DIFFUSION-LINK: DIFFUSION PROBABILISTIC MODEL FOR BRIDGING THE AUDIO-TEXT MODALITY GAP

KiHyun Nam[1*], *Jongmin Choi*[1*], *Hyeongkeun Lee*[1], *Jungwoo Heo*[2], *Joon Son Chung*[1]

[1]Korea Advanced Institute of Science and Technology, South Korea, [2]University of Seoul, South Korea

ABSTRACT

Contrastive audio–language pretraining yields powerful joint representations, yet a persistent audio–text modality gap limits the benefits of coupling multimodal encoders with large language models (LLMs). We present Diffusion-Link, a diffusion-based modality-bridging module that generatively maps audio embeddings into the text-embedding distribution. The module is trained at the output embedding from the frozen multimodal encoder and implemented as a lightweight network with three residual MLP blocks. To assess the effect of Diffusion-Link on multimodal encoder-LLM coupling, we evaluate on Automatic Audio Captioning (AAC); to our knowledge, this is the first application of diffusion-based modality bridging to AAC. We report two results. (1) Modality-gap analysis: on similarity and geometric criteria, Diffusion-Link reduces the modality gap the most among prior diffusion-based methods and shows a collective migration of audio embeddings toward the text distribution. (2) Downstream AAC: attaching Diffusion-Link to the same multimodal LLM baseline achieves state-of-the-art on AudioCaps in both zero-shot and fully supervised captioning without external knowledge, with relative gains up to 52.5% and 7.5%, respectively. These findings show that closing the modality gap is pivotal for effective coupling between multimodal encoders and LLMs, and diffusion-based modality bridging offers a promising direction beyond knowledge-retrieval-centric designs. We publish our code here[1].

***Index Terms*—** diffusion probabilistic model, modality gap, large language model, audio captioning, multimodal representation learning

1. INTRODUCTION

Large-scale audio–language models have shown strong multimodal performance across a range of multimodal tasks. In particular, CLAP [1, 2] maps natural-language descriptions and acoustic signals into a shared embedding space via contrastive learning, achieving state-of-the-art results on various audio–language multimodal tasks [3]. In parallel, advances in LLMs [4–6] enable coupling contrastive audio–language encoders with powerful decoders, already demonstrating compelling audio–language reasoning and captioning [7–9].

Yet recent studies reveal a structural modality gap in contrastive multimodal encoders. Liang et al. [10] quantified the gap and linked its magnitude to zero-shot performance and fairness, while Zhang et al. [11] analyzed embedding geometry and showed that gap reduction benefits cross-modal tasks. From an application angle, linking contrastive spaces [1, 12] via mediating modalities enables unpaired transfer [13], and broader alignment across audio–vision–text–3D yields competitive zero-shot results [14]. Taken together, these prior works suggest that addressing the modality gap is essential for improving zero-shot and cross-modal task performance.

Diffusion models [15, 16] have become a standard generative paradigm in various fields, reliably producing high-fidelity samples [17–19]. They learn a forward noising process toward an isotropic Gaussian and a reverse denoising process back to the target distribution. Viewing embedding vector as data, diffusion can learn a trajectory that bridges the embedding distributions between two modalities. We adopt this view and design a reverse process that first moves audio embeddings to a shared isotropic Gaussian waypoint and then maps them into the text-embedding distribution, thereby enabling effective modality bridging.

Recent embedding-generative works support this view. In speaker recognition, SEED [20] applies the forward process to both clean and noisy speaker embeddings and trains the reverse process to regenerate the clean speaker embeddings, introducing cross-sample prediction and demonstrating embedding-level generation. In vision–language, Diffusion-Bridge [21] trains only on CLIP text embeddings and injects image embeddings at an intermediate reverse step to convert them into text-like vectors–an early instance of embedding-space modality bridging.

