

CARD: Cross-component Audio Representation Distillation for Encoder-Free Audio Captioning

Ganesh Pavan Kartikeya Bharadwaj Kolluri¹, Yuchen Zhang^{1,2}, Michael Kampouridis¹, Ravi Shekhar^{1,2}

¹*School of Computer Science and Electronic Engineering, University of Essex*

²*Institute for Analytics and Data Science, University of Essex*

{karthik.kolluri, yuchen.zhang, mkampo, r.shekhar}@essex.ac.uk

Abstract—Modern automated audio captioning systems pair a frozen audio encoder with a large language model (LLM) via a trainable projector, incurring the encoder’s inference cost and bottlenecking the model through its fixed acoustic features. We present CARD, an encoder-free audio captioning model that removes the encoder at inference: a 13.2M projector feeds a frozen LLM with merged LoRA adapters, while the teacher used to train it is discarded. CARD distills a pretrained audio teacher (CLAP-HTSAT) into the model, but rather than injecting it into the LLM alone, it routes the teacher’s representations across components: perceptual stages to the projector and semantic stages to the LLM. This placement improves CIDEr-D by +12.18 over an LLM-only distilled model on AudioCaps and by +5.21 on Clotho, reaching 55.4 against a 66.4 encoder-kept upper bound with no encoder at inference, showing that where a teacher’s knowledge is placed matters as much as its presence.

Index Terms—audio captioning, encoder-free models, cross-component distillation, audio-language models.

I. INTRODUCTION

Automated audio captioning (AAC) is the task of generating a natural-language description of the sound events in an audio clip, such as “rain falls as thunder rumbles in the distance.” Unlike automatic speech recognition (ASR), which transcribes spoken words from the speech audio, AAC aims to describe the broader acoustic scene, capturing non-speech sounds and their relationships in free-form natural language. This ability to summarize what an audio clip contains has made AAC valuable for applications such as accessibility, audio retrieval, and content understanding [1], [2].

Most recent AAC systems follow a common architectural design: A frozen pre-trained audio encoder, such as CLAP [3], first converts the raw audio clip into a sequence of acoustic feature vectors. A trainable projector then maps these features into the input space of the Large Language Model (LLM), which decodes them into a caption. Since both the audio encoder and the LLM are pre-trained separately and usually kept frozen, only the projector is trained to bridge the two components together. This design has been widely adopted because it inherits the strengths of a powerful audio encoder and a capable language model while learning little more than the alignment between them [4]–[6]. Its main limitation, however, is that inference still depends on the dedicated audio encoder. Every clip must pass through a full encoder forward pass before the language model can act, which adds computation and memory, and confines the language model to perceiving audio only through the encoder’s fixed features.

A natural way to remove this dependency is to discard the audio encoder from the pipeline and let the model caption audio on its own. However, without the encoder, the projector and the language model must learn to perceive audio directly from the raw acoustic signal, rather than building on representations from a model already trained on large amounts of audio. Learning to extract useful acoustic features from scratch, while at the same time learning to generate captions, places a much heavier burden on training and tends to be unstable. For this reason, encoder-free models have generally struggled to reach the performance of their encoder-based counterparts [7], [8].

A promising way to ease this difficulty is knowledge distillation. Instead of learning to perceive audio entirely from scratch, the model can be guided during training by a pre-trained encoder acting as a teacher, and this teacher is then discarded at inference, so the model retains the teacher’s knowledge without paying its runtime cost [8]. Existing approaches, however, distill the teacher almost entirely into the language model, while the projector receives no representation-level supervision [9]. This is particularly important for audio captioning because the projector is responsible for converting the acoustic signal into the audio tokens consumed by the language model. Unlike the language model, however, it receives no direct representation-level supervision and must instead learn solely from the captioning objective. This raises a question that prior work has not examined. Rather than asking whether to distill, we ask how teacher knowledge should be distributed across the components of an encoder-free AAC model.

We address this question with CARD, a cross-component distillation framework for encoder-free audio captioning. The central idea is to assign teacher representations to student components according to their functional roles. Earlier teacher layers provide lower-level acoustic cues for supervising the audio projector, while later teacher layers provide more abstract information for supervising the language model. In this way, the projector learns to form audio tokens from low-level acoustic supervision, and the language model learns to interpret them with higher-level semantic supervision. At inference, the teacher and all distillation heads are removed, and the LoRA adapters are merged into the base model. The deployed model contains only the audio projector and the language model, with no audio encoder.

