Exploring Multi-Modal Large Language Models and Two-Stage Fine-Tuning for Fashion Image Retrieval

Nguyen Cao Hoang[1,2]*, Hoang Bui Le[1,2]*, Nam Vo Hoang[1,2]*, and Trung-Nghia Le[1,2]**

[1] University of Science, VNU-HCM, Ho Chi Minh, Vietnam
[2] Vietnam National University, Ho Chi Minh, Vietnam
{23125064, 23125057, 23125040}@student.hcmus.edu.vn
ltnghia@fit.hcmus.edu.vn

Abstract. Composed image retrieval retrieves a target image using a composed query of a reference image and a modified text description. In the fashion domain, this task requires understanding subtle attribute variations such as color, pattern, and texture. However, existing approaches face limitations due to scarce annotated data and simplistic negative sampling. We propose a novel framework that integrates a multi-modal large language model (LLaVA) to generate attribute-aware triplets and introduces a two-stage fine-tuning strategy to enhance contrastive learning. We leverage pretrained vision-language models, such as CLIP-ViT/B32, to generate and concatenate sentence-level prompts with the relative caption and to scale the number of negatives using static representations. Experimental results demonstrate enhanced compositional reasoning and improved fine-grained retrieval behavior, underscoring the feasibility and potential of the proposed framework for fashion retrieval.

Keywords: Composed image retrieval · Fashion image retrieval · Contrastive learning · Multi Large Language Model · LLaVa · Image Captioning · Fine-tuning

1 Introduction

Composed image retrieval (CIR) is a challenging retrieval task where a query consists of a reference image and a relative caption, aiming to locate a target image that reflects the described modifications while retaining visual similarity to the reference [4,12]. Within this paradigm, Fashion image retrieval (FIR) emerges as a specialized and fine-grained instance, tailored for fashion applications such as e-commerce, personalized shopping, and virtual try-on [12,13]. Unlike general CIR, FIR demands precise interpretation of subtle visual attributes, such as color tone, texture, pattern, and fit, based on detailed, multi-attribute user queries (e.g., "make this dress pastel blue with long sleeves and a floral pattern").

* These authors contributed equally to this work.
** Corresponding author.

Despite recent advances, FIR remains limited by shallow visual understanding and inefficient contrastive learning. Models such as CLIP [15], while powerful for general vision-language alignment, primarily capture global semantics and often overlook subtle, fine-grained attributes crucial in fashion, for instance, intricate lace textures or nuanced silhouette variations. Furthermore, the scarcity and ambiguity of annotated training data prevent contrastive objectives from effectively learning detailed visual distinctions. As a result: (i) models lack sufficient diverse positive examples, and (ii) the common practice of using random in-batch negatives fails to expose the model to genuinely hard, visually similar candidates.

To overcome these challenges, we present a framework that enhances visual representation learning through enriched image captioning and improved negative sampling. Specifically, we employ the multimodal LLM (LLaVA) [11] to generate high-quality, attribute-aware captions that enrich visual-textual alignment and mitigate data sparsity. Additionally, we introduce a two-stage fine-tuning strategy incorporating both coarse and hard-negative alignment to strengthen discriminative learning. Together, these components enable more robust, fine-grained feature representations for fashion retrieval. Experimental results on the FashionIQ dataset [22] demonstrate enhanced compositional reasoning and improved fine-grained retrieval behavior, underscoring the feasibility and potential of the proposed framework for fashion retrieval.

Our contributions are as follows:

- We employ LLaVA to generate high-quality, attribute-aware captions and triplets, enriching image-text alignment and mitigating the shortage of annotated examples.
- We design a two-stage fine-tuning strategy that combines coarse alignment with hard-negative sampling to strengthen discriminative learning and improve fine-grained retrieval accuracy.
- We integrate sentence-level prompting with relative captions using pretrained vision-language models (e.g., BLIP-2), enhancing compositional reasoning and interpretability in composed queries.

2 Related Work

Composed image retrieval (CIR) combines a query image with modifying text to retrieve target images, necessitating effective fusion of visual and textual modalities. Recent works leverage large-scale vision-language models: Liu et al. [12] introduced CIRPLANT using transformer-based adaptation, while Baldrati et al. [3,4] exploited CLIP for robust feature fusion. Xu et al. [24] proposed ComqueryFormer, a unified transformer architecture with global-local alignment, and Zhao et al. [25] incorporated progressive learning with adaptive weighting for hybrid queries. Additional innovations include sentence-level prompting [1], zero-shot methods [17], and context-aware mapping techniques [19], as well as extensions to video retrieval [21].

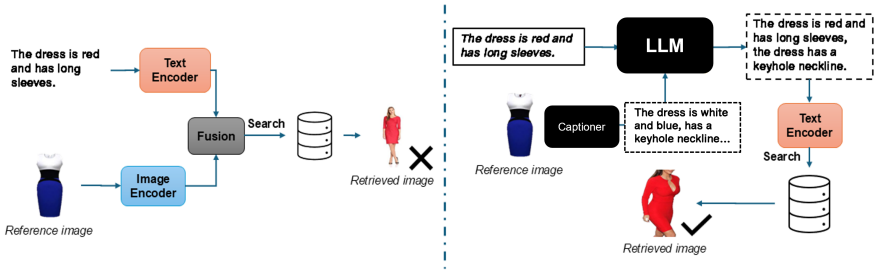

Fig. 1. Our proposed framework (Right), compared with the standard CIR [3] (Left).