We propose **Diffusion-Link**, which directly bridges the audio–text modality gap, building on prior works [20, 21]. The key idea is to (i) use paired audio–text embeddings from an audio-language multimodal encoder during training to explicitly connect the two distributions, and (ii) achieve modality bridging by enforcing that the reverse process always map to the text embedding distribution. To this end, we gradually inject Gaussian noise into both embeddings in the forward process to send them to a common isotropic Gaussian state, and train with an L2 reconstruction loss so that the reverse process consistently predicts embeddings from the text distribution. Moreover, we add a topology loss that preserves the relative geometry of the text distribution by matching the within-batch cosine similarity structure of the original text and the generated text-like embeddings. At inference, Diffusion-Link outputs a text-like embedding regardless of the input modality. Diffusion-Link is a lightweight network composed of three residual multilayer perceptron (MLP) blocks, and the multimodal encoder is frozen during training. For practical validation, we attach Diffusion-Link after multimodal encoder as a plug-in and combine it with a LLM-based decoder to evaluate audio captioning. To our knowledge, this is the first attempt to apply diffusion-based modality bridging to audio captioning.

We verify consistent gains on the AudioCaps [22] dataset along two axes: modality-gap analysis and LLM-based downstream tasks. On similarity and geometric criteria, Diffusion-Link increases the similarity of paired audio–text samples while decreasing that of unpaired, **achieving the largest gap reduction** over prior methods. Visualizations further show a clear collective migration of audio embeddings toward the text-embedding distribution after the diffusion process. In Automatic Audio Captioning (AAC), attaching

*These authors contributed equally to this work.

[1]Official code: https://github.com/DevKiHyun/Diffusion-Link

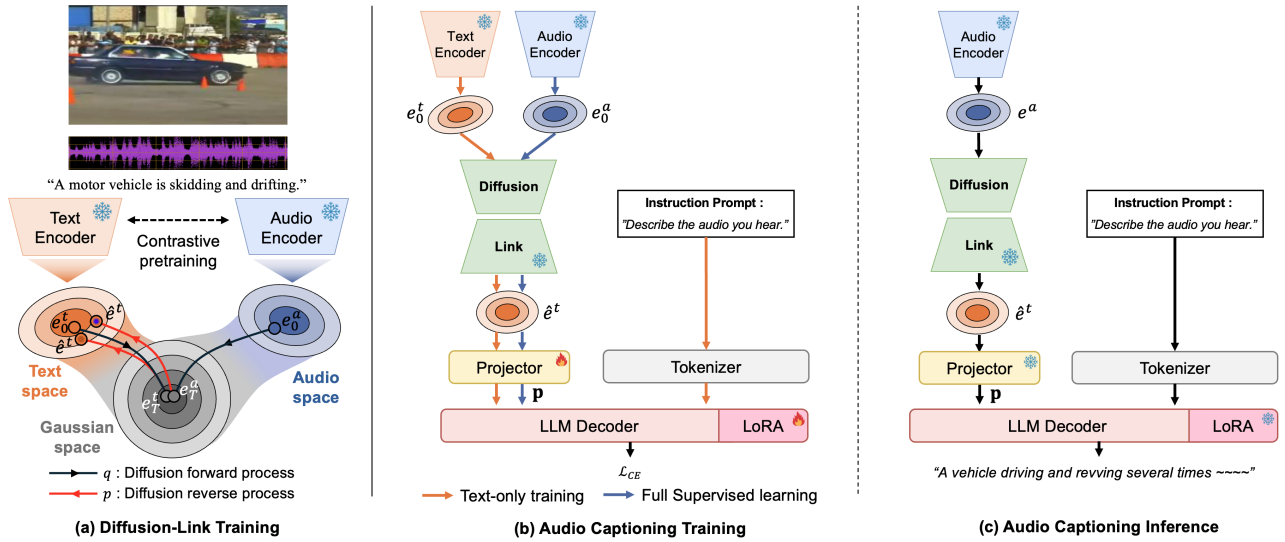

Fig. 1: (a) Overview of the proposed Diffusion-Link mechanism and (b,c) illustration of our LLM-based AAC system with Diffusion-Link.

