SpeechCoCa: Semantic Speech-to-Image Retrieval with Contrastive Captioner

Anonymous ACL submission

Abstract

Speech–image retrieval has emerged as a critical task for enabling intuitive, voice-driven access to large multimedia collections. Traditional approaches typically rely on automatic speech recognition (ASR) followed by text-based retrieval, which is prone to transcription errors and often fails to capture the full semantic intent of spoken queries. To address these challenges, we propose SpeechCoCa, a multi-modal model that directly aligns speech and visual data in a shared representation space. Our method leverages the weakly supervised nature of Whisper as a speech encoder together with the contrastive–generative capabilities of CoCa (Contrastive Captioners), enabling robust cross-modal alignment without relying on intermediate transcription. We demonstrate the effectiveness of SpeechCoCa through extensive experiments on two benchmark datasets. On Flickr8k, our model achieves state-of-the-art performance, surpassing prior methods by **+5.2%** Recall@1 in speech–image retrieval. On SpokenCOCO, SpeechCoCa achieves even larger gains, outperforming baselines by **+10.4%** Recall@1. These results highlight the potential of SpeechCoCa as a powerful framework for semantic voice-driven multimedia search, advancing both accessibility and natural human–computer interaction.

1 Introduction

The rapid growth of multimedia content ranging from digital images and instructional videos to podcasts and social media streams has created an urgent need for intuitive and efficient search technologies. Conventional search engines are largely optimized for text-based queries, which require users to provide keywords or manually transcribed descriptions of the content they seek. While effective in some domains, such approaches are not always natural or accessible, especially when the content itself is multi-modal and the user's preferred mode of interaction is voice.

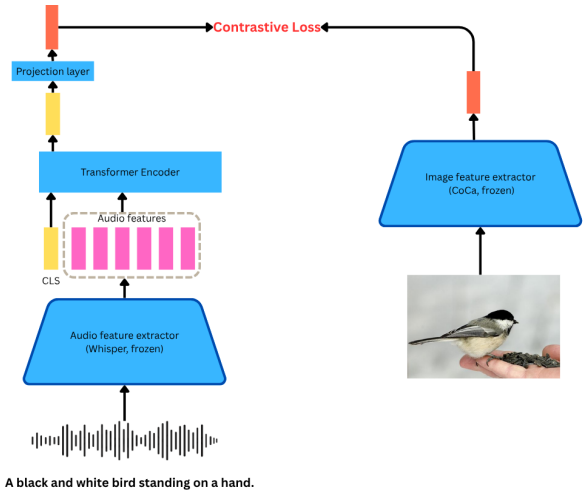

Figure 1: Overview of the SpeechCoCa architecture, utilizing a frozen Whisper encoder for speech, a frozen CoCa image encoder, and a learnable transformer block for cross-modal alignment.

Voice interfaces have become increasingly popular through virtual assistants and mobile applications, yet most systems still rely on automatic speech recognition (ASR) followed by keyword matching or text-based retrieval. This two-step pipeline introduces errors in transcription (Min et al., 2025), struggles with domain-specific vocabulary, and often fails to capture the semantic intent of the query. As a result, users face mismatches between spoken intent and retrieved multimedia items.

Recent advances in multi-modal representation learning have shown that shared embedding spaces can align text and visual data with remarkable accuracy. Models such as CLIP (Radford et al., 2021) demonstrated the strength of contrastive objective for cross-modal alignment, while CoCa (Yu et al., 2022) extended these methods by combining contrastive learning with auto-regressive captioning, yielding stronger semantic grounding. However, existing approaches primarily operate on text rather

than raw speech, limiting their applicability to truly voice-driven scenarios.

In this work, we propose SpeechCoCa, a model that bridges the gap between spoken language and visual data. Unlike traditional ASR-based pipelines, SpeechCoCa directly maps speech into a multi-modal space aligned with images and images. By adapting the CoCa model to process speech embeddings from Whisper (Radford et al., 2023) encoder, it creates a unified space where voice queries can retrieve relevant multimedia content without intermediate transcription.