The challenge of bridging the semantic gap between low-level visual features and high-level fashion concepts has been addressed in **fashion image retrieval (FIR)**. Research highlights low-level features and optimisation algorithms for semantic recognition [8, 9], while later advancements include semantic fusion networks [10] and compositional approaches [20] to capture outfit constituents. Interactive retrieval systems supported by datasets such as Fashion IQ [22] have emerged, utilizing methods like mix attention-based CNNs for brand logo recognition [12] and memory-based models for multi-turn feedback [13]. These studies underscore the importance of semantic understanding and iterative user feedback in refining retrieval results and enhancing overall retrieval efficiency.

Negative sampling plays a vital role in contrastive learning for image retrieval. Feng et al. [6] used multi-modal language models to generate triplets for CIR, addressing positive data scarcity. Zhou and Li [26] proposed a coarse-to-fine alignment framework for cross-modal image retrieval, improving performance through targeted sampling. Contrastive learning approaches, such as SimCLR [5] and non-parametric instance discrimination [23], have shown that augmentations and effective sampling of negatives are essential for learning robust representations. Additionally, conditional negative sampling [22] enhances feature transferability to new distributions, while contrastive hashing with vision transformer [16] improves retrieval performance by integrating hard negative samples.

3 Methodology

3.1 Overview

Our method builds upon the standard CIR framework [3] (Fig. 1), where a query combines a reference image and a modification text to retrieve a visually similar target image reflecting the described changes. Like prior work, we employ a dual-encoder contrastive model: the query encoder jointly embeds the reference image and modification text, while the target encoder processes candidate images. Training optimizes cosine similarity between matching query–target pairs using an in-batch contrastive loss.

The key innovation is enhancing the reference image representation through LLaVA-generated, attribute-aware captions. These captions capture fine-grained visual details, such as color, pattern, texture, and style, often missed by CLIP's global features. The generated caption is concatenated with the modification text to form a richer, contextually grounded textual input, improving fusion and alignment for fine-grained retrieval.

3.2 Enhanced Caption Generation via LLaVA

We enrich reference image descriptions using LLaVA, a vision-language model that integrates a vision encoder and language model end-to-end [11]. To generate detailed, context-aware captions highlighting fine-grained fashion attributes, we adopt a two-step prompting strategy:

Image-Conditioned Prompting For each reference image r and target image t, we input them into LLaVA alongside a structured prompt: *"Describe this fashion item in detail, focusing on color, pattern, texture, and style. Highlight any distinctive elements."* This yields caption C_r and C_t, which captures fine-grained visual attributes (e.g., "a knee-length dress with floral embroidery on a navy blue silk base").

Modified Text Synthesis To synthesize the modified text t, we concatenate C_r with the relative caption provided in the dataset (e.g., "make it pastel blue") and feed this into LLaVA with a follow-up prompt: *"Generate a concise instruction that modifies the original description based on the given change."* This produces a context-aware modified caption (e.g., "Change the navy blue silk dress to a pastel blue tone while retaining the floral embroidery"). This process ensures that t explicitly references attributes in C_r, reducing ambiguity.

3.3 Two-Stage Fine-Tuning with Augmented Triplets

We adopt a two-stage training framework to leverage both human-annotated and synthetic triplets:

- **Stage 1 (Coarse Alignment):** The query and target encoders are jointly trained with in-batch negatives to align reference images and modification texts. The query encoder combines CLIP-ViT features and LLaVA captions with relative text embeddings, enabling attribute-aware feature learning.
- **Stage 2 (Refined Alignment):** The target encoder is frozen while the query encoder is fine-tuned using hard negatives—samples with high similarity but mismatched attributes (e.g., same category, different color). This enhances the model's ability to capture subtle attribute differences for fine-grained fashion retrieval.

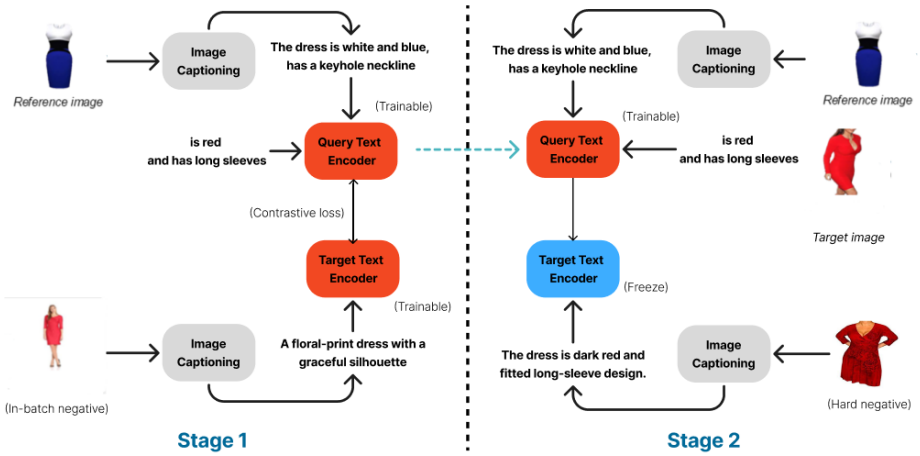

Fig. 2. Overview of our two-stage fashion retrieval training Pipeline.