Diffusion-Link as a plug-in to the same multimodal LLM baseline yields relative improvements of up to **52.5%** in zero-shot audio captioning and **7.5%** in fully supervised audio captioning, reaching **state-of-the-art in both cases** without external knowledge. Because many existing systems, especially in zero-shot, rely on external knowledge such as retrieval-augmented generation (RAG), these results establish Diffusion-Link as a new powerful solution that achieves consistent gains on the same multimodal LLM system while shifting the source of performance from knowledge retrieval to modality bridging.

2. METHOD

In this section, we describe the proposed framework (Fig. 1). We denote by $\mathbf{e}_0^a, \mathbf{e}_0^t \in \mathbb{R}^d$ the paired audio and text embeddings obtained from a multimodal encoder [1]. For brevity, we use $\mathcal{M}$ to indicate the modality, with $\mathcal{M}$ representing audio a and text t.

2.1. Background on Diffusion Probabilistic Models

We briefly review denoising diffusion probabilistic models (DDPM) [15] under sample-prediction formulation.

The forward diffusion process progressively corrupts a given sample $\mathbf{z}_0 \sim q(\mathbf{z}_0)$ at each timestep $s = 1, \ldots, T$:

$$q(\mathbf{z}_s|\mathbf{z}_0) = \mathcal{N}\big(\mathbf{z}_s; \sqrt{\bar{\alpha}_s}\,\mathbf{z}_0, (1-\bar{\alpha}_s)\mathbf{I}\big), \tag{1}$$

where $0 \le \alpha_s \le 1$ is the noise schedule, $\bar{\alpha}_s = \prod_{\tau=1}^{s} \alpha_\tau$, and $\mathbf{I}$ is the identity matrix. This also admits the following closed-form reparameterization:

$$\mathbf{z}_s = \sqrt{\bar{\alpha}_s}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_s}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

The reverse diffusion process gradually denoises $\mathbf{z}_t$ back toward the data distribution at each timestep s:

$$p_\theta(\mathbf{z}_{s-1}|\mathbf{z}_s) = \mathcal{N}\big(\mathbf{z}_{s-1};\, \mu_\theta(\mathbf{z}_s, s), \sigma_s^2\mathbf{I}\big), \tag{3}$$

where $\mu_\theta(\mathbf{z}_s, s)$ is parameterized by a neural denoiser. The denoiser $\phi_\theta(\cdot, s)$ is trained to predict the sample $\mathbf{z}_0$ at $s = 0$ via the objective

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, s, \boldsymbol{\epsilon}} \big\| \mathbf{z}_0 - \phi_\theta(\mathbf{z}_s, s) \big\|_2^2. \tag{4}$$

2.2. Modality Gap Bridging via Diffusion-Link

Diffusion-Link is a neural network denoiser trained at the output embeddings of the multimodal encoder.

2.2.1. *Training Objective*

We apply the same forward process (1) to each modality $\mathcal{M}$:

$$\mathbf{e}_s^{\mathcal{M}} = \sqrt{\bar{\alpha}_s}\,\mathbf{e}_0^{\mathcal{M}} + \sqrt{1-\bar{\alpha}_s}\,\boldsymbol{\epsilon}_{\mathcal{M}}, \quad \boldsymbol{\epsilon}_{\mathcal{M}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{5}$$

The denoiser $\phi_\theta(\cdot, s)$ is trained under the sample-prediction formulation to map both noised text and audio embeddings to the *text* embedding distribution at $s = 0$. This yields the cross-sample prediction loss [20]:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}\Big[\underbrace{\|\mathbf{e}_0^t - \phi_\theta(\mathbf{e}_s^t, s)\|_2^2}_{\text{text}\to\text{text}} + \underbrace{\|\mathbf{e}_0^t - \phi_\theta(\mathbf{e}_s^a, s)\|_2^2}_{\text{audio}\to\text{text}} \Big], \tag{6}$$

where the first term enforces high-fidelity reconstruction of text-like embeddings, while the second term encourages audio embeddings toward the text distribution.