Our main contributions are summarized as follows: We build upon the parallel architecture proposed in the SpeechCLIP (Shih et al., 2023), while introducing key modifications to improve cross-modal alignment. Specifically, we replace self-supervised audio encoders such as HuBERT (Hsu et al., 2021) with the Whisper encoder, leveraging its weakly supervised training for more robust speech representations. In addition, we adopt the CoCa image and text encoders instead of the widely used CLIP encoders, thereby benefiting from CoCa's generative capability in multi-modal modeling. Finally, we evaluate SpeechCoCa against SOTA approaches on the Flickr8k (Harwath and Glass, 2015) and SpokenCOCO (Hsu et al., 2020) benchmarks, demonstrating clear improvements in multi-modal retrieval performance.

2 Related Work

We briefly mention the pre-trained models and prior works most relevant to SpeechCoCa. **SpeechCLIP**. (Shih et al., 2023) was proposed as a framework for bridging speech and text through visual grounding. By leveraging pre-trained HuBERT (Hsu et al., 2021) for speech and CLIP (Radford et al., 2021) for vision–language alignment, SpeechCLIP aligns spoken captions with image representations without requiring transcriptions. It introduced both parallel and cascaded architectures, showing that even frozen pre-trained encoders can be enhanced by small contrastive modules. The model demonstrated competitive performance on Flickr8k and SpokenCOCO, and also showed the capability of zero-shot speech–text retrieval, highlighting the feasibility of using images as an intermediate bridge between speech and text.

Listening for "You". More recently, Yang et al. (Yang et al., 2025) addressed the problem of speech–image retrieval in multi-speaker scenarios by introducing the Target Speaker Speech-Image Retrieval task. Their framework incorporated a Target Speaker Extraction and Retrieval (TSRE) module, allowing the system to focus on a specific speaker within overlapping conversations. By conditioning on target speaker embeddings, their approach achieved substantial gains in Recall@1 on SpokenCOCO2Mix and SpokenCOCO3Mix compared to single-speaker baselines. This work is relevant to our setting as it also evaluates on the standard single-speaker retrieval mode, demonstrating competitive results; however, its main emphasis lies in handling multi-speaker mixtures, whereas our model focuses on improving retrieval accuracy in the conventional single-speaker setting.

CoCa. Contrastive Captioners (CoCa) (Yu et al., 2022) represented a breakthrough in multimodal modeling by combining two key objectives: contrastive learning for alignment and autoregressive captioning for generative grounding. Unlike CLIP, which relies solely on contrastive objectives, CoCa extends the training signal to predict captions, thereby enriching the semantic representations and improving transfer across modalities.

Whisper. Whisper (Radford et al., 2023) is a large-scale speech model trained in a weakly supervised manner. Crucially, Whisper's encoder produces semantic-rich speech embeddings that capture both acoustic detail and higher-level linguistic information, making them highly transferable to downstream applications. Its release in multiple model sizes (from small to large variants) allows researchers to balance accuracy with computational cost. For cross-modal retrieval, these properties make Whisper a natural choice as the speech encoder in SpeechCoCa, where robust and semantically meaningful representations are essential for aligning speech with vision.

3 Method

The SpeechCoCa approach extends the parallel architecture of SpeechCLIP (Shih et al., 2023), leveraging the strength of the dual-encoder design while modifying its components to achieve stronger cross-modal alignment. In particular, SpeechCoCa replaces the HuBERT encoder in SpeechCLIP with Whisper, a large-scale weakly supervised speech model that produces semantic-rich embeddings robust to noise and accents. On the visual side, SpeechCoCa adopts the image encoder from CoCa rather than CLIP, benefiting from CoCa's large-

scale pretraining with both contrastive and generative objectives.

As illustrated in Figure 1, SpeechCoCa follows a simple yet effective parallel design, where Whisper encodes spoken queries and CoCa's vision backbone encodes images, with both modalities projected into a shared semantic space. By combining Whisper and CoCa's vision encoder within this contrastive framework, SpeechCoCa inherits the scalability of SpeechCLIP while significantly improving semantic grounding. The speech and image encoders are trained jointly with a contrastive loss, projecting both modalities into a shared space. Unlike CoCa, which includes a multimodal decoder for caption generation, SpeechCoCa is purely contrastive, focusing entirely on retrieval improvement.

4 Experiments

4.1 Datasets

We evaluate SpeechCoCa on two widely used benchmarks for speech–image retrieval: the **Flickr8k Audio Captions Corpus** (Harwath and Glass, 2015) and **SpokenCOCO** (Hsu et al., 2020). Each image in both datasets is paired with five spoken captions produced by humans reading uttering text captions. Flickr8k consists of approximately 8,000 images and 46 hours of speech, providing a relatively small-scale but challenging benchmark. SpokenCOCO contains around 123K images and 742 hours of speech, making it a large-scale dataset for training and evaluation. Following SpeechCLIP, we adopt the Karpathy split for SpokenCOCO to ensure consistency and fair comparison across systems. For Flickr8k, we use the standard training, validation, and test partitions provided with the dataset.

