (2024)	19.5	50.5	**14.9**	32.7	32.7
LOAE (2024)	19.7	51.3	14.7	33.0	33.0
EnCLAP++ (2024b)	**19.9**	48.0	14.8	31.4	31.4
Ours					
Multi2Cap$_{small}$	19.2	48.4	13.9	31.1	31.1
Multi2Cap$_{base}$	19.7	**53.0**	14.4	**33.7**	**33.6**

a) Clotho

	ME	CD	SP	SD	SD-F
Human (2019)	28.8	91.3	21.6	-	-
EnCLAP (2024c)	25.5	80.3	18.8	49.5	-
LOAE (2024)	26.7	81.6	19.3	50.5	50.4
AutoCap (2024)	25.3	83.2	18.2	50.7	-
EnCLAP++ (2024a)	**26.9**	82.3	**19.7**	**51.0**	-
Ours					
Multi2Cap$_{small}$	25.0	79.3	17.2	48.2	48.1
Multi2Cap$_{base}$	25.9	**83.7**	18.2	50.9	**50.9**

b) AudioCaps

Table 3: **Performance Comparison on Clotho and AudioCaps**: This table shows the comparison of the fine-tuning results of pre-trained Multi2Cap on each AAC benchmark dataset with the performance of previous studies. It can be seen that Multi2Cap achieved state-of-the-art performance in most metrics. Best performance for each metric is in **Bold**, and the second-best is Underlined.

similarity using TF-IDF weighting with reference sentences. SPICE evaluates semantic similarity by considering objects, relationships, and attributes. SPIDEr balances lexical and semantic evaluation by averaging CIDEr and SPICE, and SPIDEr-FL and FENSE further assess fluency and grammatical correctness. Sentence-BERT measures semantic similarity through cosine similarity between sentence embeddings, providing a comprehensive analysis of model performance.

Audio Contents Augmentation. in this study, four audio augmentation techniques are used to compensate for the insufficient data compared to WavCaps(Mei et al., 2024). The techniques employed are *Adding white noise*, *Shifting*, *Stretching*, and *Flipping*. Each technique dynamically adjusts its application and order during training to maximize diversity. The operation of each technique is described in Appendix A.2.

Implementation Details. in this study, we use two versions of our model: Multi2Cap$_{small}$ and Multi2Cap$_{base}$. Both model uses CED as the audio encoder, Llama2-7B as the text decoder. Then, ResNet is used as the Multi2Cap$_{small}$ and ViT-Large as the Multi2Cap$_{base}$ for the image encoder. The models were trained using the AdamW optimizer(Loshchilov, 2017), with a learning rate of 5e-5 for the pre-training phase and 1e-4 for the fine-tuning phase. In pre-training, a batch size of 320 was used with 15 epochs and 2 warm-up epochs, while in fine-tuning, a batch size of 384 was used, and training was conducted for 30 epochs. Additional implementation details are provided in Appendix A.1.

5.2 Overall Performance Comparison

5.2.1 Performance of VggCaps

The pre-training performance of the proposed method is compared with previous research(Liu et al., 2024) and the human performance conducted by ourselves. The results are presented in table 2. The most noticeable finding is that both humans and AI models show improved performance when the Image Modality is added. Additionally, Human evaluators performed well on relatively simple n-gram-based performance metrics, indicating that the VggCaps dataset is clearly and naturally structured for humans to generate captions. On the other hand, Multi2Cap demonstrated outstanding performance on more complex metrics based on semantic relationships, such as CIDEr (CD), SPIDEr-FL (SD-F), Sentence-BERT (SB), and FENSE (FS). This is the result of effectively learning the potential representations between audio and images through Cross-modal Feature Distillation (CFD), suggesting that the interaction between the two modalities played a crucial role in caption generation. In particular, the significant improvement in Multi2Cap's performance over previous research in CIDEr and SPIDEr (SD) demonstrates that the addition of image modality played an important role in enhancing the semantic performance of audio captioning. In conclusion, it can be confirmed that the VggCaps dataset is well-constructed for caption generation, and that the Multi2Cap framework has significantly contributed to improving semantic performance in audio captioning tasks.