Furthermore, we introduce a batch-level topology loss to preserve the relative geometry of the text distribution. Let $\mathbf{X} = [\mathbf{e}_{0,i}^t]_{i\in\mathcal{B}}$ and $\hat{\mathbf{X}} = [\hat{\mathbf{e}}_i^t]_{i\in\mathcal{B}}$ denote the text and text-like embedding matrices. Row-wise ℓ_2-normalized matrices $\mathbf{X}'$ and $\hat{\mathbf{X}}'$ are obtained from $\mathbf{X}$ and $\hat{\mathbf{X}}$, respectively, yielding similarity matrices $\mathbf{S}_{xx} = \mathbf{X}'\mathbf{X}'^\top \in \mathbb{R}^{\mathcal{B}\times\mathcal{B}}$ and $\mathbf{S}_{x\hat{x}} = \mathbf{X}'\hat{\mathbf{X}}'^\top \in \mathbb{R}^{\mathcal{B}\times\mathcal{B}}$, and the topology loss is the squared Frobenius distance:

$$\mathcal{L}_{\text{topo}} = \|\mathbf{S}_{xx} - \mathbf{S}_{x\hat{x}}\|_F^2. \tag{7}$$

The total training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{topo}}. \tag{8}$$

2.2.2. *Inference to generate text-like embedding*

At inference, given $\mathbf{e}^{\mathcal{M}}$, we optionally apply forward noising at step s_* with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and then run the learned reverse trajectory to

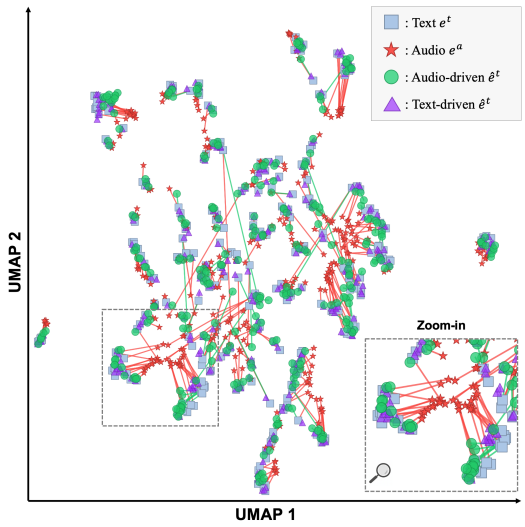

Fig. 2: Visualization of embeddings on AudioCaps using UMAP. Red line means the pair of audio and text embeddings. Green line means the pair of text-like and original text embeddings.

$s=0$ using DDIM sampler [16]:

$$\mathbf{e}_{s_*}^{\mathcal{M}} = \sqrt{\bar{\alpha}_{s_*}}\,\mathbf{e}^{\mathcal{M}} + \sqrt{1-\bar{\alpha}_{s_*}}\,\boldsymbol{\epsilon}, \quad \textit{(Forward)} \tag{9}$$

$$\hat{\mathbf{e}}^t = \mathrm{DDIM}\big(\phi_\theta, \mathbf{e}_{s_*}^{\mathcal{M}},\, s_* \to 0\big), \quad \textit{(Reverse)} \tag{10}$$

The output $\hat{\mathbf{e}}^t$ is a *text-like* embedding.

