Learning to See through Sound: From VggCaps to Multi2Cap for Richer Automated Audio Captioning

Anonymous ACL submission

Abstract

Automated Audio Captioning (AAC) aims to generate natural language descriptions of audio content, enabling machines to interpret and communicate complex acoustic scenes. However, current AAC datasets often suffer from short and simplistic captions, limiting model expressiveness and semantic depth. To address this, we introduce **VggCaps**, a new multimodal dataset that pairs audio with correpsonding video and leverages large language models (LLMs) to generate rich, descriptive captions. VggCaps significantly outperforms existing benchmarks in caption length, lexical diversity, and human-rated quality. Furthermore, we propose **Multi2Cap**, a novel AAC framework that learns audio-visual representations through a AV-grounding module during pre-training and reconstructs visual semantics using audio alone at inference. This enables visually grounded captioning in audio-only scenarios. Experimental results on Clotho and AudioCaps demonstrate that Multi2Cap achieves state-of-the-art performance across multiple metrics, validating the effectiveness of cross-modal supervision and LLM-based generation in advancing AAC.

1 Introduction

Automated Audio Captioning (AAC) (Drossos et al., 2017) is a task that generates natural language descriptions of audio content, emerging as a significant challenge in artificial intelligence. Unlike Automatic Speech Recognition (ASR) (Benesty et al., 2008), which solely converts speech to text, AAC requires comprehensive understanding and description of both linguistic elements and non-verbal audio signals, including environmental sounds, animal vocalizations, and musical content. Despite being a relatively recent research direction, AAC has garnered increasing attention due to growing demands in complex audio applications, particularly in audio interaction and retrieval systems (Mei et al., 2022; Xu et al., 2024). Such demand has catalyzed continuous technological advancement in efficient processing and description of diverse audio information (Liu et al., 2024).

Despite recent progress, existing AAC systems face two core limitations. First, current datasets such as AudioCaps (Kim et al., 2019) and WavCaps (Mei et al., 2024), which contain only short, template-like caption–typically fewer than 10 words–that fail to reflect the complexity of real-world auditory scenes(Table 1). These overly concise descriptions not only lack semantic richness but also lead to increased risk of model overfitting (Eldan and Li, 2023), as the limited lexical variation constrains the diversity of training signals. Second, although AAC is inherently defined as the task of generating captions from audio alone, this poses a fundamental challenge: many acoustic scenes are inherently ambiguous without additional contextual information (Chen et al., 2021). For example, the sound of cheering could correspond to a sports event, a concert, or a public demonstration—distinctions that are difficult to resolve from audio alone but easily clarified with visual cues (Holmes et al., 2024). This observation motivates our approach: rather than modifying the AAC task to accept visual input at inference time, we propose to train the model to internalize visual semantics during training, enabling it to infer richer and more grounded descriptions from audio alone.

To address these challenges, we introduce **VggCaps**, a large-scale multi-modal audio captioning dataset. Built upon the VGGSound (Chen et al., 2020) corpus, we pairs audio segments with corresponding video frames and generates initial captions based on the audio content, which are then refined and enriched using visual context. This process leverages large language models (LLMs) to produce high-quality captions that capture both auditory and visual semantics. Compared to prior datasets, VggCaps features significantly longer captions (21.1 words on average), richer vocabulary, and higher readability complexity, encouraging the development of more expressive AAC models. Human evaluation confirms the clarity and fidelity of these captions.

Furthermore, we propose **Multi2Cap**, a novel framework that leverages visual supervision only during pre-training to enhance the semantic richness of audio representations. Specifically, Multi2Cap learns to align audio and visual features through an Audio-Visual Grounding module and recovers visual semantics from audio alone through a dedicated Visual Feature Reconstructor. This enables the model to indirectly leverage visual information during inference, resulting in higher-quality and more semantically grounded captions. We validate our approach on standard AAC benchmarks, including Clotho (Drossos et al., 2020) and AudioCaps (Kim et al., 2019), where Multi2Cap consistently outperforms existing methods across both lexical and semantic evaluation metrics. Our ablation studies further demonstrate the semantic fidelity of the reconstructed visual features and the effectiveness of grounding-based training.

Our contributions are threefold: (1) We propose a new paradigm for AAC by incorporating visual context during pre-training and reconstructing it from audio at inference time.
(2)We introduce **VggCaps**, a large-scale multimodal dataset with LLM-generated captions that are longer, more diverse, and semantically richer than existing AAC corpora.
(3)We present **Multi2Cap**, a grounding-based AAC model that achieves state-of-the-art performance on Clotho and AudioCaps, and demonstrate its ability to preserve and utilize visual semantics even in audio-only scenarios.