5.2.2 Performance of Benchmark

The proposed method is compared with previous top models on a per-dataset basis. The comparison results are presented in table 3. In the case of the Clotho dataset (table 3a), despite the absolute shortage of pre-training data compared to studies with similar settings(Liu et al., 2024), it shows superior performance across most metrics. Additionally, for the AudioCaps dataset (table 3b), it demonstrates performance that is either similar to or better than

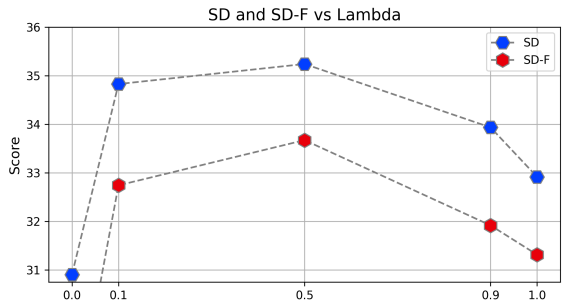


Figure 6: **Comparisons with Different λ in CFD**

	Smooth L1	MSE	InfoNCE	Cosine	Triplet
ME	12.6	13.8	13.7	13.8	**13.9**
CD	50.5	51.8	52.7	52.8	**55.0**
SP	14.3	15.3	15.2	15.3	**15.5**
SD	32.4	33.6	34.0	34.1	**35.2**
SD-F	30.2	31.6	31.9	31.8	**33.7**
SB	62.8	62.8	63.0	63.1	**63.5**
FS	58.0	58.6	58.8	58.9	**59.7**

Table 4: **Comparisons CFD by Objective Function Type**

previous models. In particular, when compared to previous research(Kim et al., 2024a), it shows improved performance in CIDEr(CD).

These results indicate that Multi2Cap effectively captures and utilizes the potential interactions between the audio and visual modalities that previous models could not. The impact of introducing a multi-modal perspective is particularly evident in specific metrics like CIDEr(CD), enabling the generation of more detailed and contextually rich audio captions. Through these results, we propose the potential for further development of multi-modal learning frameworks for audio captioning in the future.

5.3 Ablation Study

5.3.1 Comparisons $\mathcal{L}_{cfd}$ by λ difference

The pre-training performance according to λ in the objective function $\mathcal{L}_{cfd}$ is compared. The comparison results are presented in figure 6. The results are based on the SPIDEr(SD) and SPIDEr-FL(SD-F) performance of the model, with λ varying from 0 to 1 during training. Other fixed settings include $\mathcal{L}_{cfd}$ based on `TripletMarginLoss`$(\cdot)$, with experiments conducted on Multi2Cap$_{\text{base}}$. The baseline shows significant performance improvement with a relatively small weight value (0.1) at $\lambda = 0$. This demonstrates that the addition of another modality can benefit AAC tasks, as is the case in other tasks. However, when the weight is set to the maximum value (1.0), the learning effect is diluted, and the influence diminishes.

5.3.2 Comparisons $\mathcal{L}_{cfd}$ by objective function

The results based on the type of objective function for $\mathcal{L}_{cfd}$ are provided in table 4. With fixed settings, $\lambda = 0.5$ in the Multi2Cap$_{\text{base}}$ model. In the case of Smooth L1(Huber, 1992) and MSE(Bishop and Nasrabadi, 2006), which are regression-based loss functions, they are very simple functions aimed at minimizing the difference between features. As a result, they show relatively low performance improvement. On the other hand, in the case of contrastive learning-based methods like InfoNCE(Oord et al., 2018), Cosine Similarity(Radford et al., 2021), and Triplet(Schroff et al., 2015), they seem to capture potential interactions more effectively, as they can extract additional information from the comparison target.

6 Conclusion

In this study, we propose the Multi2Cap framework, which applies Cross-modal Feature Distillation (CFD) to overcome the limitations of Automated Audio Captioning (AAC) tasks that have relied on a single modality. We confirmed that Multi2Cap showed superior performance compared to existing models by expanding AAC into a multi-modal approach. This suggests that the multi-modal approach is a key factor in improving performance in AAC tasks. Overall, Multi2Cap experimentally demonstrated that expanding multi-modal learning can produce richer and more accurate results in AAC tasks. Additionally, to address the limitations of short and concise captions in existing datasets, we constructed a new VggCaps dataset. By using LLM as a guide and supplementing audio information with visual information, this dataset provides contextually rich captions with longer and more diverse descriptions. Through various analyses, we confirmed that it outperforms existing datasets in terms of lexical diversity and complexity of expression, and we validated its robustness and validity through human evaluation. Overall, VggCaps presents a higher level of challenge in AAC tasks.