2.3. LLM-based Text Decoding

Given a text-like embedding $\hat{\mathbf{e}}^t$, a projection head maps it to a soft-prefix vector $\mathbf{p} \in \mathbb{R}^{mh}$. Here m denotes the number of learnable soft tokens and h denotes the decoder hidden size. We then project $\mathbf{p}$ into soft-prefix tokens sequence $\mathbf{p} \in \mathbb{R}^{m\times h}$ and feed this sequence to the decoder. If we optionally prepend a fixed instruction prompt of n tokens, the resulting input becomes $\mathbf{p} \in \mathbb{R}^{(m+n)\times h}$.

We consider two training options for our LLM decoder framework: **(i) text-only training**: using text-driven $\hat{\mathbf{e}}^t$ with an instruction prompt. **(ii) fully supervised training**: using audio-driven $\hat{\mathbf{e}}^t$ without an instruction prompt. We train the LLM decoder with the standard autoregressive cross-entropy objective

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{l=1}^{L} \log p_\psi\big(w_l | w_{<l}, \mathbf{p}\big), \tag{11}$$

where ψ denotes the LLM decoder's learnable parameters and $\mathbf{w} = (w_1, \ldots, w_L)$ means the target caption tokens. At inference, given audio data only with optional instruction prompt, and decode target caption. When the LLM decoder is trained under the text-only training, this evaluation corresponds to zero-shot captioning.

3. EXPERIMENTAL SETTINGS

3.1. Datasets

For training and evaluation, we conduct all experiments on AudioCaps [22], a corpus of ten-second audio clips paired with human-written captions. We use 48,595 training clips and 944 test clips. The *train* split provides one caption per clip, whereas the *test* split provides up to five. All audio is resampled to 48kHz. For audio preprocessing, we compute STFTs with a 1,024 window size and a 480 hop length, and then form mel-spectrograms with 64 mel-bins. We train on the *train* split and report results on the *test* split.

Table 1: Average cosine similarity scores for various embedding pairs on AudioCaps. For the CLAP, no transformation is applied, so $\hat{e}^{\mathcal{M}} = e^{\mathcal{M}}$. We report $e^t \cdot \hat{e}^{t\leftarrow\mathcal{M}}$, where $\hat{e}^{t\leftarrow\mathcal{M}}$ denotes a text-like embedding generated from modality $\mathcal{M}$. Transformations are obtained via C3 [11], DB (Diffusion-Bridge) [21], DG (DiffGap) [23], and our DL (Diffusion-Link). Here, $\sim$ and $\not\sim$ indicate matched and non-matched pairs, respectively.

Comparison Pair	Cosine Similarity				
	CLAP	C3	DB	DG	DL
(a) $e^t \cdot \hat{e}^{t\leftarrow a}(\sim)$	0.486	0.547	0.528	0.110	**0.688**
(b) $e^t \cdot \hat{e}^{t\leftarrow t}(\sim)$	**1.000**	1.000	0.999	0.334	0.945
(c) $e^t \cdot \hat{e}^{t\leftarrow a}(\not\sim)$	0.030	0.092	**0.000**	0.007	**0.000**
(d) $e^t \cdot \hat{e}^{t\leftarrow t}(\not\sim)$	0.098	0.158	0.002	0.043	**0.001**

Table 2: Average cosine similarity scores for various inference forward timestep s_* during diffusion process on AudioCaps.

Diffusion-Link	Inference forward timestep s_*				
	100	200	300	400	500
Cosine Similarity	**0.688**	0.654	0.596	0.510	0.404