2 Related works

2.1 Automated Audio Captioning

Automated Audio Captioning (AAC) is a task that generates natural language descriptions for audio content, requiring a comprehensive understanding of both verbal and non-verbal acoustic events such as environmental sounds, music, or animal vocalizations. Unlike Automatic Speech Recognition (ASR) (Benesty et al., 2008), AAC involves a deeper semantic interpretation of auditory scenes (Narisetty et al., 2022). Due to these characteristics, AAC introduces unique challenges in capturing complex acoustic contexts and abstract concepts. AAC architectures have evolved from early CNN-RNN hybrids (Drossos et al., 2017; Mei et al., 2022) to transformer-based models (Xu et al., 2024). The integration of Large Language Models (LLMs) marked a significant advancement in caption generation quality (Bommasani et al., 2021). However, current approaches remain constrained to audio-text modalities (Liu et al., 2024; Kim et al., 2024a), prompting our investigation into multi-modal AAC frameworks.

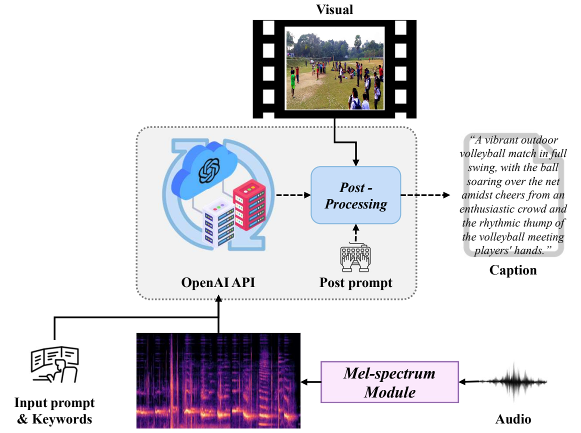

Visual
Post - Processing
OpenAI API
Post prompt
"A vibrant outdoor volleyball match in full swing, with the ball soaring over the net amidst cheers from an enthusiastic crowd and the rhythmic thump of the volleyball meeting players' hands."
Caption
Input prompt & Keywords
Mel-spectrum Module
Audio

Figure 1: **Pipeline of VggCaps Data Processing.** Each data sample consists of an video frames and a corresponding 10-second audio segment, which are processed into a Mel-spectrogram and combined as input to GPT-4o. The generated captions undergo post-processing for refinement and clarity.

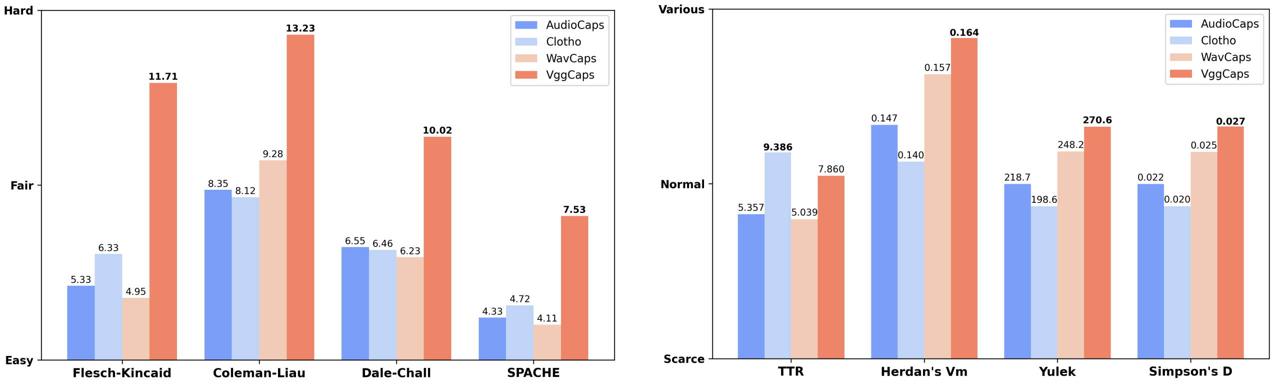
2.2 Audio-Visual Representation Learning

In recent years, the field of multimodal learning has made notable progress in audio-visual representation learning, aiming to integrate information from both audio and visual modalities for more informative representations. Prior research largely follows two main approaches. The first learns a shared embedding space to directly model semantic relationships across modalities (Radford et al., 2021; Jia et al., 2021; Guzhov et al., 2021), aligning representations for strong downstream performance. The second employs cross-attention to capture contextual interactions between audio and visual inputs (Jaegle et al., 2022; Nagrani et al., 2022; Shi et al., 2022), enabling dynamic cross-modal dependency modeling.

Building on these approaches, our research introduces a method that combines these two strategies, employing a Grounding Token mechanism and a reconstruction process where visual information is indirectly represented through audio. This design

Dataset	num. of row	num. of audio	avg(std). audio length(s)	num. of caption	avg(std). caption length	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** VggCaps shows higher linguistic complexity and vocabulary diversity than prior AAC datasets, demonstrating its potential to support richer and more expressive audio captions.

allows the model to make use of visual context even in downstream tasks, providing a flexible and efficient approach to audio-visual representation learning.