4.2 Model

Our SpeechCoCa configuration consists of a transformer block with 4 layers, applied on top of the speech encoder to map its output into a shared embedding space with the visual encoder. we train with a per-device batch size of 168. Optimization is performed using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}. SpeechCoCa are trained on 2 X 80 GB NVIDIA A100 GPU and our largest experiment lasts 12 hours.

Method	Speech→Image			Image→Speech		
	R@1	R@5	R@10	R@1	R@5	R@10
Flickr8k						
MILAN (Sanabria et al., 2021)	33.2	62.7	73.9	49.6	79.2	87.5
SpeechCLIP (Shih et al., 2023)	39.1	72.0	83.0	54.5	84.5	93.2
SpeechCLIP+ (Wang et al., 2024)	41.7	73.7	84.1	54.2	86.8	94.2
CMD-SpeechCLIP (Zhou et al., 2024)	40.7	75.1	85.8	56.8	86.2	94.2
HuBERT (Yang et al., 2025)	51.3	82.1	90.2	69.5	93.2	96.8
WavLM (Yang et al., 2025)	52.1	81.8	90.1	71.2	93.7	97.6
SpeechCoCa* (ours)	**57.3**	**86.6**	**93.3**	**76.3**	**95.7**	**98.5**
SpokenCOCO						
SpeechCLIP (Shih et al., 2023)	35.8	66.5	78.0	50.6	80.9	89.1
SpeechCLIP+ (Wang et al., 2024)	36.5	66.3	77.9	51.0	80.0	88.5
CMD-SpeechCLIP (Zhou et al., 2024)	37.5	67.3	78.6	52.3	81.4	89.7
HuBERT (Yang et al., 2025)	37.8	66.9	78.0	55.1	82.7	90.0
WavLM (Yang et al., 2025)	38.0	67.7	78.5	58.5	83.5	90.9
SpeechCoCa* (ours)	**48.4**	**76.5**	**85.3**	**63.1**	**87.8**	**93.4**

Table 1: Speech↔Image retrieval results (Recall@K). Best scores in **bold**, second best in underline.

Method	Speech→Text			Text→Speech		
	R@1	R@5	R@10	R@1	R@5	R@10
Flickr8k						
SpeechCLIP (Shih et al., 2023)	19.6	44.1	58.5	22.5	44.1	54.5
SpeechCoCa* (ours)	**68.6**	**92.2**	**95.6**	**70.0**	**94.6**	**97.2**

Table 2: Speech↔Image retrieval results. These results are produced by substituting the image encoder with the text encoder, without further fine-tuning on speech–text pairs, to assess the inherent semantic correspondence between speech and text representations in SpeechCoCa.

4.3 Speech-image retrieval

We provide evaluation of SpeechCoCa on the task of speech-to-image retrieval, where the goal is to retrieve the correct image given a spoken caption query.

Table 1 summarizes our results on Flickr8k and SpokenCOCO. SpeechCoCa achieves consistent improvements over prior methods. On Flickr8k, our model improves Recall@1 by +5.2% compared to the best existing baseline. On SpokenCOCO, SpeechCoCa demonstrates even larger gains, outperforming previous methods by +10.4% Recall@1.

4.4 Speech-text retrieval

To further assess the semantic quality of the learned speech representations, we evaluate SpeechCoCa on the speech–text retrieval task as seen in Table 2. In this setting, a spoken caption is used as a query to retrieve its corresponding textual caption from a pool of candidates, and vice versa. This task measures how well the model captures linguistic meaning in speech embeddings, independent of visual information.

Method	Speech→Text			Text→Speech		
	R@1	R@5	R@10	R@1	R@5	R@10
HuBERT Large	43.9	76.1	86.4	59.6	90.8	96.3
Whisper Small (n = 4)	48.5	78.9	88.9	68.2	92.1	96.4
Whisper Large (n = 1)	**57.6**	85.6	93.0	74.2	95.5	98.4
Whisper Large (n = 4)	57.3	**86.6**	**93.3**	**76.3**	**95.7**	**98.5**

Table 3: Ablation results on Flickr8k showing the effect of varying the speech encoder while fixing the CoCa ViT-L visual backbone. n indicates the number of transformer layers in the projection block.