Our contributions are as follows:

- We propose **CARD**, an encoder-free model for audio captioning that removes the audio encoder at inference through teacher-guided distillation, leaving only a frozen, LoRA-adapted language model and an audio projector.
- We introduce the **cross-component distillation** framework, which assigns the teacher’s early representations to the audio projector and its later representations to the language model, matching each component to the level of knowledge it is best suited to learn.
- We provide **comprehensive experiments** showing that the placement of teacher supervision is critical for encoder-free AAC, and that CARD substantially outperforms LLM-only distillation strategies that supervise the language model alone.

II. RELATED WORK

A. Encoder-based automated audio captioning

Recent LLM-based AAC systems mostly share a common architectural design, in which a frozen pretrained audio encoder first extracts acoustic representations, and a trainable projector subsequently aligns these representations with a language decoder. EnCLAP [4] employs a frozen CLAP encoder together with a BART decoder. DRCap [6] further incorporates retrieval-augmented prompting, while SLAM-AAC [5] couples the EAT encoder with an LLM through a lightweight projector. LOAE [10] follows a similar design using the CED [11] encoder.

These systems have demonstrated strong captioning performance. However, maintaining a separate audio encoder introduces several limitations. First, the encoder remains active during inference, increasing computational and memory requirements beyond the language model itself [12]. Second, the language model is restricted to operating within the encoder’s frozen representation space, making acoustic information discarded by the encoder inaccessible downstream [8], [13]. Third, bridging pretrained audio encoders and language models commonly relies on training a projector on paired audio-text data to align encoder outputs with the language model embedding space [4], [5]. These considerations motivate the question of whether the audio encoder can be removed entirely.

B. Encoder-free audio-language models

Encoder-free multimodal large language models aim to internalize modality understanding directly within the language model, eliminating the need for dedicated modality encoders at inference time. This paradigm was first explored in the vision domain, where Fuyu-8B [13], EVE [7], and Mono-InternVL [14] demonstrated that visual representations can be learned within an LLM through lightweight projections of raw image inputs, while VoRA [8] further introduced teacher-guided distillation as an alternative route to encoder-free adaptation.

Such encoder-free designs have only recently begun to be explored for audio understanding. Early studies mainly rely

on downstream supervision to induce acoustic understanding within an LLM. Mel-LLM [12] projects Mel-spectrogram patches through a linear projector into a LoRA-adapted LLM and learns acoustic representations from speech-oriented tasks such as ASR and Text to Speech (TTS), without explicit representational supervision. Gemma4-12B [15] follows a similar philosophy at a larger scale by removing dedicated audio encoders altogether and projecting raw audio directly into the LLM, again without supervision of the audio frontend.

More recently, AuRA [9], concurrent with our work, introduced teacher-guided adaptation for encoder-free audio understanding. It distills a frozen ASR encoder into a LoRA-adapted LLM through layer-wise hidden-state alignment. Its supervision, however, is applied entirely on the language-model side, leaving the audio frontend unsupervised. Moreover, AuRA focuses on speech-oriented tasks such as transcription, whereas AAC requires recognizing and describing diverse environmental and non-speech acoustic events in natural language.

Together, these studies demonstrate that audio understanding can be internalized into an LLM. Existing approaches differ primarily in how acoustic knowledge is acquired. Mel-LLM and Gemma 4 rely solely on downstream objectives to induce acoustic representations, whereas AuRA introduces supervision from a pretrained teacher. However, even in the latter case, supervision is applied only to language-model layers, leaving the audio frontend that first transforms raw signals into tokens unsupervised. Consequently, the question of where teacher knowledge should be injected within an encoder-free audio-LLM remains largely unexplored.

C. Distillation for encoder-free multimodal models

Knowledge distillation [16] transfers information from a teacher model to a student model through output- or representation-level supervision. In the LLM era, distillation has evolved from a model compression technique into a mechanism that enables encoder-free multimodal adaptation, allowing modality-specific knowledge to be internalized into language models without retaining dedicated encoders at inference [17].

Recent approaches mainly employ distillation strategies based on hidden-state alignment. VoRA [8] distills a frozen vision transformer into LoRA adapters inserted within an LLM via block-wise supervision, after which the adapters are merged and the teacher discarded. AuRA [9] adopts a similar strategy for speech, aligning hidden representations from a frozen ASR encoder with LoRA-adapted language layers. In this view, distillation is not merely a compression technique applied after architectural design, but the mechanism by which an encoder-free architecture becomes achievable.

Despite their effectiveness, existing cross-modal distillation methods inject supervision into a single class of locations within the student, typically the language-model blocks, while treating the modality frontend as an unsupervised projection. Consequently, whether teacher knowledge should instead be distributed across multiple components of an encoder-free

audio-LLM according to their distinct functional roles remains unexplored.