3 Proposed Dataset: VggCaps

We present VggCaps, a novel dataset for multi-modal audio captioning research. This section details the dataset construction methodology, analysis metrics, and human evaluation protocols.

VggCaps builds upon VggSound (Chen et al., 2020), a large-scale audio-visual dataset originally designed for sound event classification in videos. While VggSound has been widely adopted in audio-visual research (Senocak et al., 2021; Wang et al., 2023), it lacks descriptive captions for its audio content and does not explicitly account for the semantic interplay between auditory and visual signals. Moreover, existing audio captioning datasets often feature short, context-agnostic descriptions that fail to capture the complexity of real-world scenes. To address these limitations, we introduce **VggCaps**, a multi-modal dataset that provides rich, semantically grounded captions aligned with both audio and visual content.

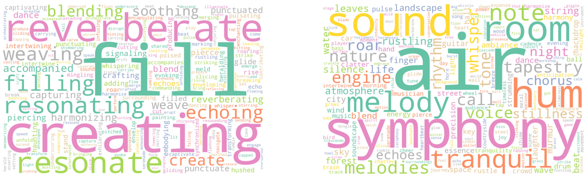
3.1 Data Processing

Our core objective is to enable more expressive and context-aware audio captioning by incorporating visual cues during caption generation. To this end, we utilize large language models (LLMs) as a practical tool to generate high-quality captions that reflect both modalities as illustrated in figure 1. For each data point, we extract a representative video frames and its corresponding 10-second audio segment. The audio is converted to a Mel-spectrogram (Hannun et al., 2014) and paired with the image frame as input to GPT-4o (Shahriar et al., 2024). Additionally, we provide the input with a prompt to generate suitable captions. Generated captions undergo post-processing to eliminate redundant expressions and enhance linguistic clarity.

In this study, we obtained a total of 173,494 VggCaps data samples. This dataset is used as the pre-training 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. VggCaps significantly differs from existing AAC datasets in two key aspects: caption length and



a) VERB b) NOUN

Figure 3: **Wordcloud in VggCaps**

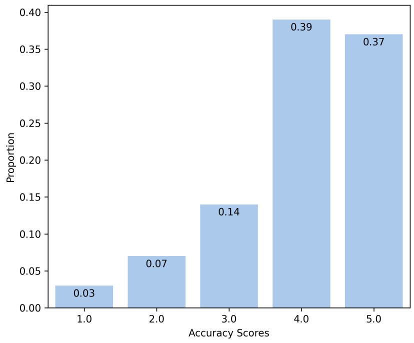


Figure 4: **Mean Opinion Score (MOS)**: This table shows the distribution of MOS for VggCaps calculated through human evaluation. It indicates that the vast majority of samples have appropriate captions.

modality. Our captions are approximately twice the length of current benchmarks, enabled by LLM-guided generation. Additionally, VggCaps incorporates corresponding visual information, facilitating multi-modal AAC research.

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 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) 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

