From Pixels to Voice: A Simple and Efficient End-to-End Spoken Image Description Approach via Vision Codec Language Models

Chung Tran
Graduate School of Science and Technology
Nara Institute of Science and Technology
Ikoma, Japan
tran.quang_chung.tq9@naist.ac.jp

Sakriani Sakti
Graduate School of Science and Technology
Nara Institute of Science and Technology
Ikoma, Japan
ssakti@is.naist.jp

***Abstract*—Neural audio codecs provide a powerful tool for compressing audio signals into discrete codec representations. This compact discrete representation has made it possible to successfully apply a natural language processing (NLP) model to various audio and speech processing tasks, including text-to-speech (e.g., VALL-E, VALL-E X) and multimodal audio-text generation (e.g., LauraGPT, VioLA). While these models excel at handling sequential data like text and speech, their potential for processing non-sequential data, such as images, remains unexplored. In this paper, we introduce PixVoxLM, a simple and efficient end-to-end framework that combines vision-language models with neural audio codecs to tackle the Image-to-Speech (I2S) problem. Experiments on the Flickr8k dataset demonstrate that PixVoxLM delivers promising results compared to existing I2S methods. Furthermore, this research is the first to explore a new capability: visual-guided speech completion in I2S model, paving the way for new practical applications in everyday communication, such as speech prompt-based instruction.**

***Index Terms*—Spoken Image Description, Vision-Codec Language Models, Neural audio codecs**

I. Introduction

Large language models like ChatGPT [1] and LLaMA [2] have brought breakthroughs to the field of NLP. These impressive advances have been driven by the application of the Transformer with attention mechanisms [3], a powerful model for tasks such as text understanding and text generation.

The success of generative models in NLP has also extended to various speech processing tasks. Specifically, by using quantization models like Encodec [4] to convert continuous speech signals into discrete codes and then reconstruct them, many traditional speech processing pipelines have been re-defined as conditional codec language modeling problems. For example, while traditional Text-to-Speech (TTS) models use Mel-spectrograms as intermediate representations, VALL-E [5] and VALL-E X [6] offer an innovative solution by using generative models to produce discrete codes from phonemes and then convert these discrete codes back into speech using the Encodec model (phoneme $\mapsto$ discrete code $\mapsto$ waveform). Furthermore, unified decoder models like VioLA [7] and LauraGPT [8] have extended this concept further by performing multiple tasks such as audio generation, speech generation, and speech translation.

While these models excel in processing sequential inputs such as text or speech, the potential of neural audio codecs for handling non-sequential data, like images, remains underexplored. This limitation is significant, especially when considering the inherently multimodal nature of human communication, which involves not only text and speech but also visual information. Therefore, the integration of visual data processing with models is essential. Several studies have made significant advances in developing vision-language models like BLIP [9] and ALIGN [10]. Unfortunately, these models heavily rely on text-based inputs, which poses challenges for many languages that lack a written form [11].

Recent studies have proposed models that bypass text, but they often rely on joint training of multiple components, resulting in increased complexity. For example, SAT [12] and Im2Sp [13] involve training Image-to-Unit (I2U) and Unit-to-Speech (U2S) models, while E-I2S [14] trains I2U with VQ-VAE. To address this limitation, our research introduces PixVoxLM, a novel Image-to-Speech (I2S) pipeline that enables end-to-end training with a single component (I2U). To accomplish this, we use an Encodec [4] model to convert audio into discrete codes, and then employ a vision-language model to learn the mapping between images and the discrete codes. This is the first study to explore the potential of vision-language models with neural audio codecs, which opens up significant opportunities for unwritten languages as well as practical applications, such as spoken image descriptions for individuals with visual impairment. Additionally, this research introduces a novel capability in the I2S model—visual-guided speech completion. This feature has the potential to enable various applications, including speech prompt-based instruction based on image content.

Our contributions are as follows:

- To the best of our knowledge, this is the first study to explore the potential of neural audio codecs using images as input.

- We introduce PixVoxLM, a simple and efficient end-to-end framework designed specifically for the I2S task.
- This is also the first study to explore a novel capability of the I2S model for visual-guided speech completion.