3.2. Implementation Details and Metrics

For audio-language multimodal encoder, we use the LAION-CLAP pretrained model [2] and keep it frozen. Following prior work [21], we apply the same normalization process to the output embeddings of CLAP. For Diffusion-Link, we adopt three residual MLP blocks [32]. We train Diffusion-Link with the Adam [33] optimizer and a batch size of 128. The base learning rate is set to 1×10^{-4} and follows a step-decay schedule, multiplying the rate by 0.97 every 200 steps. We employ an exponential moving average (EMA) of the model parameters with a decay of 0.995 and use the EMA weights for inference. We adopt a cosine noise schedule with a total of T=1000 timesteps. At inference, we employ DDIM [16] sampling with 5 iteration steps. Before denoising, we apply a shallow forward noising to s_*=100 and then run the reverse process. For LLM-based text decoder, we adopt LLaMA2(7B) [5] as the LLM decoder. In the text-only training, we employ a linear layer with soft prefix tokens m=1 for the project head and prepend a short instruction prompt; in the fully supervised training, we use 2 linear layers with m=10 for the project head and no hard prompt. We fine-tune project head and the LLM using LoRA [34]. LLM training uses AdamW [35] optimizer with batch size 4 for 50 epochs: the learning rate warms up over the first 2 epochs with max learning rate 5×10^{-6}, then use a cosine decay. We also train a baseline multimodal LLM system to verify the effectiveness of Diffusion-Link, we adopt same setting but detach only Diffusion-Link module. For evaluation, we adopt the metrics for modality gap analyzing, including cosine similarity and visualization using UMAP [36]. For AAC, we use the metrics, METEOR (ME) [37], CIDEr (CD) [38], SPICE (SP) [39], and SPIDEr (SD) [40].

4. RESULTS

4.1. Main Results

Effectiveness of Diffusion-Link for Modality Bridging. As shown in Table 1, Diffusion-Link attains the highest cosine similarity on

Table 3: Performance comparison of AAC models on AudioCaps. **External knowledge #** is the number of non-audio samples used by the LLM at test time. For a fair comparison on the embedding-level modality-gap problem, † results use only embedding-level RAG without external k-caption selection.

Method	Encoder output dim.	External knowledge #	ME↑	CD↑	SP↑	SD↑
Zero-shot Captioning						
ZerAuCap [24]	$1 \times D$	527	12.3	28.1	8.6	18.3
DRCap† [25]	$1 \times D$	450,000	21.8	59.5	15.7	37.6
Zhang *et al.* [26]	$1 \times D$	No	22.0	64.4	15.6	40.0
WSAC [27]	$1 \times D$	46,000	24.1	63.3	17.3	40.3
Ours	$1 \times D$	No	**24.2**	**73.2**	**17.5**	**45.4**
Fully Supervised Captioning						
Prefix AAC [28]	$T \times D$	No	24.0	73.3	17.7	45.5
RECAP [29]	$T \times D$	600,000	25.6	75.1	18.6	47.1
EnCLAP-large [30]	$T \times D$	No	25.5	80.3	18.8	49.5
CLAP-ART [31]	$T \times D$	No	25.6	80.7	18.8	49.8
Ours	$1 \times D$	No	**25.6**	**82.5**	**18.9**	**50.7**

Table 4: Ablation study to analyze the effectiveness of diffusion-based modality bridging method.

Method	ME↑	CD↑	SP↑	SD↑
Zero-shot Captioning				
Baseline (CLAP & LLaMa2-7B)	21.2	48.0	14.4	31.2
+ Diffusion-Bridge [21]	23.3	62.6	16.5	39.5
+ **Diffusion-Link (Ours)**	**24.2**	**73.2**	**17.5**	**45.4**
Fully Supervised Captioning				
Baseline (CLAP & LLaMa2-7B)	25.0	76.9	18.6	47.7
+ Diffusion-Bridge [21]	25.2	77.1	18.0	47.4
+ **Diffusion-Link (Ours)**	**25.6**	**82.5**	**18.9**	**50.7**

matched audio–text pairs. While most approaches improve over CLAP, DiffGap underperforms, because it generates from pure Gaussian noise with the input embedding condition, which weakens information reconstruction. By contrast, Diffusion-Link treats the input embedding as residing at an intermediate reverse step, thereby minimizing information loss and ensuring high-quality generation along the reverse trajectory. Importantly, Diffusion-Link also yields the lowest similarity on non-matched pairs, indicating not merely a global contraction of the space but maintaining semantic information. Figure 2 visualizes this effect. Both the generated text-like embeddings from audio and text embeddings all move toward the ground-truth text embedding distribution, demonstrating that Diffusion-Link has learned a stable generative modality bridge for the text embedding distribution, regardless of the input modality.