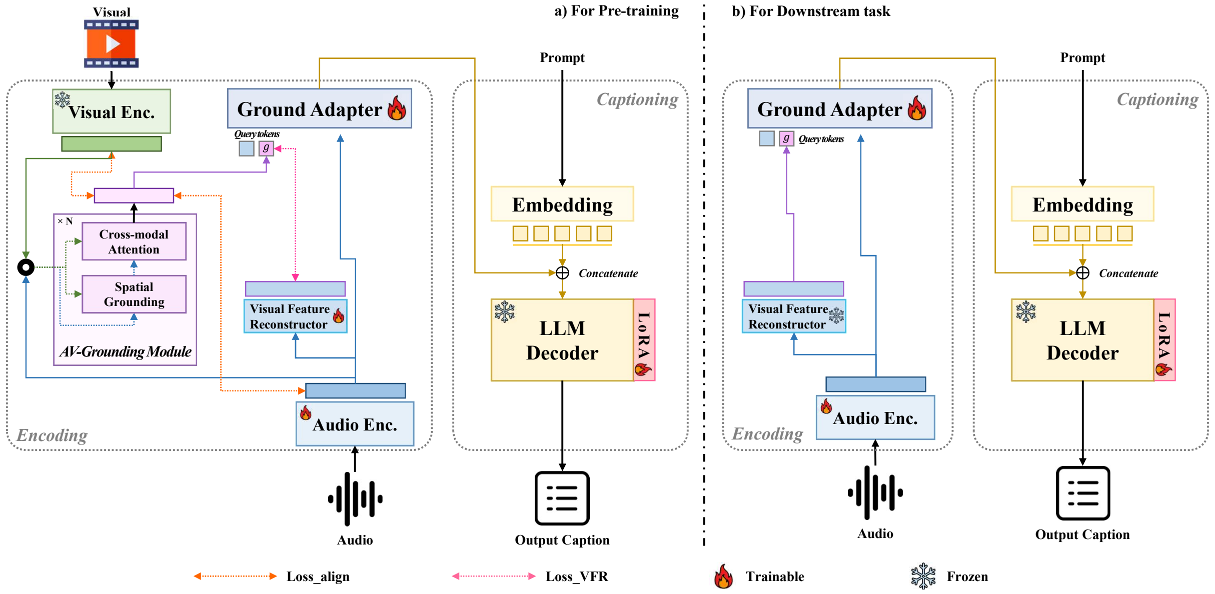


Figure 5: **Overview of the Multi2Cap architecture with Audio-Visual Grounding and Visual Feature Reconstruction.** (a) During pre-training, audio and visual inputs are fused via the AV-Grounding Module to produce a compressed representation g, which is passed to a trainable Ground Adapter. Simultaneously, a Visual Feature Reconstructor (VFR) learns to reconstruct g from audio alone. Both the audio encoder and LLM decoder are optimized using LoRA. (b) During downstream task, only the audio is provided. The VFR reconstructs $\hat{g}$, which is used by the Ground Adapter to generate grounded representations, enabling the LLM decoder to produce captions with visual grounding, even without image input.

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 Proposed Framework: Multi2Cap

This section provides a detailed description of the architecture and training procedure of the proposed Multi2Cap model. Multi2Cap is designed to generate descriptive captions from audio in scenarios where visual information is available during pre-training. It leverages visual cues to learn rich audio-visual representations in the pre-training phase, while being structured to effectively utilize the learned representations even in downstream tasks where visual inputs are not provided. A detailed illustration of the overall workflow is presented in the accompanying figure. 5.

4.1 Creating Caption

Multi2Cap is trained to convert audio inputs into textual captions by minimizing the cross-entropy loss with respect to the ground-truth captions. Given an audio input A, it is first encoded into a feature representation through the Audio Encoder. The resulting representation is then passed through a Ground Adapter and fed into the LLM Decoder. Based on the encoded input, the decoder generates a natural language caption, and the model is optimized by minimizing the cross-entropy loss L_{cap} between the generated caption and the ground-truth reference:

$$\mathcal{L}_{\text{cap}} = -\sum_{t=1}^{T} \log p(y_t | y_{<t}, A; \theta) \qquad (1)$$

where, y_t denotes the t-th word, and θ represents the trainable parameters of the model.

4.2 Audio-Visual Grounding

The Audio-Visual Grounding Module proposed in this work fuses audio and visual information into a single compact token, denoted as g. In this process, visual input (V) is first encoded through a visual encoder and then spatially grounded with the audio input (A). Cross-attention is applied between A and V to emphasize the visual characteristics embedded within the audio. The resulting representations are aggregated via average pooling to form the compact grounding token g, which densely encapsulates the joint audio-visual information.

To further enhance alignment between audio and visual representations during training, an additional

Method	LM	BL	RG-L	ME	CD	SP	SD	SD-F	SB	FS
Pre-training w/o visual										
Human	-	46.1	29.6	13.5	15.4	7.5	11.4	11.4	48.6	48.6
Whisper (2023)	-	29.3	21.9	9.7	32.1	9.3	20.7	19.2	51.4	49.3
AutoCap (2024)	-	31.3	27.2	10.9	46.7	12.9	29.8	27.3	59.2	54.3
EnCLAP++ (2024a)	BART	33.8	29.2	13.1	48.4	14.8	31.6	28.2	62.3	56.3
LOAE (2024)	LLaMA2-7B	33.0	29.1	12.4	47.6	14.2	30.9	27.6	61.8	55.3
Pre-training w/ visual										
Human		49.2	**32.9**	14.8	19.4	7.5	13.4	13.4	48.9	48.9
Multi2Cap(Ours)	LLaMA2-7B	35.8	29.6	14.2	52.5	15.4	34.0	**33.7**	63.3	57.5
	LLaMA3.1-8B	34.5	29.4	15.0	52.8	15.1	33.9	33.1	63.1	56.6
	LLaMA3.2-3B	34.9	29.6	15.1	53.7	15.1	34.4	33.3	63.5	57.6
	Mistral-7B	36.5	29.1	15.3	53.9	15.3	34.6	32.9	64.0	**59.6**
	Qwen2.5-3B	34.4	29.2	15.1	53.1	15.2	34.1	32.8	64.5	59.5
	Qwen2.5-7B	**36.9**	29.5	15.1	52.7	15.8	34.3	33.2	64.4	58.5
	DeepSeek-R1-1.5B	34.9	29.6	15.3	53.9	15.2	34.6	32.6	64.9	57.7
	DeepSeek-R1-7B	34.3	29.1	**15.4**	**54.2**	**15.9**	**35.1**	33.3	**65.4**	58.9

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. The performance shows superior results across all metrics. Best performance for each metric is in **Bold**, and the second-best is Underlined.

loss term L_{align} is introduced. This objective promotes semantic consistency across modalities and encourages the grounding token to effectively capture multimodal context:

$$\mathcal{L}_{\text{align}} = \frac{\alpha}{2K} \sum_{k=1}^{K} (\text{CE}(g_k, A_{\text{mean}}) + \text{CE}(g_k, V_{\text{cls}})) \quad (2)$$

where, CE denotes the cross-entropy loss, g_k refers to the k-th audio-visual grounding token, A_{mean} is the mean-pooled audio representation, and V_{cls} is the [CLS] token derived from the visual encoder. We empirically set $K = 4$ based on optimal performance observed during experimentation.