III. METHODOLOGY

CARD is an audio captioning model that eliminates the need for a dedicated audio encoder. Unlike conventional architectures [5] that rely on a pretrained audio encoder, projector, and LLM decoder, CARD learns acoustic knowledge through representation-level distillation from a frozen teacher model. The student model consists only of a audio projector and a LoRA-adapted LLM. After training, the teacher is removed, and the LoRA adapters are merged into the LLM, resulting in an efficient architecture comprising only the LLM backbone and audio projector while maintaining audio-captioning capability. The overview of CARD is shown in Figure 1.

A. Model Architecture

Student model. The student model consists of a lightweight audio projector followed by a LoRA-adapted large language model. Given an input audio clip, the projector converts the audio into a sequence of audio tokens, which are prepended to the text prompt embeddings and consumed by the language model to generate the output caption.

Specifically, the audio is first resampled to 48 kHz and converted into a 64-band log-Mel spectrogram using the CLAP processor [3]. The audio projector consists of a stack of strided one-dimensional convolutional layers, followed by a linear projection to the LLM hidden size and layer normalization. The output is a sequence of audio tokens used as the prefix to the language model. The resulting audio-token sequence forms the prefix presented to the language model.

We adopt Qwen3-4B [18] as the language backbone. Adaptation is performed using LoRA [19] on the linear layers of every transformer block, including the query, key, value, output, gate, up, and down projections. All other parameters are frozen, so only the LoRA parameters are optimized during training. After training, the LoRA weights are merged into the backbone, introducing no additional parameters or computational overhead at inference.

Teacher model. We employ a frozen CLAP [20] as the teacher throughout training. CLAP is contrastively pretrained on paired audio-text data and serves as the audio encoder in several state-of-the-art AAC systems [4], [6]. Its representations therefore provide caption-relevant acoustic knowledge for supervising the student model.

The CLAP audio branch adopts an HTSAT backbone [21], implemented as a hierarchical Swin Transformer [22]. The backbone produces a hierarchy of intermediate representations with progressively increasing semantic abstraction. During training, the teacher receives the same log-Mel spectrogram as the student. The resulting intermediate representations are used as supervision signals in the component-aware distillation described next.

B. Cross Component Distillation

In existing teacher-guided encoder-free audio models [9], teacher supervision is typically applied only to language-model layers, leaving the audio projector trained solely through the downstream captioning objective. Since every audio token consumed by the language model is produced by the projector, insufficient supervision at the projector may limit the quality of all downstream representations.

CARD addresses this mismatch by aligning the teacher hierarchy with the functional structure of the student. The CLAP teacher provides acoustic representations at multiple levels of abstraction through its hierarchical HTSAT backbone. The student, in contrast, separates computation across two components: the projector transforms raw acoustic signals into language-compatible representations, and the language model interprets these representations for semantic reasoning and caption generation. Teacher supervision is therefore assigned according to component function, with lower-level acoustic representations supervising the projector and higher-level semantic representations supervising the language model.

Specifically, we extract the outputs of the four HTSAT stages, $\{t_0, t_1, t_2, t_3\}$, as teacher representations. Since the semantic abstraction of these representations increases with network depth, we use the earlier stages to supervise the audio projector and the later stages to supervise the language model.

For projector supervision, the audio projector produces a single sequence of audio tokens. Let u_{proj} denote the mean-pooled projector representation. We attach one linear distillation head for t_0 and another for t_1 , both operating on the same mean-pooled projector representation:

$$\hat{u}_0 = W_0 u_{\text{proj}}, \quad \hat{u}_1 = W_1 u_{\text{proj}}. \quad (1)$$

The projected representations $\hat{u}_0$ and $\hat{u}_1$ are then matched to the mean-pooled teacher features from t_0 and t_1 using cosine-based distillation.

For language-model supervision, we use the hidden representations at the audio-token positions from the lower 12 transformer blocks. Let u_{llm}^1 denote the average representation over blocks 0–8, and let u_{llm}^2 denote the average representation over blocks 9–11. Two linear distillation heads then project these representations to the dimensionalities of the corresponding teacher stages:

$$\hat{u}_2 = W_2 u_{\text{llm}}^1, \quad \hat{u}_3 = W_3 u_{\text{llm}}^2. \quad (2)$$

Similarly, the projected representations $\hat{u}_2$ and $\hat{u}_3$ are compared with the corresponding teacher representations t_2 and t_3 using the cosine distillation loss. The remaining transformer blocks are not directly supervised by the teacher and are optimized only through the captioning objective.