Diffusion-Link Amplifies Multimodal Encoder-LLM Coupling.

Table 3 compares a range of AAC systems. In contrast to many prior methods that leverage longer audio representations or external knowledge (e.g., RAG), our multimodal LLM system captures input audio feature with only a single $1 \times D$ text-like embedding produced by Diffusion-Link, and achieves SOTA in both zero-shot and fully supervised captioning *without external knowledge*. Notably, considering that most prior zero-shot models rely heavily on external knowledge, outperforming them without any external knowledge demonstrates the significant efficiency of Diffusion-Link.

According to Table 4, our baseline LLM-based AAC system is not competitive relative to prior AAC systems in Table 3. However, applying Diffusion-Link markedly improves the same backbone. In zero-shot audio captioning, we observe a **52.5**% relative increase in CIDEr together with substantial gains on the other metrics. These dramatic gains demonstrate that Diffusion-Link is the key factor and reaffirm the primacy of *modality-gap reduction* over using longer audio representations or external knowledge. Moreover, in fully supervised audio captioning we observe up to **7.3**% relative improvement, underscoring our method's applicability.

4.2. Ablation Studies

We conduct ablations to analyze how the depth of forward noising affects modality bridging and high-quality generation. According to Table 2, increasing the inference forward timestep s_* from shallow levels initially keeps similarity quite stable; beyond a threshold, the similarity drops sharply as s_* increases. This indicates that over-noising pushes representations deeper into the common Gaussian space and *erases information*, thereby degrading semantic preservation in the reconstructed text-like embeddings.

This finding is consistent in AAC results. In Table 1, the similarity score of Diffusion-Bridge is similar to that of Diffusion-Link when s_* is between 300 and 400. This suggests that the performance of Diffusion-Bridge corresponds to over-noising of Diffusion-Link, which aligns with the observed *semantic information loss*. Furthermore, under the same multimodal LLM system, the AAC results in Table 4 follow the same pattern: attaching Diffusion-Link yields large gains, whereas using Diffusion-Bridge provides only limited improvements. Together, the three tables show that *excessive* forward noising reduces similarity and weakens bridging, which in turn harms downstream performance; conversely, choosing an appropriate s_* maximizes content preservation in the text-like embedding, strengthens conditioning-distribution alignment for the LLM decoder, and translates into AAC gains.

5. CONCLUSION

We introduced **Diffusion-Link**, a lightweight residual MLP diffusion module that bridges audio embeddings to the text embedding distribution without keeping the multimodal encoder frozen. The method aligns the conditioning input by increasing matched similarity and decreasing mismatched similarity. On AAC, it improves the same multimodal LLM baseline by **52.5%** and **7.3%** without external knowledge for zero-shot and fully supervised AAC, respectively. This plug-and-play approach of Diffusion-Link is expected to generalize beyond audio captioning and enable effective zero-shot performance in a variety of multimodal LLMs.

Acknowledgment

This work was supported by IITP grants funded by the Korean government (MSIT, RS-2025-02263169, Detection and Prediction of Emerging and Undiscovered Voice Phishing).

References

[1] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. ICASSP*, 2023.

[2] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*, 2023.

[3] Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha, "Reclap: Improving zero shot audio classification by describing sounds," in *Proc. ICASSP*, 2025.

[4] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al., "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, pp. 6, 2023.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al., "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, 2024.

[7] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang, "SALMONN: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[8] Kyeongha Rho, Hyeongkeun Lee, Valentio Iverson, and Joon Son Chung, "Lavcap: Llm-based audio-visual captioning using optimal transport," in *Proc. ICASSP*, 2025.