4.3 Visual Feature Reconstructor

Since visual inputs are not available in downstream tasks, it is necessary to compress and store visual information into the grounding token during pre-training and reconstruct it later. To achieve this, we introduce a Visual Feature Reconstructor (VFR), denoted as $\psi(A)$, which is an MLP-based module designed to infer visual representations from audio alone.

The VFR is trained by minimizing the mean squared error (MSE) loss between the reconstructed representation $\hat{g} = \psi(A)$ and the original grounding token g generated from actual visual inputs:

$$\mathcal{L}_{\text{vfr}} = \|g - \hat{g}\|^2 \quad (3)$$

This allows the model to recover semantically meaningful visual context solely from audio, enabling effective representation learning even in the absence of images during downstream inference.

4.4 Objective of Multi2Cap

The final Multi2Cap model is trained using the following combined loss function:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{cap}} + \alpha\mathcal{L}_{\text{align}} + \beta\mathcal{L}_{\text{vfr}}, & \text{in pre-training} \\ \mathcal{L}_{\text{cap}} & \text{otherwise} \end{cases} \quad (4)$$

where, α and β are hyperparameters that control the relative importance of each loss term. In this study, we empirically set $\alpha = 0.02$ and $\beta = 0.05$ based on optimal performance observed during experimentation.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate our model on two standard AAC benchmarks: Clotho(Drossos et al., 2020) and AudioCaps(Kim et al., 2019). Clotho comprises 6,000 audio clips (15-30 seconds) with five captions per clip, while AudioCaps contains 50,000 clips (10 seconds) with one caption for training and five for validation/testing. Clotho serves as our primary benchmark, with AudioCaps providing additional validation.

Evaluation 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)

Method		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	**LLMs**					
Multi2Cap	LLaMA3.1-8B	20.7	52.1	15.0	33.6	33.5
	Mistral-7B	19.6	**53.0**	14.2	33.6	33.6
	Qwen2.5-7B	19.9	51.7	14.9	33.3	33.3
	DeepSeek-R1-7B	**20.8**	52.5	**15.3**	**33.9**	**33.9**

a) Clotho

Method		ME	CD	SP	SD	SD-F
Human		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	**LLMs**					
Multi2Cap	LLaMA3.1-8B	28.6	83.2	20.4	51.8	51.7
	Mistral-7B	27.5	82.9	19.7	51.3	51.2
	Qwen2.5-7B	27.8	82.7	19.5	51.1	51.0
	DeepSeek-R1-7B	**29.0**	**83.6**	**20.8**	**52.2**	**52.2**

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.

(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 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.

Implementation Details. In this study, we evaluate the performance of Multi2Cap using a variety of backbone networks. For the audio encoder, we adopt CED (Dinkel et al., 2024); for the visual encoder, we utilize CLIP-ViT-Large (Radford et al., 2021); and for the text decoder, we experiment with relatively lightweight LLMs, including LLaMA (Touvron et al., 2023), Qwen (Qwen et al., 2025), and DeepSeek (DeepSeek-AI et al., 2025). 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 prior studies and human performance assessed in our own evaluation. The results are summarized in Table 2. A key observation is that the inclusion of image modality consistently improves performance for both humans and AI models. Notably, human evaluators exhibited strong performance on relatively simple n-gram-based metrics (e.g., BLEU, ROUGE-L), indicating that the VggCaps dataset is well-structured and intuitively understandable for human caption generation.

In contrast, our proposed Multi2Cap framework achieves superior performance on semantically oriented metrics, including CIDEr, SPIDEr-FL, Sentence-BERT, and FENSE. These improvements are attributed to the introduction of the Audio-Visual Grounding Module and Visual Feature Reconstructor (VFR). During pre-training, the model encodes visual context into a compact grounding token g and learns to reconstruct it from audio alone, enabling Multi2Cap to retain visual semantics and generate contextually coherent captions even in audio-only downstream scenarios. In summary, the VggCaps dataset provides a robust foundation for training multimodal models, and the Multi2Cap framework, through its novel grounding-based architecture, significantly advances the state of semantic audio captioning.