The pairwise distillation loss for each teacher-student representation pair is defined as

$$\ell_{\text{distill}}(\hat{u}_i, t_i) = 1 - \cos(\hat{u}_i, t_i), \quad (3)$$

where $i \in \{0, 1, 2, 3\}$ indexes the four teacher stages, and $\hat{u}_i$ and t_i denote the projected student representation and the corresponding teacher representation, respectively.

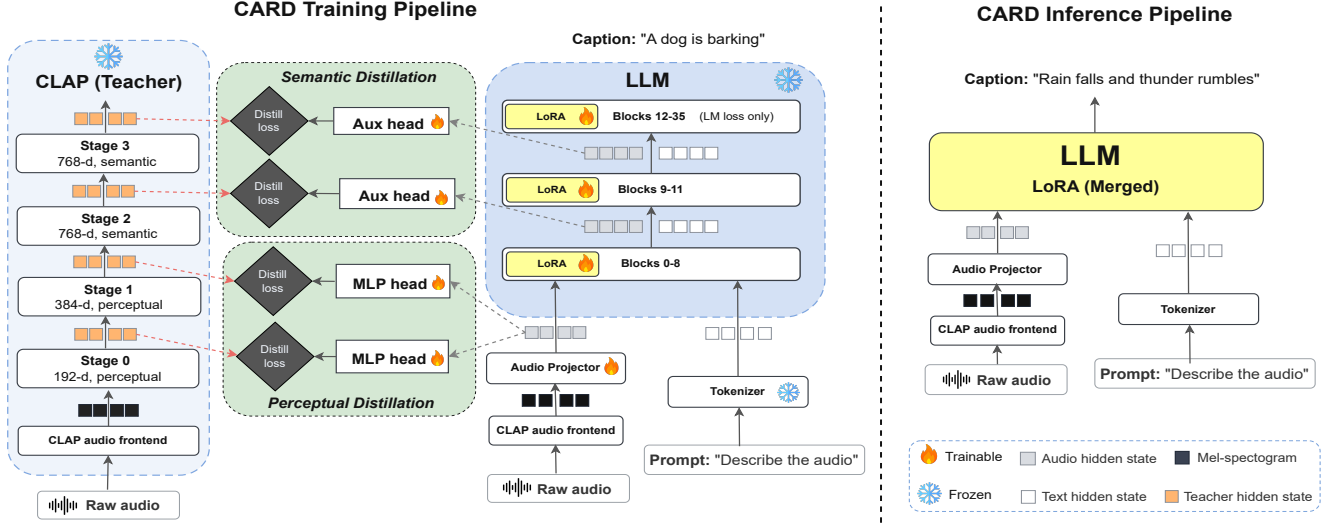


Fig. 1. Overview of CARD. During training (left), a frozen CLAP teacher supervises the two student components by role: its early, perceptual stages (0-1) distill into the audio projector, and its later, semantic stages (2-3) into the LLM’s LoRA adapters. At inference (right), the teacher and distillation heads are dropped, and the LoRA is merged into the LLM, leaving only the projector and the LLM, with no audio encoder.

C. Training Objective

CARD is trained in two consecutive phases with different optimization objectives. During Phase 1, the student model is jointly optimized with a captioning loss, $\mathcal{L}_{\text{cap}}$, a projector distillation loss, $\mathcal{L}_{\text{proj}}$, and a language-model distillation loss, $\mathcal{L}_{\text{llm}}$, while the CLAP teacher remains frozen. The audio projector, LoRA adapters, and distillation heads are jointly updated to transfer caption-relevant acoustic knowledge from the teacher into the encoder-free student. After Phase 1, the teacher model and all distillation heads are removed.

Specifically, the projector distillation loss is obtained by averaging the pairwise distillation losses associated with the lower teacher stages,

$$\mathcal{L}_{\text{proj}} = \frac{1}{2} (\ell_{\text{distill}}(\hat{u}_0, t_0) + \ell_{\text{distill}}(\hat{u}_1, t_1)). \quad (4)$$

The language-model distillation loss is computed from the higher teacher stages,

$$\mathcal{L}_{\text{llm}} = \frac{1}{2} (\ell_{\text{distill}}(\hat{u}_2, t_2) + \ell_{\text{distill}}(\hat{u}_3, t_3)). \quad (5)$$

The captioning loss is the standard autoregressive cross-entropy loss over the target caption tokens,

$$\mathcal{L}_{\text{cap}} = - \sum_{i=1}^N \log p(y_i | y_{<i}, x), \quad (6)$$

where x denotes the input audio tokens and $y = \{y_1, \dots, y_N\}$ is the target caption.