In conclusion, Multi2Cap effectively leverages multi-modal learning to propose a new paradigm for AAC research, highlighting the importance of a multi-modal approach in future AAC tasks. This study demonstrates the potential to build more sophisticated and rich AI systems by effectively utilizing multi-modal data beyond a single modality, marking an important milestone in the future development of multi-modal AI research.

7 Limitations

While Multi2Cap enhances audio captioning performance by utilizing visual features, several limitations exist. First, the visual context derived from static image frames in the video clips of the VggSound dataset may not adequately capture the complex interactions between visual and auditory elements. For instance, in dynamic and evolving scenes, relying solely on a single static frame may lead to visual information that either does not align with the audio or provides insufficient context, potentially resulting in inaccurate captions. This limitation is particularly pronounced in scenarios requiring a more nuanced and multi-layered understanding of the interaction between audio and visual elements, ultimately impacting caption accuracy.

Furthermore, enabling the video modality to leverage visual information more effectively could dilute the primary goal of Automated Audio Captioning (AAC). The essence of AAC is to generate meaningful descriptions with audio as the central medium. Visual information should remain auxiliary to the audio, and over-reliance on visual cues risks undermining the primary purpose of AAC research. Therefore, in cases where complex interactions between visual and auditory elements are essential, careful validation is required to ensure that visual inputs are used only as supplementary information when appropriate.

Lastly, while Multi2Cap has demonstrated the ability to generate longer captions, this can lead to overly detailed descriptions for simple audio events. Excessively verbose captions may introduce unnecessary information in real-world applications, creating confusion in the model's learning process. Particularly for simple audio events, generating complex descriptions can cause the model to miss essential information, negatively impacting the practicality and efficiency of AAC systems. As such, future research should focus on distinguishing between simple and complex audio inputs and generating appropriate captions accordingly.

8 Risks and Ethics

Using LLMs and Cross-modal Feature Distillation(CFD) techniques can potentially raise ethical concerns, especially when the dataset originates from biased sources. There is a risk of generating biased or inaccurate captions, specifically due to the possibility that biases inherent in the image encoder could be distilled into the audio encoder and, ultimately, reflected in the LLM text decoder. This is particularly problematic for sensitive or subjective content, where stereotypes or incorrect representations may be amplified. Furthermore, in real-world applications of AAC technology, users are likely to trust and accept the generated captions as factual. If the generated captions contain incorrect information or biased perspectives, there is a risk of instilling biased viewpoints in users. These biased outputs could then be further processed by users and potentially re-enter the learning pipeline, circulating as social network data and contributing to future training datasets. Therefore, when interpreting the results of this research and deploying AAC-based services, it is essential to recognize these potential risks and challenges and to approach them with caution.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

J Benesty, J Chen, and Y Huang. 2008. Automatic speech recognition: A deep learning approach.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. 2024. Ced: Consistent ensemble distillation for audio tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 291–295. IEEE.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. 2017. Automated audio captioning with recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 374–378. IEEE.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Ronen Eldan and Yi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, Sergey Tulyakov, and Vicente Ordonez. 2024. Taming data and transformers for audio generation. *arXiv preprint arXiv:2406.19388*.

Awni Hannun et al. 2014. Deep speech: Scaling up end-to-end speech recognition. In *Proceedings of the 31st International Conference on Machine Learning*, pages 382–390.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE.

T. Heittola et al. 2021. Sound event detection and context recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Gustav Herdan. 1960. *Quantitative linguistics*. Butterworths.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Nicholas P. Holmes, Gemma A. Calvert, and Charles Spence. 2024. *Multimodal Integration*. Springer.

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*.

Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. 2024. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015.

Dino Ienco and Cassio Fraga Dantas. 2024. Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning. *arXiv preprint arXiv:2408.07080*.

Marek Kadlčík, Adam Hájek, Jürgen Kieslich, and Radosław Winiecki. 2023. A whisper transformer for audio captioning trained with synthetic captions and transfer learning. *arXiv preprint arXiv:2305.09690*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Jaeyeon Kim, Minjeon Jeon, Jaeyoon Jung, Sang Hoon Woo, and Jinjoo Lee. 2024a. Enclap++: Analyzing the enclap framework for optimizing automated audio captioning performance. *arXiv preprint arXiv:2409.01201*.

Jaeyeon Kim, Jaeyoon Jung, Minjeong Jeon, Sang Hoon Woo, and Jinjoo Lee. 2024b. Expanding on enclap with auxiliary retrieval model for automated audio captioning. Technical Report 108, DCASE2024 Challenge. Ranked 2/12 in the DCASE2024 Challenge Task 6 with FENSE score of 0.544.

Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. 2024c. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6735–6739. IEEE.

Emilie Labbe, Thomas Pellegrini, and Julien Pinquier. 2022. Spider-fl: An extension of spider for evaluating fluency and linguistic diversity.

Etienne Labbé, Thomas Pellegrini, and Julien Pinquier. 2023. Irit-ups dcase 2023 audio captioning and retrieval system. In *Proc. Conf. Detection Classification Acoust. Scenes Events Challenge*, pages 1–5.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. 2024a. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*.

Lin Li et al. 2024b. Generalist multimodal ai: A review of architectures, challenges, and opportunities. *arXiv preprint arXiv:2406.05496*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jizhong Liu, Gang Li, Junbo Zhang, Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Yujun Wang, and Bin Wang. 2024. Enhancing automated audio captioning via large language models with optimized audio encoding. *arXiv preprint arXiv:2406.13275*.

Peng Liu et al. 2022. Multi-target knowledge distillation via student self-reflection. *International Journal of Computer Vision*, 130:2159–2175.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Viorica Marian, Sayuri Hayakawa, and Scott R. Schroeder. 2021. Cross-modal interaction between auditory and visual input impacts memory retrieval. *Frontiers in Neuroscience*, 15.

Xinhao Mei, Xubo Liu, Mark D Plumbley, and Wenwu Wang. 2022. Automated audio captioning: An overview of recent progress and new challenges. *EURASIP journal on audio, speech, and music processing*, 2022(1):26.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Chaitanya Narisetty, Emiru Tsunoo, Xuankai Chang, Yosuke Kashiwagi, Michael Hentschel, and Shinji Watanabe. 2022. Joint speech recognition and audio captioning. *Preprint*, arXiv:2202.01405.

Wenqi Niu, Yingchao Wang, et al. 2024. Efficient and robust knowledge distillation from a stronger teacher based on correlation matching. *arXiv preprint arXiv:2410.06561*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Zhongyi Sang, Kotaro Funakoshi, and Manabu Okumura. 2024. Contrastive knowledge distillation for robust multimodal sentiment analysis. *arXiv preprint arXiv:2410.08692*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Arda Senocak, Tae-Hyun Oh, Jean-Charles Kim, et al. 2021. Localizing visual sounds the hard way. *arXiv preprint arXiv:2104.02691*.

Sakib Shahriar et al. 2024. Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *arXiv preprint arXiv:2407.09519*.

EH Simpson. 1949. Measurement of diversity. *Nature*, 163.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22(2):213–227.

Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yu Wang, Ya Zhang, Jinxiang Liu, et al. 2023. A unified audio-visual learning framework for localization, separation, and recognition. *arXiv preprint arXiv:2305.19458*.

Shih-Lun Wu, Xuankai Chang, Gordon Wichern, Jee-weon Jung, François Germain, Jonathan Le Roux, and Shinji Watanabe. 2024. Improving audio captioning models with fine-grained audio features, text

embedding supervision, and llm mix-up augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320. IEEE.

Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kai Yu. 2023. Beyond the status quo: A contemporary survey of advances and challenges in audio captioning. *Preprint*, arXiv:2205.05357.

Han Yu et al. 2023. Aligning in a compact space: Contrastive knowledge distillation between heterogeneous architectures. In *Proceedings of the 2023 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

C Udny Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.

Yi Zhang et al. 2024. Voiceldm: Text-to-speech with environmental context. *arXiv preprint arXiv:2309.13664*.

Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. 2022. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 981–985. IEEE.

A Appendix

A.1 Additional Details

A.1.1 Pre-training Implementation Details

For reproducibility, the implementation details used in pre-training are presented in table 5. Based on Multi2Cap$_{\text{base}}$, the AdamW optimizer is used, with the base learning rate set to 5×10^{-5} and the weight decay set to 1×10^{-6} to prevent overfitting. The batch size is 320, and training is conducted for a total of 15 epochs, with the first 2 epochs set as a warm-up phase to stabilize initial training. The β parameters of the Adam optimizer are set to $(0.9, 0.999)$. The sampling rate of the audio input is fixed at 16,000Hz, and four audio augmentation techniques—*AddWhiteNoise*, *Shifting*, *Stretching*, and *Flipping*—are applied. The visual information is processed using ResNet50 and ViT-Large image encoders, with the image resolution set to 224×224 pixels. Additionally, the *RandomResizedCrop* technique is used for image augmentation.