5.2.2 Performance of Benchmark

We compare our proposed method with state-of-the-art baselines across two standard AAC benchmarks, as shown in Table 3. On the Clotho dataset (Table 3a), despite using comparatively less pre-training data than prior methods, Multi2Cap consistently outperforms existing approaches across most evaluation metrics. On the AudioCaps dataset (Table 3b), our model achieves competitive or superior results, particularly excelling in CIDEr, SPIDEr

Token	R@1	R@5	CLIP Score	Attn-Entropy
g	89.5	99.7	44.5	0.38
$\hat{g}$	**71.8**	**91.3**	**38.3**	**0.44**
audio	1.2	7.3	16.8	1.62
random	0	0.3	8.3	6.14

Table 4: **Retrieval and semantic alignment performance** of reconstructed AV-Ground Token $\hat{g}$ compared to original g and baseline embeddings.

and SPIDEr-FL scores.

These improvements can be attributed to our redesigned architecture, which effectively encodes and reconstructs visual context through grounding, even when visual inputs are absent during downstream inference. Unlike previous methods that rely solely on audio-text alignment, Multi2Cap learns semantically enriched representations by leveraging visual supervision during pre-training, allowing it to generate more descriptive, coherent, and context-aware captions.

5.3 Ablation Study

5.3.1 Semantic Fidelity of Reconstructed g

This case study investigates whether the Visual Feature Reconstructor (VFR) in Multi2Cap can effectively reconstruct visual semantics from audio alone. Specifically, we evaluate how well the reconstructed AV-Ground Token $\hat{g}$ preserves the original visual information, using the CLS token from the visual encoder (visual-CLS) as a reference. The comparison results are presented in table 4.

First, we measure Recall@1 and Recall@5 in an image retrieval task, where each embedding is used as a query and the visual-CLS token serves as the key in the gallery. The results show that g performs on par with visual-CLS in both metrics, while $\hat{g}$ retains approximately 80% of g's performance. Next, we project each embedding into the CLIP ViT-L/14 text embedding space and compute the CLIP Score (Hessel et al., 2022) based on cosine similarity with predefined category prompts. In this setting as well, $\hat{g}$ exhibits semantic consistency comparable to g, indicating that the reconstructed token successfully preserves semantic class information even in the absence of visual input. Additionally, we evaluate the attention entropy (Zhang et al., 2024) of each embedding to assess the degree of information concentration.

These findings collectively demonstrate that the combination of the AV-Grounding module and the VFR enables the audio encoder to effectively internalize latent visual semantics.

	CIDEr		V-CLS	FLOPs (G)	Latency(ms)
K	VggCaps	Clotho	R@1		
1	53.5	51.7	66.3	0.67	23.2
4	**54.2**	52.5	**71.8**	0.77	24.7
16	54.1	**52.6**	71.6	1.18	28.8
64	53.7	52.1	71.3	2.83	34.3
256	53.4	51.8	70.7	9.72	41.9

Table 5: **Effect of varying the number of Ground Tokens** (K) on captioning performance, visual retrieval accuracy, and inference efficiency.

5.3.2 Token g Granularity vs. Performance

In this ablation study, we analyze the effect of varying the number of Ground Tokens (K) on both performance and computational efficiency. As shown in Table 5, increasing K initially improves performance—reaching the highest CIDEr and retrieval accuracy (V-CLS R@1) at $K = 4$—but further increases lead to a gradual decline. This suggests a potential trade-off, where excessively large K values may dilute the model's attention over relevant visual cues. One possible explanation is that the number of effective grounding tokens correlates with the number of salient visual perspectives the model attends to. Based on this observation, we select $K = 4$ as the final setting, offering the best balance between performance and efficiency.

6 Conclusion

To address the limitations of existing AAC datasets—particularly their short and simplistic captions—we introduce **VggCaps**, a large-scale multi-modal dataset that pairs audio with static video frames. Captions are generated using large language models (LLMs) to reflect both auditory and visual semantics, resulting in significantly longer and more linguistically rich descriptions than prior datasets. Human evaluation confirms their clarity and expressive quality.

Furthermore, we present **Multi2Cap**, a framework that incorporates visual supervision during training but generates captions from audio alone at inference. It employs an Audio-Visual Grounding Module and a Visual Feature Reconstructor to encode and recover visual semantics from audio. Multi2Cap achieves state-of-the-art performance on Clotho and AudioCaps benchmarks, with further analysis showing strong semantic alignment between reconstructed and actual visual representations.

7 Limitations

While Multi2Cap demonstrates strong performance across various automated audio captioning (AAC) benchmarks, several limitations remain that warrant further exploration.

First, the model leverages visual context only during pre-training and relies on reconstructing it from audio during inference. Although our ablation study shows that the reconstructed AV-ground token ($\hat{g}$) retains a substantial portion of the original visual semantics, this audio-only reconstruction is inherently limited. As AAC fundamentally aims to generate captions solely from audio inputs, further discussion is needed to delineate the boundary between permissible auxiliary information during training and the core objective of maintaining audio-only inference.