The overall training objective during Phase 1 is

$$\mathcal{L}_{\text{phase1}} = \mathcal{L}_{\text{cap}} + \lambda_{\text{proj}} \mathcal{L}_{\text{proj}} + \lambda_{\text{llm}} \mathcal{L}_{\text{llm}}, \quad (7)$$

where $\lambda_{\text{proj}} = \lambda_{\text{llm}} = 1$ in all experiments. The teacher model and all distillation heads are used only during training

and are discarded afterwards. At inference, only the audio projector and the LoRA-merged language model are retained.

During Phase 2, the student model is fine-tuned separately for each evaluation dataset using only the captioning loss $\mathcal{L}_{\text{cap}}$, allowing the model to adapt to the target caption distribution without additional teacher supervision:

$$\mathcal{L}_{\text{phase2}} = \mathcal{L}_{\text{cap}}. \quad (8)$$

After Phase 2, the LoRA adapters are merged into the language-model backbone, yielding the final deployed model with only the lightweight audio projector and the language model.

IV. EXPERIMENTS

A. Experimental Setting

Datasets. Table I summarizes the datasets used in the two training phases. Phase 1 uses a large-scale mixture of captioned-audio datasets for representation learning, including WavCaps [23], Auto-ACD [24], AudioCaps [1], Clotho [2], and MACS [25]. To preserve the language understanding capability of the backbone LLM, text-only instruction data from Infinity-Instruct [26] are interleaved during training.

Phase 2 fine-tunes the pretrained student separately for each evaluation benchmark, following the two-stage setup of SLAM-AAC [5]: For AudioCaps evaluation, the model is fine-tuned on the prompt-expanded AudioCaps training set [1], while Clotho evaluation uses the corresponding prompt-expanded Clotho training set [2]. We evaluate on the standard AudioCaps and Clotho benchmarks.

Implementation details. CARD uses Qwen3-4B [18] as the language backbone and CLAP [20] as the Phase 1 teacher. The CLAP teacher is kept frozen throughout training. LoRA is applied to all seven linear projections of all 36 transformer

TABLE I
DATASET SUMMARY.

Phase	Dataset	Samples
Phase 1	WavCaps [23]	140,750
	Auto-ACD [24]	82,268
	AudioCaps [1]	45,178
	Clotho [2]	3,839
	MACS [25]	3,930
	Text-instruction mix	50,000
Phase 2	AudioCaps	225,890
	Clotho	95,975
Evaluation	AudioCaps test	975
	Clotho evaluation	1,045

blocks of Qwen3-4B. The audio projector consists of two strided 1-D convolutional layers with kernel size 3 and GELU activations. The first and second convolutional layers project the 64-channel Mel spectrogram to 512 and 2560 channels, respectively.

Phase 1 optimizes the audio projector, LoRA adapters, and distillation heads for one epoch. Optimization is performed using AdamW with $\beta = (0.9, 0.95)$, an effective batch size of 64, an initial learning rate of 2×10^{-4} , a constant-with-warmup learning-rate schedule, and gradient clipping with a maximum norm of 1.0. LoRA is configured with rank $r = 16$ and scaling factor $\alpha = 16$. Text-only instruction examples are interleaved as modality-pure batches, where audio and text examples are not mixed within a batch.

In Phase 2, the teacher and all distillation heads are removed, and the model is fine-tuned for captioning on the target dataset without distillation. A fresh LoRA adapter ($r=16$, $\alpha=16$) is then trained on this consolidated backbone for one epoch, while the projector continues from its Phase 1 checkpoint. Building captioning on top of the merged backbone lets Phase 2 specialize while leaving the Phase 1 alignment intact.

At evaluation, captions are generated using beam search with a beam size of 4 and a maximum length of 40 tokens. All experiments are conducted on $2 \times A6000$ GPUs with 48 GB memory each.

Baselines. We use SLAM-AAC as the encoder-based baseline. For a fair comparison, we replace its original audio encoder with the same CLAP encoder used in our teacher model. We report two settings, without LoRA and with LoRA, both of which retain the frozen CLAP encoder during inference.

Metrics. We use CIDEr-D (CD) [27] as the primary metric, as it measures consensus with human references and is widely reported in AAC benchmarks. We also report SPIDEr (SD) [28], SPICE (SC) [29], METEOR (MT) [30] to provide a comprehensive assessment of caption quality.