Hyper-parameters	**Value**
Optimizer	AdamW
Base learning rate	5×10^{-5}
Weight decay	1×10^{-6}
Adam β	(0.9, 0.999)
Batch size	320
Training epochs	15
Warmup epochs	2
Audio sample rate	16000
Audio Augmentation	AddWhiteNoise
	Shifting
	Stretching
	Flipping
Image Teacher	ResNet50, ViT-Large
Image resolution	224×224
Image augmentation	RandomResizedCrop

Table 5: Default Pre-training Setting

A.1.2 Fine-tuning Implementation Details

The implementation details for the fine-tuning phase of the Multi2Cap model on benchmark datasets are presented in table 6. Based on Multi2Cap$_{\text{base}}$, the AdamW optimizer is used. The base learning rate is set to 1×10^{-4}, and the weight decay is set to 1×10^{-6}. The β parameters of the Adam optimizer are specified as $(0.9, 0.999)$. The batch size is 384, and training is conducted for a total of 30 epochs. Of these, the first 2 epochs are

used as a warm-up phase to stabilize the model during the initial stages of training. The audio input is processed at a sampling rate of 16,000Hz, and four audio augmentation techniques—*AddWhiteNoise*, *Shifting*, *Stretching*, and *Flipping*—are applied.

Hyper-parameters	Value
Optimizer	AdamW
Base learning rate	1×10^{-4}
Weight decay	1×10^{-6}
Adam β	(0.9, 0.999)
Batch size	384
Training epochs	30
Warmup epochs	2
Audio sample rate	16,000
Audio Augmentation	AddWhiteNoise Shifting Stretching Flipping

Table 6: Default fine-tuning setting

A.2 Additional Experiments

A.2.1 Overall Comparisons by λ Difference

The results of the pre-training performance for each λ of $\mathcal{L}_{cfd}$, as presented in Figure 6, are summarized based on all metrics in table 7. When $\lambda = 0.5$, the best performance was observed across all metrics. Specifically, a score of 55.0 in CIDEr(CD), 15.5 in SPICE(SP), and 35.2 in SPIDEr(SD) was recorded, which is significantly higher compared to other λ values. Additionally, the highest performance was also achieved in the SD-F and Sentence-BERT(SB) metrics with scores of 33.7 and 63.5, respectively, while the FENSE(FS) also showed excellent results with a score of 59.7. Therefore, it can be confirmed that $\lambda = 0.5$ is the most suitable value for deriving the optimal performance of the Multi2Cap model.

λ	ME	CD	SP	SD	SD-F	SB	FS
0	12.4	47.6	14.2	30.9	27.6	61.8	55.3
0.1	13.7	54.4	15.3	34.8	32.7	63.1	59.1
0.5	**13.9**	**55.0**	**15.5**	**35.2**	**33.7**	**63.5**	**59.7**
0.9	13.8	52.4	15.4	33.9	31.9	63.0	59.1
1	13.7	50.6	15.2	32.9	31.3	63.0	59.9

Table 7: Performance comparison by lambda (all metrics)

A.2.2 Comparison About Audio Content Augmentation

Additionally, the results based on whether audio content augmentation was applied are examined. The results are presented in table 8. The augmentation techniques used for pre-training in Multi2Cap are *AddWhiteNoise*, *Shifting*, *Stretching*, and *Flipping*.

Adding white noise is the simplest method, which involves adding random noise to the audio signal. This technique helps improve the model's generalization performance in environments with various noise by applying slight noise to the input data.

Shifting shifts the start point of the audio by a certain amount of time forward or backward. This contributes to increasing the model's robustness to temporal shifts in the data.

Stretching changes the playback speed of the audio while maintaining the pitch. This provides the model with generalization capabilities to handle audio data at different speeds.

Flipping inverts the phase of the audio waveform. While this results in no perceptible change to the human ear, it alters the mathematical structure of the signal, providing additional data diversity.

These four augmentation techniques are dynamically set in terms of whether to apply them during training and in what order, to maximize diversity during learning. This can be expressed in the following formula. Although augmentation generally contributes positively to performance improvement, it does not show consistent performance gains across all evaluation metrics. Specifically, in BLEU scores, the impact of augmentation on performance is minimal or nearly nonexistent, whereas clear performance improvements can be observed in metrics such as CIDEr(CD) and SPICE(SP). This suggests that audio content augmentation techniques do not significantly affect simple n-gram-based performance but have a positive effect on semantic consistency and the generation of sophisticated captions.