II. Related Works

One notable breakthrough in I2S research was introduced by Wei-Ning Hsu [12], who developed a method to convert images to speech without using text. This approach used the pre-trained ResDAVEnet-VQ [15] model to extract speech units, and it trained two separate models: I2U and U2S. This multi-model approach was further explored by Minsu Kim *et al.* [13], who fine-tuned a pre-trained image captioning model to enhance I2U performance, and Johannes *et al.* [14] , who proposed a pipeline that trained both VQ-VAE [16], [17] and I2U models simultaneously. However, these multi-model approaches present two major challenges: joint training of multiple components and the need to retrain all models when speech-unit representations change. Xin-sheng Wang *et al.* [18] addressed these issues with the Show and Speak (SAS) model, a modified Tacotron2 [19] architecture that uses image features extracted by a pre-trained Faster-RCNN [20] to synthesize mel-spectrograms, which are then converted into speech using a pre-trained neural vocoder, WavGlow. However, this end-to-end approach struggles with performance due to the limitations of Faster R-CNN, including missed detections, misidentifications, challenges with overlapping objects, and inadequate contextual understanding.

Unlike previous work, ours is the first to use an off-the-shelf audio codec model to extract discrete representations and reconstruct them into speech. This simplifies I2S training by focusing exclusively on the vision-language model. Additionally, the model uses a vision transformer to learn image features end-to-end, reducing the need for external or hand-crafted feature extraction. Experiments on the Flickr8k [21] dataset show that our model is easier to train and infer while also delivering promising results compared to existing I2S methods. Furthermore, this research pioneers the exploration of a new capability of the I2S model: visual-guided speech completion. This capability allows the model to synthesize speech based on both speech prompts and image inputs.

III. Proposed Method

A. Problem Formulation

To train an end-to-end I2S model, we formulate the problem as follows, Fig. 1:

1) Speech Encoding and Reconstruction: To compress speech into an intermediate discrete representation, we employ an off-the-shelf neural Encodec [4] model that adheres to the following conditions:

$$U = \text{Encodec-Enc}(S), \tag{1}$$

$$\hat{S} = \text{Encodec-Dec}(U), \tag{2}$$

where S represents the input speech, U denotes the discrete unit representation (or discrete codebook representation), and

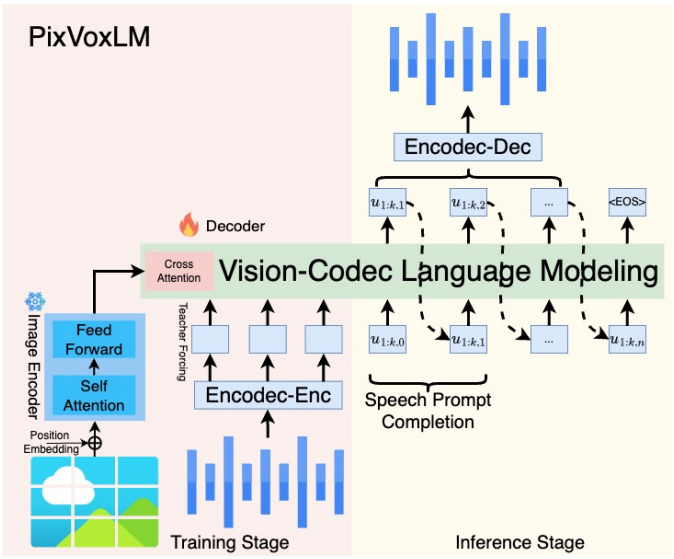

Fig. 1. Overview of PixVoxLM: novel, simple, and efficient I2S pipeline

$\hat{S}$ indicates the reconstructed speech derived from the U representation.

2) Image-to-Unit Mapping: The model processes an input image I to generate an output $\hat{U}$. Initially, it transforms the image into a visual hidden feature using an Image Encoder Transformer. Subsequently, the Unit Decoder Transformer adopts this feature to produce the final output $\hat{U}$.

$$\hat{U} = \text{I2U-Transformation}(I). \tag{3}$$

After the model has learned to transform the input image I into the discrete representation $\hat{U}$: $P(\hat{U}|I)$, the next step involves converting $\hat{U}$ into speech. This is accomplished using the Encodec-Dec model $P(\hat{S}|\hat{U})$, as described in Equation (2).

B. Neural Encodec Model

In this study, we use EnCodec [4], an advanced convolutional autoencoder specifically designed for audio tokenization, to convert continuous audio signals into discrete representations. EnCodec employs Residual Vector Quantization (RVQ) to quantize the latent space into multiple hierarchical codebooks. This multi-codebook quantization enables the model to effectively capture various details within the audio signal, preserving both fine-grained and broader features throughout the encoding process. By transforming audio into discrete tokens, EnCodec achieves efficient compression while maintaining a high level of fidelity in the reconstructed audio output, making it well-suited for downstream tasks that depend on discrete audio tokens.