B. Overall Results

Table II compares audio captioning performance on the AudioCaps and Clotho benchmarks between the encoder-based SLAM-AAC baseline and the proposed CARD framework. Within CARD, we compare several distillation configurations. *No Distill* removes teacher supervision entirely. *LLM Distill* applies teacher supervision only to the language model. *Proj_Full* applies teacher supervision only to the audio projector, using teacher representations from all four HTSAT stages, $\{t_0, t_1, t_2, t_3\}$. *Proj_Early* also supervises only the audio projector, but uses only the early teacher stages t_0, t_1 . *CARD $^\diamond$* combines full-stage projector supervision with language-model supervision. *CARD** combines early-stage projector supervision with language-model supervision, corresponding to the projector supervision described in Section III-B.

Encoder-based baseline. The encoder-based SLAM-AAC baseline achieves the strongest overall performance. With LoRA, it reaches 66.4% CIDEr-D on AudioCaps and 39.0% on Clotho. This is expected because the CLAP encoder is retained during inference and directly provides acoustic representations to the language model.

Effect of distillation placement. Removing the encoder leads to a substantial performance drop. Without teacher supervision, *No Distill* achieves 43.2% CIDEr-D on AudioCaps and 22.3% on Clotho. Applying teacher supervision only to the language model gives almost no improvement, with *LLM Distill* reaching 43.5% and 22.5%, respectively. This suggests that language-model supervision alone cannot compensate for the absence of direct supervision on the audio projector.

Projector supervision is more effective, but only when the teacher representation matches the role of the projector. *Proj_Full* reaches 40.7% CIDEr-D on AudioCaps, while *Proj_Early* improves this to 52.5%. The same trend appears on Clotho, where *Proj_Early* improves over *Proj_Full* by 3.1%. This shows that the projector benefits more from early-stage teacher representations than from the full teacher hierarchy.

Full CARD model. Combining projector and language-model supervision gives the best encoder-free results. *CARD $^\diamond$* reaches 49.9% CIDEr-D on AudioCaps, while *CARD** reaches 55.4%. On Clotho, *CARD** also achieves the best encoder-free result, reaching 27.5% CIDEr-D. Compared with *No Distill*, *CARD** improves CIDEr-D by 12.2 percentage points on AudioCaps and 5.2 percentage points on Clotho, with consistent gains across SPIDEr, SPICE, and METEOR.

Overall, the results show that teacher supervision is most effective when early teacher representations supervise the audio projector and later teacher representations supervise the language model. This supports the component-aware distillation strategy used in *CARD**.

C. Model Analysis

Effect of LoRA Capacity. Table III studies the effect of LoRA capacity under different training settings. With cross-component distillation, the model is sensitive to the LoRA rank. Using rank 8 gives 33.3% CIDEr-D on AudioCaps and

TABLE II
AUDIO CAPTIONING PERFORMANCE ON THE AUDIOCAPS AND CLOTHO EVALUATION SPLITS. ALL VALUES ARE REPORTED AS PERCENTAGES (%).

Model	Variant	Enc.	LoRA	Distill	AudioCaps				Clotho			
					CD	SD	SC	MT	CD	SD	SC	MT
SLAM-AAC	No LoRA	✓	✗	–	59.8	38.7	17.5	23.7	32.0	22.3	12.2	16.4
	LoRA	✓	✓	–	66.4	41.9	17.3	23.3	39.0	25.8	12.7	16.2
CARD	<i>No Distill</i>	✗	✓	No teacher	43.2	27.9	12.6	18.4	22.3	15.5	8.6	13.0
	<i>LLM Distill</i>	✗	✓	LLM only	43.5	27.8	12.2	18.4	22.5	15.1	8.4	13.0
	<i>Proj_Full</i>	✗	✓	All stages→projector	40.7	26.4	12.1	17.8	21.2	14.6	8.1	12.7
	<i>Proj_Early</i>	✗	✓	Early stages→Projector	52.5	33.3	14.1	20.1	24.3	16.6	8.9	13.1
	<i>CARD</i> [◦]	✗	✓	Projector (all stages) + LLM	49.9	32.0	14.1	20.3	24.8	17.1	9.4	13.1
	<i>CARD</i> [*]	✗	✓	Projector (early stages) + LLM	55.4	35.2	15.1	21.2	27.5	18.8	10.1	14.2

17.8% on Clotho, below the *No Distill* setting at rank 16. This suggests that the adapter capacity is insufficient for jointly fitting the captioning and distillation objectives. Increasing the rank to 16 gives the best results on both datasets, reaching 55.4% on AudioCaps and 27.5% on Clotho. Increasing the rank further to 32 reduces performance to 50.0% and 24.9%, showing that larger adapters do not necessarily improve transfer.