A.3 VggCaps

A.3.1 Prompt Templates

In Figure 1, the input-prompt and post-prompt are introduced. The input-prompt is provided in listing 1. The input-prompt requests the creation of an initial caption using both audio and images. Additionally, rules are applied to set conditions for generating the raw-caption. Furthermore, the post-

	w/o augment	w/ augment
B-1	**34.1**	**34.1**
B-2	19.1	**19.2**
B-3	11.5	**11.6**
B-4	7.5	**7.6**
RG-L	29.6	**29.9**
ME	12.7	**13.9**
CD	51.8	**55.0**
SP	14.2	**15.5**
SD	33.0	**35.2**
SD-F	30.9	**33.7**
SB	62.6	**63.5**
FS	58.2	**59.7**

Table 8: Comparison about audio content augmentation

prompt is provided in listing 2. The post-prompt modifies the raw-caption into more general and descriptive sentences. Similarly, rules are applied to impose generation conditions, and examples are provided to ensure that refined captions are generated.

A.3.2 Examples

Table 9 shows samples from the constructed VggCaps.

```
The two images are a video snapshot and an audio spectrum,
   categorized into {}. Write a caption that includes the visual and
   auditory elements that these images suggest :

[Rules]
1. Focus on audio-visual elements
2. Write with a rich vocabulary
3. Exclude anything other than captions

>>> Caption :
```

Listing 1: Input Prompt Template for VggCaps

```
Revise the sentence below to a more general and descriptive sentence,
    and one that is appropriate for captioning images and audio:

[Rules]
1. Focus on the background sound element
2. Exclude hashtags (like "#peoplemarching")
3. Avoid captions that read like film titles.(like "A Joyful Heart's
   Muffled Cough-terlude ~~~")
4. Write with a rich vocabulary
5. Exclude anything other than captions

[example]
1. A sea of determined faces marches in unison, their boots
   thundering in harmony across the ground, echoing the orchestrated
   rhythm of military discipline and resolve.
2. Amidst the whispering reeds, the pheasant's crowing breaks the
   silence of the meadow, a resonant call that echoes the wild heart
   of nature.
3. Explosive bursts light up the night, resonating with the jubilant
   cracking and booming of celebratory fireworks.


>>> Caption :
```

Listing 2: Post Prompt Template for VggCaps

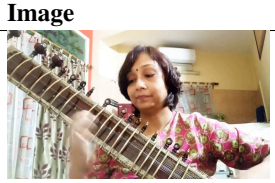
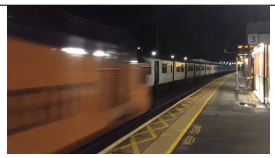
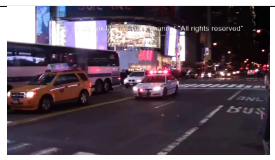
ID	Image	Caption	Category
1v5mmZoJJ50		Her fingers dance gracefully on the sitar strings, weaving a tapestry of sound that resonates with the serene and soulful essence of the music.	playing sitar
5IuRzJRrRpQ		The joyful bleats of sheep mingle with the soft rustling of grass and the occasional bark of an energetic dog, bringing a lively atmosphere to the green pastures.	sheep bleating
0fTwdhslb6E		The thunderous crack of the ball against the wall reverberates as two players immerse themselves in the rhythm of their squash game.	playing squash
1tPjBLXRHqM		The lively hum of the festival is accompanied by the drummer's rhythmic beats, their sticks creating a pulsating rhythm that resonates through the crowd.	playing drum kit
2zJiY9Mqhtc		A canopy alive with song as the harmonious tweets and chirps of birds enliven the surroundings, weaving a vibrant tapestry of nature's own symphony.	bird chirping, tweeting
3ymE2QOPRCA		The invigorating sounds of a volleyball match fill the air, blending cheers with the sharp slap of the ball.	playing volleyball
1mpFmBJ3nv0		The thunderous crescendo of a train horn slices through the stillness of the night, a wild call that reverberates along the tracks.	train horning
1t3sNHA0Vd4		In the cacophony of urban sounds, the police car's siren pierces the night air, signaling urgency and command.	police car (siren)
P0Mzdxr6F58I		In the stillness of the night, a lone frog's persistent ribbit pierces the quiet, adding a rhythm to the tranquil scene.	cattle mooing

Table 9: Examples of VggCaps