Specifically, EnCodec takes an audio input S and compresses it into a lower-dimensional feature vector representation. These vectors are then quantized into integer vectors in the RVQ layer. Ultimately, this process produces an integer matrix representing the audio, denoted as $U = [u_{kn}]$, where k represents the number of codebooks, which is four in our

study, and n is the frame step with a sequence length of N. In this matrix, $U[:, n]$ contains the four integer codes corresponding to frame step n, while each column $U[k, :]$ reflects the length after quantization, with $k \in [1:4]$.

C. Vision-Codec Language Model

1) Image Encoder: The image encoder begins by resizing the input image to a fixed resolution and then divides it into smaller patches. Each patch is then transformed into a 1D vector, and positional embeddings are added to preserve the spatial relationships among the patches. These patch embeddings are processed through multiple Transformer layers, which include components such as self-attention and feed-forward layers. Finally, the image encoder produces a high-level visual feature that serves as input for the decoding stage.

2) Unit Decoder: The decoder side consists of multiple identical transformer layers, each comprising self-attention, cross-attention, and feed-forward layers. The cross-attention layer combines high-level visual feature with discrete unit feature embeddings, allowing the model to learn the relationships between them effectively. The output of the last decoder layer is then used to predict the discrete unit sequence, corresponding to the image input.

3) Objective function: The objective function for the vision-codec language modeling focuses on minimizing the cross-entropy loss between the predicted unit sequence and the ground truth labels across multiple codebooks. The total loss $\mathcal{L}$ can be simplified as:

$$\mathcal{L} = \sum_{i=1}^{K} \sum_{n=1}^{N} -\hat{u} \log p(\hat{u}|u), \tag{4}$$

where K is the number of codebooks, N is the unit sequence length, and $p(\hat{u}|u)$ is the predicted probability of the true token $\hat{u}$ given the input u.

D. Codebook Patterns

This study employs two codebook representations: the parallel pattern and the delay pattern, both inspired by the research done for MusicGen [22] research and illustrated in Fig. 2. In the parallel pattern, each frame step n contains four integer codebooks, $U[:, n]$, with a sequence length of N.

On the other hand, the delay pattern shifts each codebook one step to the right, filling the resulting empty positions with zeros. This staggering of codebooks introduces temporal dependencies, enhancing the model's ability to capture sequential relationships. Note that during inference, the delay pattern must be converted back into the parallel pattern.

IV. EXPERIMENT SETUP

A. Dataset

In this experiment, we adopt the Flickr8k dataset [21], which is tailored for spoken captioning research. This dataset contains 8,000 images, each paired with five spoken captions, focusing on diverse scenes and everyday situations. While the original dataset includes recordings from over 100 different speakers, we reuse speech generated in the SAS study [18], produced by a TTS model with a single speaker. The dataset is divided into three subsets: 6,000 images for training and 1,000 images each for validation and testing.

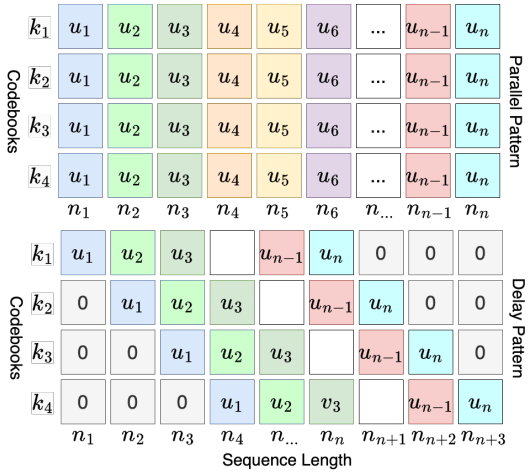

Fig. 2. Two methods of representing the codebook: parallel and delay pattern.

B. Implementation Details

We employed a pre-trained BLIP [9] architecture for the Vision-Codec language modeling. To optimize training efficiency and preserve the model's image understanding capabilities, we froze the image encoder transformer and focused on fine-tuning the token language model. The trainable parameters total 125 M out of the 211 M in the entire model. We used 100 epochs, the AdamW [23] optimizer with a learning rate of 5e-5, and a batch size of 60. Early stopping was implemented to avoid overfitting if the validation loss increased, and selected the checkpoint with the lowest validation loss to assess the performance of the I2S model.

For the quantization model, we utilized a pre-trained 24-kHz Encodec model[1], trained on a diverse array of data, including speech, audio, and music. We employed a bandwidth of 3 kbps, corresponding to a quantization codebook size K of 4.