5 Ablation study

We perform an ablation study on the Flickr8k test set to isolate the effect of each major component in SpeechCoCa, focusing on the speech encoder choice and the number of transformer layers used in the projection module. All models use the same CoCa ViT-L image encoder, and only the audio-side configuration is varied. Results are summarized in Table 3. The ablation confirms that audio supervision and scale play the dominant role in performance. Transitioning from HuBERT to Whisper particularly from unsupervised to weakly supervised pretraining significantly narrows the modality gap, while projection depth has a smaller but consistent effect on alignment quality.

6 Conclusion

SpeechCoCa introduces a simple yet effective dual-encoder framework for bridging speech with visual and textual modalities. By integrating Whisper, a weakly supervised speech encoder, with the CoCa ViT-L vision backbone, the model achieves strong semantic alignment across modalities without requiring ASR or generative decoding. Experiments on Flickr8k and SpokenCOCO show consistent improvements in both speech-image and speech-text retrieval tasks, with Recall@1 gains reaching 10.4 % over existing approaches. The ablation analysis highlights that Whisper's semantic representations and the CoCa pretrained vision encoder are the main contributors to this performance, while the number of transformer layers in the projection block provides smaller but consistent benefits.

Overall, the results demonstrate that combining weakly supervised speech modeling with large-scale vision–language pretraining offers a powerful and efficient path toward robust cross-modal understanding.

7 Limitations

Although SpeechCoCa achieves strong retrieval performance, it still inherits several limitations from its underlying components. The reliance on large-scale pretrained backbones such as Whisper and CoCa introduces high computational and memory demands, making the model less practical for resource-constrained environments. While Whisper provides semantically rich and robust speech representations, its limited coverage of low-resource languages and accented speech may hinder multilingual generalization. Moreover, the current evaluation assumes single-speaker and relatively clean audio, leaving the model's robustness in multi-speaker or noisy conditions untested. Finally, SpeechCoCa is trained solely with a contrastive learning objective, which restricts it from leveraging complementary generative alignment signals that could further enhance the richness of its multimodal representations.

8 Future Work

Future research will aim to extend SpeechCoCa beyond static image retrieval toward spoken–video understanding and temporal alignment in audiovisual streams, enabling richer multimodal reasoning. Another promising direction involves exploring multilingual and cross-accent adaptation, leveraging Whisper's multilingual foundation to improve performance on under-represented languages and dialects. To make the model more accessible for real-world use, future work will also investigate lightweight and efficient architectures that maintain retrieval accuracy while reducing computational cost for on-device deployment. Finally, incorporating generative, such as speech captioning, could provide deeper semantic grounding and enhance cross-modal transfer capabilities.

References

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. In *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*, pages 237–244. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Wei-Ning Hsu, David Harwath, Christopher Song, and James Glass. 2020. Text-free image-to-speech synthesis using learned segmental units. *arXiv preprint arXiv:2012.15454*.

Do June Min, Karel Mundnich, Andy Lapastora, Erfan Soltanmohammadi, Srikanth Ronanki, and Kyu Han. 2025. Speech retrieval-augmented generation without automatic speech recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Ramon Sanabria, Austin Waters, and Jason Baldridge. 2021. Talk, don't write: A study of direct speech-based image retrieval. *arXiv preprint arXiv:2104.01894*.

Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath. 2023. Speechclip: Integrating speech with pre-trained vision and language model. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 715–722. IEEE.

Hsuan-Fu Wang, Yi-Jen Shih, Heng-Jui Chang, Layne Berry, Puyuan Peng, Hung-yi Lee, Hsin-Min Wang, and David Harwath. 2024. Speechclip+: Self-supervised multi-task representation learning for speech via clip and speech-image data. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 465–469. IEEE.

Wenhao Yang, Jianguo Wei, Wenhuan Lu, Xinyue Song, and Xianghu Yue. 2025. Listening for" you": Enhancing speech image retrieval via target speaker extraction. *arXiv preprint arXiv:2509.09306*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models.

Lifeng Zhou, Yuke Li, Rui Deng, Yuting Yang, and Haoqi Zhu. 2024. Cross-modal denoising: A novel training paradigm for enhancing speech-image retrieval. *arXiv preprint arXiv:2408.13705*.