[9] Hyeongkeun Lee, Jongmin Choi, KiHyun Nam, and Joon Son Chung, "Lamb: Llm-based audio captioning with modality gap bridging via cauchy-schwarz divergence," *arXiv preprint arXiv:2601.04658*, 2026.

[10] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *NeurIPS*, 2022.

[11] Yuhui Zhang, Elaine Sui, and Serena Yeung, "Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data," in *Proc. ICLR*, 2024.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.

[13] Zehan Wang, Yang Zhao, et al., "Connecting multi-modal contrastive representations," in *NeurIPS*, 2023.

[14] Ziang Zhang, Zehan Wang, Luping Liu, Rongjie Huang, Xize Cheng, Zhenhui Ye, Wang Lin, Huadai Liu, Haifeng Huang, Yang Zhao, Tao Jin, Siqi Zheng, and Zhou Zhao, "Extending multi-modal contrastive representations," in *NeurIPS*, 2024.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[16] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," in *Proc. ICLR*, 2021.

[17] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, pp. 3, 2022.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022.

[19] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley, "Audioldm: text-to-audio generation with latent diffusion models," in *Proc. ICML*, 2023.

[20] Kihyun Nam, Jungwoo Heo, Jee weon Jung, Gangin Park, Chaeyoung Jung, Ha-Jin Yu, and Joon Son Chung, "SEED: Speaker Embedding Enhancement Diffusion Model," in *Proc. Interspeech*, 2025.

[21] Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang, "Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning," in *Proc. CVPR*, 2025.

[22] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL-HLT*, 2019.

[23] Shentong Mo, Zehua Chen, Fan Bao, and Jun Zhu, "Diffgap: A lightweight diffusion module in contrastive space for bridging cross-model gap," in *Proc. ICASSP*, 2025.

[24] Leonard Salewski, Stefan Fauth, A Koepke, and Zeynep Akata, "Zero-shot audio captioning with audio-language model guidance and audio context keywords," in *Proc. NeurIPS ML4Audio Workshop*, 2023.

[25] Xiquan Li, Wenxi Chen, Ziyang Ma, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Qiuqiang Kong, and Xie Chen, "Drcap: Decoding clap latents with retrieval-augmented generation for zero-shot audio captioning," in *Proc. ICASSP*, 2025.

[26] Yiming Zhang, Xuenan Xu, Ruoyi Du, Haohe Liu, Yuan Dong, Zheng-Hua Tan, Wenwu Wang, and Zhanyu Ma, "Zero-shot audio captioning using soft and hard prompts," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2025.

[27] Theodoros Kouzelis and Vassilis Katsouros, "Weakly-supervised automated audio captioning via text only training," in *Proc. DCASE Workshop*, 2023.

[28] Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh, "Prefix tuning for automated audio captioning," in *Proc. ICASSP*, 2023.

[29] Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Ramani Duraiswami, and Dinesh Manocha, "Recap: Retrieval-augmented audio captioning," in *Proc. ICASSP*, 2024.

[30] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo, "EnCLAP: Combining neural audio codec and audio-text joint embedding for automated audio captioning," in *Proc. ICASSP*, 2024.

[31] Daiki Takeuchi, Binh Thien Nguyen, Masahiro Yasuda, Yasunori Ohishi, Daisuke Niizumi, and Noboru Harada, "Clapart: Automated audio captioning with semantic-rich audio representation tokenizer," *arXiv preprint arXiv:2506.00800*, 2025.

[32] Tianhong Li, Dina Katabi, and Kaiming He, "Return of unconditional generation: A self-supervised representation generation method," in *NeurIPS*, 2024.

[33] Diederik P Kingma, Jimmy Ba, Y Bengio, and Y LeCun, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

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

[35] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[36] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, 2018.

[37] Satanjeev Banerjee and Alon Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *Proc. CVPR*, 2015.

[39] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. ECCV*, 2016.

[40] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proc. ICCV*, 2017.