C. Evaluation metrics

Evaluating the generated content of I2S models directly by human assessment is a subjective and labor-intensive task. To tackle this challenge, most studies use ASR models trained on large datasets to transcribe speech into text and apply metrics similar to those in image captioning. In this study, we employ bilingual evaluation understudy BLEU (B) [24] with four n-grams (B1, B2, B3, and B4) and other machine translation metrics such as METEOR (M) [25], ROUGE-L (R) [26], and CIDEr (C) [27], similar to metrics used in prior work: SAS [18]. We compare the ASR model's output with five ground truths and calculate scores using the acc-metric library [28], where higher scores indicate better caption quality. Although these scores depend on the ASR model, we treat it as an

[1]https://github.com/facebookresearch/encodec

TABLE I
PERFORMANCE COMPARISON OF PIXVOXLM WITH EXISTING I2S MODELS ACROSS VARIOUS EVALUATION METRICS

Methods	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑
Multiple-Model Training							
SAT [12]	-	-	-	11.60	14.10	39.00	23.20
SAT-FT [12]	-	-	-	12.60	14.50	39.10	24.20
E-I2S [14]	-	-	-	14.78	17.40	45.75	32.89
Single-Model Training							
SAS [18]	29.60	14.70	7.20	3.50	11.30	23.20	8.00
PixVoxLM-Parallel	34.52	18.75	10.65	6.22	10.51	26.30	9.43
PixVoxLM-Delay	48.08	30.59	18.92	11.49	15.19	35.76	25.54

independent component of the I2S model, allowing us to focus on improving the I2S model's quality. The ASR model used is a pre-trained Wav2Vec2 architecture, trained on 960 hours of Libri-Light and Librispeech.

V. RESULTS AND DISCUSSION

A. Codebook Patterns and Model Comparison

1) Codebook Patterns: The results in Table I illustrate the superior performance of the delay pattern in the PixVoxLM model compared to the parallel pattern. Specifically, the delay pattern achieves higher BLEU scores, such as B3 (18.92 vs. 10.65) and B4 (11.49 vs. 6.22). Additionally, improvements in the M, R and C scores, with increases of 15.19, 35.76 and 25.54 respectively, highlight the delay pattern's effectiveness in capturing temporal dependencies crucial for generating coherent and contextually accurate speech from visual data.

2) Model Comparison: The baselines we selected—SAT [12], E-I2S5 [14], and SAS [18]—represent the most recent methods available that exclusively use Flickr8k. Among these, SAS serves as the primary end-to-end baseline and represents the standard for comparison. PixVoxLM outperforms the end-to-end SAS model in all evaluation metrics. Its performance is competitive with multi-model training approaches like SAT and SAT-FT, achieving scores that are nearly equivalent in some cases. Notably, the PixVoxLM model with the delay pattern closely matches the SAT on the B4 metric (11.49 vs. 11.60) and exceeds SAT and SAT-FT in the M and C metrics. These results demonstrate that even with a single-model training process, PixVoxLM delivers competitive performance against models reliant on complex multi-model training architectures.

B. Visual-guided speech completion

We evaluate the speech completion capabilities of our models by providing speech prompts at varying levels of completeness. Specifically, we test three scenarios: 0% (no speech information), 25% (a quarter of the speech), and 50% (half of the speech). Results in Table II show that PixVoxLM-Delay consistently outperforms PixVoxLM-Parallel, especially as the information ratio increases. At 50%, PixVoxLM-Delay significantly improves B4, M, R, and C scores, demonstrating superior prompt completion ability.

TABLE II
SPEECH PROMPT COMPLETION AT VARIOUS INFORMATION LEVELS

PixVoxLM	Prompt	B1↑	B2↑	B3↑	B4↑	M↑	R↑	C↑
Parallel	0%	34.52	18.75	10.65	6.22	10.51	26.30	9.43
	25%	37.11	23.30	14.58	8.87	13.05	29.71	15.24
	50%	46.26	34.00	26.25	20.42	20.54	40.83	37.64
Delay	0%	48.08	30.59	18.92	11.49	15.19	35.76	25.54
	25%	49.76	34.18	23.04	14.90	18.00	39.04	32.68
	50%	57.31	44.3	35.31	28.11	24.19	48.35	60.10

GT: Two dogs play in the grass
ASR: Two dogs running in grass

GT: Three children playing in sand at beach
ASR: Thre children playing in the sand

GT: A man climbs icy rocks
ASR: Clamber or climbing a neste

Fig. 3. Sample of speech generated using our proposal, PixVoxLM with delay pattern. Note that the text transcript is generated using an off-the-shelf ASR model. Images courtesy of Flickr8k.