This sensitivity is much weaker without cross-component distillation. Under *LLM Distill*, increasing α from 16 to 32 reduces CIDEr-D from 43.5% to 41.4% on AudioCaps and from 22.5% to 20.8% on Clotho. Without teacher supervision, increasing the LoRA rank from 16 to 32 changes CIDEr-D only slightly, from 43.2% to 43.9% on AudioCaps and from 22.3% to 22.6% on Clotho. These results suggest that LoRA capacity matters most when the adapter must accommodate both the captioning objective and the cross-component teacher signal.

TABLE III
EFFECT OF LORA RANK AND SCALING ON CIDEr-D UNDER DIFFERENT DISTILLATION SETTINGS.

Model	Variant	r	α	AudioCaps	Clotho
CARD	<i>CARD</i> [*]	8	8	33.3	17.8
		16	16	55.4	27.5
		32	32	50.0	24.9
	<i>LLM Distill</i>	16	16	43.5	22.5
		16	32	41.4	20.8
		16	16	43.2	22.3
	<i>No Distill</i>	32	32	43.9	22.6

Effect of Two-Phase Training. Table IV evaluates the contribution of the two training phases. Using only Phase 2 fine-tuning achieves 37.2% CIDEr-D on AudioCaps and 16.5% on Clotho. This suggests that caption supervision alone is not sufficient for learning strong audio representations in the encoder-free setting. Using only Phase 1 pretraining performs much worse, reaching 10.3% on AudioCaps and 6.3% on Clotho. This indicates that distillation pretraining alone does not adapt the model to the target captioning format.

TABLE IV
EFFECT OF THE TWO TRAINING PHASES AT $r = 16$.

Setting	AudioCaps				Clotho			
	CD	SD	SC	MT	CD	SD	SC	MT
Phase 2 Only	37.2	24.2	11.2	17.2	16.5	13.3	6.8	12.1
Phase 1 Only	10.3	10.6	11.0	16.2	6.3	6.2	6.0	11.2
Phase 1+2	55.4	35.2	15.1	21.2	27.5	18.8	10.1	14.2

The full two-stage training pipeline performs best across all reported metrics. It reaches 55.4% CIDEr-D, 35.2% SPIDEr, 15.1% SPICE, and 21.2% METEOR on AudioCaps. On Clotho, it reaches 27.5% CIDEr-D, 18.8% SPIDEr, 10.1% SPICE, and 14.2% METEOR. Compared with Phase 2 alone, the full pipeline improves CIDEr-D by 18.2 percentage points on AudioCaps and 11.0 points on Clotho. Compared with Phase 1 alone, the gains are much larger. These results show that the two phases are complementary. Phase 1 transfers acoustic knowledge from the teacher, while Phase 2 adapts the student for caption generation.

V. CONCLUSION

We presented CARD, an encoder-free framework for automated audio captioning. CARD transfers acoustic knowledge from a frozen CLAP teacher into a lightweight audio projector and a LoRA-adapted LLM during training. Our results show that knowledge distillation is important for recovering the performance lost by removing the audio encoder, but its effectiveness depends strongly on where teacher supervision is applied. LLM-only distillation gives negligible gains over the non-distilled encoder-free model, while projector-only distillation with the full teacher hierarchy degrades performance. In contrast, supervising the projector with early teacher stages substantially improves performance, showing that the projector benefits most from lower-level acoustic representations. The full CARD model combines early-stage projector supervision with later-stage language-model supervision and achieves the

best performance among all encoder-free variants. It improves AudioCaps CIDEr-D from 43.2% to 55.4% and Clotho CIDEr-D from 22.3% to 27.5%, with consistent gains across other captioning metrics. These results suggest that projector-level and language-model-level distillation provide complementary supervision. Overall, CARD demonstrates a promising approach to encoder-free audio captioning by transferring perceptual and semantic audio knowledge into the student model during training, removing the need for a separate audio encoder at inference.

AI-GENERATED CONTENT DISCLOSURE

The authors used generative AI tools solely for grammar correction and language refinement. No AI-generated content was used for the development of research ideas, analysis, interpretation of results, or scientific conclusions. The authors reviewed and approved all revised text and take full responsibility for the content of the manuscript.