Second, the VggCaps dataset, while significantly more descriptive and diverse than prior AAC datasets, is constructed using synthetic captions generated by large language models (LLMs). Although human evaluation confirms the general quality of these captions, reliance on LLM-generated annotations may introduce stylistic artifacts or latent biases that diverge from human-authored content, potentially affecting model generalization in real-world applications.

Third, the current framework has been primarily validated on English-language benchmarks. The adaptability of both Multi2Cap and VggCaps to non-English or multilingual settings remains unexplored, which poses a limitation in terms of cross-linguistic applicability and inclusivity.

8 Risks and Ethics

We acknowledge potential risks associated with the use of large-scale audio and visual data, particularly when paired with powerful language models.

Our model is trained on multimodal data that may inherently reflect socio-cultural biases present in both the audio content and the captions generated by large language models (LLMs). While we implemented filtering and human evaluation procedures to ensure overall quality, we acknowledge that such measures cannot fully eliminate the presence of biased or inappropriate content—particularly when scaling to more diverse or less curated datasets.

In addition, the methodology of reconstructing visual context from audio introduces potential privacy concerns. For instance, audio recordings captured in public or semi-private spaces may enable the model to infer or hallucinate visual scenarios that were neither recorded nor consented to. Such capabilities could be misused in surveillance or profiling applications.

To mitigate these risks, no personally identifiable or sensitive data were used in the construction of the VggCaps dataset, and all human evaluators confirmed the absence of sensitive content in the validation subset. Furthermore, we advocate for the responsible use and deployment of Multi2Cap within ethical frameworks that emphasize transparency, data governance, and informed user consent.

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.

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

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2021. Localizing visual sounds the hard way. *Preprint*, arXiv:2104.02691.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

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.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Audioclip: Extending clip to image, text and audio. *Preprint*, arXiv:2106.13043.

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.

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

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

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

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2022. Perceiver io: A general architecture for structured inputs outputs. *Preprint*, arXiv:2107.14795.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *Preprint*, arXiv:2102.05918.

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.

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*.

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*.

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*.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2022. Attention bottlenecks for multimodal fusion. *Preprint*, arXiv:2107.00135.

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

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

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

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*.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *Preprint*, arXiv:2201.02184.

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.

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, Jeeweon 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. 2024. Beyond the status quo: A contemporary survey of advances and challenges in audio captioning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:95–112.

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

Zhisong Zhang, Yan Wang, Xinting Huang, Tianqing Fang, Hongming Zhang, Chenlong Deng, Shuaiyi Li, and Dong Yu. 2024. Attention entropy is a key factor: An analysis of parallel context encoding with full-attention-based pre-trained language models. *Preprint*, arXiv:2412.16545.

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 6. Based on Multi2Cap, 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 CLIP-ViT-Large visual encoders, with the visual resolution set to 224×224 pixels. Additionally, the *RandomResizedCrop* technique is used for visual 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
Visual encoder	CLIP-ViT-Large
Visual resolution	224×224
Visual augmentation	RandomResizedCrop

Table 6: 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 7. Based on Multi2Cap, 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 7: Default fine-tuning setting

A.2 Additional Experiments

A.2.1 Hyperparameter Optimization for α and β

To balance the relative contributions of each loss term in the Multi2Cap framework, we introduce two hyperparameters: α for the alignment loss $\mathcal{L}_{\text{align}}$ and β for the reconstruction loss $\mathcal{L}_{\text{vfr}}$. The overall training objective is defined as Eq 4

We perform a grid search over various combinations of α and β to empirically determine the optimal setting. Table 8 presents the CIDEr and SPICE scores for each pair of hyperparameter values. The results indicate that moderate weighting values strike a better trade-off: overly small values underutilize auxiliary supervision, while excessively large values degrade caption quality by overemphasizing alignment or reconstruction. We observe that $\alpha = 0.02$ and $\beta = 0.05$ yield the best overall performance, achieving a CIDEr score of 54.2 and a SPICE score of 15.9. We therefore adopt this configuration as the final setting in all subsequent experiments.

α / β	0.01	0.02	0.05	0.09	0.10
0.01	53.5 / 14.8	53.7 / 15.0	54.0 / 15.4	53.6 / 15.2	53.4 / 15.0
0.02	53.6 / 15.0	54.0 / 15.6	**54.2 / 15.9**	53.9 / 15.4	53.7 / 15.2
0.05	53.4 / 14.7	53.8 / 15.3	54.0 / 15.7	53.6 / 15.2	53.3 / 15.0
0.09	53.0 / 14.6	53.4 / 15.0	53.7 / 15.4	53.5 / 15.1	53.2 / 14.9
0.10	52.9 / 14.5	53.3 / 14.9	53.6 / 15.3	53.4 / 15.0	53.5 / 15.2

Table 8: Performance comparison by α & β (CIDEr SPICE)