C. Subjective Results

We selected several images along with the text transcripts generated by the ASR model to conduct a deeper and more thorough analysis of the results produced by the PixVoxLM model with delay pattern. In Fig. 3-A, the generated output is understandable in terms of content; however, the absence of articles like "the" affects the overall score. In Fig. 3-B, the output is generally good, but the spelling error "thre" from the ASR model negatively impacts the overall quality. In Fig. 3-C, the output is completely unintelligible, even though some words from the ASR transcript appear in the Ground Truth (GT). These cases highlight the need to improve the I2S model's performance in the future.

VI. CONCLUSION

The end-to-end PixVoxLM framework offers a simple and efficient solution for generating speech directly from images without relying on text as an intermediate step. Experimental results on the Flickr8k dataset demonstrate that our model outperforms the recent end-to-end SAS model. Furthermore, we are the first to explore the model's capability for visual-guided speech completion. However, subjective evaluations highlight several issues, particularly the intelligibility of the generated speech. In future work, we will enhance the quality of the ITS model to improve performance and broaden its applicability across various datasets and real-world scenarios.

ACKNOWLEDGMENTS

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681.

REFERENCES

[1] OpenAI, "GPT-4 Technical Report," *CoRR*, vol. abs/2303.08774, 2023, arXiv: 2303.08774. [Online]. Available: https://doi.org/10.48550/arXiv.2303.08774

[2] A. Dubey, A. Jauhri, and et al., "The Llama 3 Herd of Models," *CoRR*, vol. abs/2407.21783, 2024, arXiv: 2407.21783. [Online]. Available: https://doi.org/10.48550/arXiv.2407.21783

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Trans. Mach. Learn. Res.*, 2023. [Online]. Available: https://openreview.net/forum?id=ivCd8z8zR2

[5] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," *CoRR*, vol. abs/2301.02111, 2023, arXiv: 2301.02111. [Online]. Available: https://doi.org/10.48550/arXiv.2301.02111

[6] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak Foreign Languages with Your Own Voice: Cross-Lingual Neural Codec Language Modeling," *CoRR*, vol. abs/2303.03926, 2023, arXiv: 2303.03926. [Online]. Available: https://doi.org/10.48550/arXiv.2303.03926

[7] T. Wang, L. Zhou, Z. Zhang, Y. Wu, S. Liu, Y. Gaur, Z. Chen, J. Li, and F. Wei, "Viola: Unified codec language models for speech recognition, synthesis, and translation," *arXiv preprint arXiv:2305.16107*, 2023.

[8] J. Wang, Z. Du, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma, W. Wang, S. Zheng, C. Zhou, Z. Yan, and S. Zhang, "LauraGPT: Listen, Attend, Understand, and Regenerate Audio with GPT," 2024. [Online]. Available: https://openreview.net/forum?id=jDy2Djjrge

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022.

[10] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021.

[11] M. P. Lewis, Fennig, and G. F. Simon, "Ethnologue: Languages of the world," 2016. [Online]. Available: http://www.ethnologue.com, 2016

[12] W.-N. Hsu, D. Harwath, T. Miller, C. Song, and J. R. Glass, "Text-Free Image-to-Speech Synthesis Using Learned Segmental Units," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*.

[13] M. Kim, J. Choi, S. Maiti, J. H. Yeo, S. Watanabe, and Y. M. Ro, "Towards practical and efficient image-to-speech captioning with vision-language pre-training and multi-modal tokens," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[14] J. Effendi, S. Sakti, and S. Nakamura, "End-to-end image-to-speech generation for untranscribed unknown languages," *IEEE Access*, 2021.

[15] D. Harwath, W.-N. Hsu, and J. R. Glass, "Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech," in *8th International Conference on Learning Representations, ICLR 2020*.

[16] A. Van Den Oord, O. Vinyals, and others, "Neural discrete representation learning," *Advances in neural information processing systems*, 2017.

[17] A. Tjandra, S. Sakti, and S. Nakamura, "Transformer VQ-VAE for Unsupervised Unit Discovery and Speech Synthesis: ZeroSpeech 2020 Challenge," 2020.

[18] X. Wang, S. Feng, J. Zhu, M. Hasegawa-Johnson, and O. Scharenborg, "Show and Speak: Directly Synthesize Spoken Description of Images," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*.

[19] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, and others, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, 2015.

[21] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015.

[22] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[25] S. Banerjee and A. 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, pp. 65–72.

[26] L. Chin-Yew, "Rouge: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*, 2004.

[27] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[28] E. Labbé, "aac-metrics," Mar. 2024. [Online]. Available: https://github.com/Labbeti/aac-metrics/