REFERENCES

- [1] C. D. Kim, B. Kim, H. Lee, and G. Kim, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 119–132. [Online]. Available: <https://aclanthology.org/N19-1011/>
- [2] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 736–740. [Online]. Available: <https://doi.org/10.1109/ICASSP40776.2020.9052990>
- [3] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/ICASSP49357.2023.10095889>
- [4] J. Kim, J. Jung, J. Lee, and S. H. Woo, "Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 6735–6739. [Online]. Available: <https://doi.org/10.1109/ICASSP48485.2024.10446672>
- [5] W. Chen, Z. Ma, X. Li, X. Xu, Y. Liang, Z. Zheng, K. Yu, and X. Chen, "SLAM-AAC: enhancing audio captioning with paraphrasing augmentation and clap-refine through llms," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. IEEE, 2025, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/ICASSP49660.2025.10889071>
- [6] X. Li, W. Chen, Z. Ma, X. Xu, Y. Liang, Z. Zheng, Q. Kong, and X. Chen, "Drcap: Decoding CLAP latents with retrieval-augmented generation for zero-shot audio captioning," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*. IEEE, 2025, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/ICASSP49660.2025.10890325>
- [7] H. Diao, Y. Cui, X. Li, Y. Wang, H. Lu, and X. Wang, "Unveiling encoder-free vision-language models," *ArXiv*, vol. abs/2406.11832, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270559398>
- [8] H. Wang, Y. Ye, B. Li, Y. Nie, J. Lu, J. Tang, Y. Wang, and C. Huang, "Vision as lora," *ArXiv*, vol. abs/2503.20680, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:277322798>
- [9] B. Cheng, L. Shi, Z. Ma, Y. Wu, J. Xu, J. Gao, J. Hao, and R. He, "Aura: Internalizing audio understanding into llms as lora," 2026. [Online]. Available: <https://api.semanticscholar.org/CorpusID:289131922>
- [10] J. Liu, G. Li, J. Zhang, H. Dinkel, Y. Wang, Z. Yan, Y. Wang, and B. Wang, "Enhancing automated audio captioning via large language models with optimized audio encoding," *ArXiv*, vol. abs/2406.13275, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:270621008>
- [11] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "CED: consistent ensemble distillation for audio tagging," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 291–295. [Online]. Available: <https://doi.org/10.1109/ICASSP48485.2024.10446348>
- [12] R. Fan, Y. Wang, Y. Hu, B. Ren, Y. Xia, X. Wang, Y. Qian, and J. Li, "Llm can read spectrogram: Encoder-free speech-language modeling," 2026. [Online]. Available: <https://api.semanticscholar.org/CorpusID:289132537>
- [13] Adept AI, "Fuyu-8B: A multimodal architecture for AI agents," <https://www.adapt.ai/blog/fuyu-8b/>, 2023, blog post.
- [14] G. Luo, X. Yang, W. Dou, Z. Wang, J. Dai, Y. Qiao, and X. Zhu, "Mono-intervl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24 960–24 971, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:273233458>
- [15] Google DeepMind, "Gemma 4 model card," <https://developers.google.com/gemma-4-12b-the-developer-guide/>, 2026, online.
- [16] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: <http://arxiv.org/abs/1503.02531>
- [17] Y. Gu, L. Dong, F. Wei, and M. Huang, "Minillm: On-policy distillation of large language models," 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259164722>
- [18] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv et al., "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>
- [20] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/ICASSP49357.2023.10095969>
- [21] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 646–650. [Online]. Available: <https://doi.org/10.1109/ICASSP43922.2022.9746312>
- [22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 9992–10 002. [Online]. Available: <https://doi.org/10.1109/ICCV48922.2021.00986>
- [23] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 3339–3354, 2024. [Online]. Available: <https://doi.org/10.1109/TASLP.2024.3419446>
- [24] L. Sun, X. Xu, M. Wu, and W. Xie, "Auto-act: A large-scale dataset for audio-language representation learning," in *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, J. Cai, M. S. Kankanhalli, B. Prabhakaran, S. Boll, R. Subramanian, L. Zheng, V. K. Singh, P. César, L. Xie, and D. Xu, Eds. ACM, 2024, pp. 5025–5034. [Online]. Available: <https://doi.org/10.1145/3664647.3681472>
- [25] I. M. Morató and A. Mesaros, "Macs - multi-annotator captioned soundscapes," 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:244972894>
- [26] J. Li, L. Du, H. Zhao, B. wen Zhang, L. Wang, B. Gao, G. Liu, and Y. Lin, "Infinity instruct: Scaling instruction selection and synthesis to

- enhance language models,” *ArXiv*, vol. abs/2506.11116, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:279391576>
- [27] R. Vedantam, C. L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 4566–4575. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7299087>
 - [28] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, “Improved image captioning via policy gradient optimization of spider,” in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 873–881. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.100>
 - [29] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9909. Springer, 2016, pp. 382–398. [Online]. Available: https://doi.org/10.1007/978-3-319-46454-1_24
 - [30] S. Banerjee and A. Lavie, “METEOR: an automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, J. Goldstein, A. Lavie, C. Lin, and C. R. Voss, Eds. Association for Computational Linguistics, 2005, pp. 65–72. [Online]. Available: <https://aclanthology.org/W05-0909/>