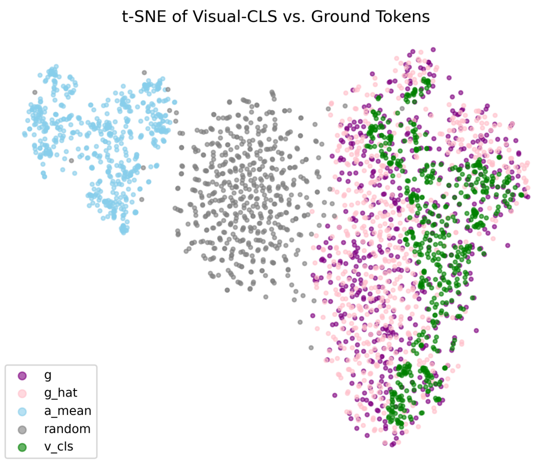


Figure 6: **t-SNE visualization of Visual-CLS and ground token variants in 2D space.**

A.2.2 t-SNE Visualization of Ground Token Semantics

We project different embeddings—including the original AV-Ground Token (g), its audio-reconstructed counterpart ($\hat{g}$), the average-pooled audio representation (a_mean), random vectors, and the visual encoder's CLS token (v_cls)—into a 2D space using t-SNE. The figure shows that g and $\hat{g}$ are closely distributed around v_cls, indicating strong semantic alignment. In contrast, a_mean and random embeddings are clustered far from v_cls, suggesting limited visual semantic relevance.

A.2.3 Comparison About Audio Content Augmentation

Additionally, the results based on whether audio content augmentation was applied are examined. The results are presented in table 9. 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.

	w/o augment	**w/ augment**
B-1	34.1	**34.3**
B-2	18.7	**19.2**
B-3	10.9	**11.6**
B-4	6.9	**7.6**
RG-L	29.1	**29.1**
ME	12.7	**15.4**
CD	51.8	**54.2**
SP	14.2	**15.9**
SD	33.0	**35.1**
SD-F	30.9	**33.3**
SB	62.6	**65.4**
FS	58.2	**58.9**

Table 9: Comparison about audio content augmentation

A.3 VggCaps

A.3.1 Prompt Templates

In Figure 1, the input-prompt and post-prompt used in VggCaps data generation are illustrated. The input-prompt provided in listing 1 receives only the audio spectrum as input and is designed to generate an initial caption that broadly describes the auditory content. Basic rules are applied to guide the captioning process, focusing solely on audio-based elements. In contrast, the post-prompt provided in listing 2 incorporates a visual frame in addition to the audio input, and refines the raw caption into a more general, descriptive sentence. This stage applies additional constraints and provides examples to ensure the generation of contextually rich and visually grounded captions.

A.3.2 Examples

Table 10 shows samples from the constructed VggCaps.

```
The image is an audio spectrum, categorized into {} category.
Write a caption that accurately describes the sounds represented in
   this spectrum:

[Rules]
1. Focus solely on auditory elements.
2. Use clear and descriptive language.
3. Avoid any references to visual content.
4. Exclude anything outside of the caption.

>>> Caption:
```

Listing 1: Input Prompt Template for VggCaps

```
Revise the caption below to a more descriptive and contextually
   accurate sentence, making appropriate use of visual details:

[Rules]
1. Focus on both audio and visual elements.
2. Avoid repetitive expressions and redundant descriptions.
3. Use expressive and idiomatic language for vivid imagery.
4. Maintain clarity and coherence in sentence structure.
5. Exclude anything other than captions.

[example]
1. A vibrant outdoor volleyball match in full swing, with the ball
   soaring over the net amidst cheers from an enthusiastic crowd and
   the rhythmic thump of the volleyball meeting players' hands.
2. A quiet forest, where a gentle breeze rustles the leaves and a
   bird's melodious song weaves through the tranquil air.
3. An energetic street performance, the rhythmic beat of drums
   accompanied by the cheering crowd, and the colorful dance moves
   creating a festive atmosphere.

>>> Caption:
```

Listing 2: Post Prompt Template for VggCaps

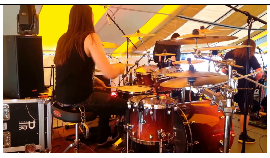
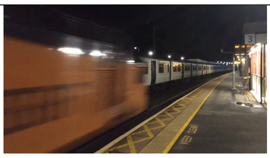
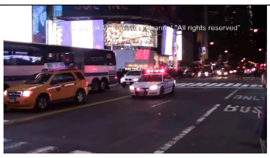
ID	Image	Caption	Category
1v5mmZoJJ50		Her fingers dance gracefully on the sitar strings, weaving a tapestry of sound that resonates with the serene and soulful essence of the music.	playing sitar
5IuRzJRrRpQ		The joyful bleats of sheep mingle with the soft rustling of grass and the occasional bark of 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 10: Examples of VggCaps