3.4 Negative Sampling Strategy

To address the scarcity of negative samples, we adopt a **hybrid negative sampling strategy** inspired by Feng *et al.* [6]. Each training triplet is represented as (c_t, c_r, t_u), where c_t denotes the *target caption*, c_r the *reference caption* generated by **LLaVA**, and t_u the *user modification text* describing the desired change. We define three complementary negative types:

- **In-Batch Negatives:** Drawn from other samples within the same mini-batch, providing baseline diversity by contrasting each query (c_r, t_u) against all non-matching c_t.
- **Synthetic Negatives:** Plausible but incorrect targets generated by perturbing the target caption c_t. For example, attribute terms within c_t are modified (e.g., replacing "pastel blue" with "emerald green"), producing semantically close but mismatched descriptions. This challenges the model's ability to discriminate fine-grained differences.
- **Augmented Negatives:** Formed by cross-combining elements from different triplets, such as pairing an unrelated c_t' with a similar c_r or t_u, generating hard and diverse negatives that reduce spurious correlations.

This hybrid strategy exposes the model to both easy and hard negatives, thereby strengthening contrastive alignment and improving sensitivity to subtle attribute variations.

4 Experiments

4.1 Implementation Details

We used CLIP-ViT-B/32 for image encoding and LLaVA-v1.5-13b-3GB for text encoding. The fused image-text representations are processed through a 2-layer

Table 1. Performance of our proposed method compared with SOTA methods. The best and second best methods are shown in **bold** and *italic*.

Method	Dress		Shirt		Top-Tee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CompoDiff [7]	40.65	57.14	36.87	57.39	43.93	61.17	40.48	58.57
SPRC [2]	47.80	*72.70*	55.84	74.37	58.89	78.99	54.17	75.35
MAPNet [18]	**51.17**	**74.12**	*56.37*	*75.17*	**59.56**	*79.30*	**55.70**	**76.20**
Ours	*49.62*	71.83	**57.02**	**75.66**	*59.31*	**79.50**	*55.32*	*75.67*

transformer to produce query embeddings. We froze CLIP text encoder, used to encode target images captioning from LLaVa consistently across all experiments.

4.2 Experimental Settings

All experiments were performed on the *FashionIQ* [22], a natural language-based interactive fashion product retrieval dataset. It contains 77,684 images crawled from Amazon.com, covering three categories: Dresses, Tops & Tees, and Shirts. Among the 46,609 training images, there are 18,000 image pairs. Each pair is accompanied by an average of two natural language sentences that describe one or multiple visual properties to modify in the reference image, such as "is shiny" or "is blue in color and floral, and with white base."

We employed *Recall@K*($R@K$) [14] as the primary evaluation metric, which measures the proportion of queries for which the retrieved top K images include the correct target image.

4.3 Experimental Results

Comparison with SOTA Methods On the full FashionIQ dataset, our CLIP-4CIR model with LLaVA-enhanced captioning achieves better quantitative results at Table 1 than coarse-grained baselines like standard CIR models (e.g., CompoDiff [7] and SPRC [2]), though it remains below recent SOTA method (e.g., MAPNet [18]). The framework demonstrates stronger attribute expressiveness and compositional reasoning, particularly for fine-grained, multi-attribute queries. While minor alignment noise arises from the linguistic variability of generated captions, these enriched descriptions enhance visual-text alignment and retrieval interpretability. Overall, the results confirm the effectiveness of integrating multimodal caption refinement to advance fine-grained fashion retrieval.

Qualitative Results As shown in Figure 3, our method accurately retrieves targets with clear geometry and distinct patterns, showing strong reasoning over attributes like neckline, sleeves, and graphics. LLaVA-enhanced captions further improve attention to fine-grained details beyond category-level semantics.

Fig. 3. Illustrative positive examples of our method's performance. The captions for the queries are as follows: (a) "has a v neck and has a flower pattern", (b) "it has a floral print and long sleeves and has longer sleeves and is leopard print", (c) "has a more fun graphic and has more arrows on it", and (d) "is red in color and is red with different facial drawings".

Fig. 4. Illustrative negative examples of our method's performance. The captions for the queries are as follows: (a) "has no sleeves and only pink color and is more pink and sleeveless", (b) "it has a floral print and long sleeves and has longer sleeves and is leopard print", (c) "is gray with figures on it and is more vogue", and (d) "is more tighter and is more fitted and black".

Conversely, Figure 4 presents failure cases where the model struggles to resolve subtle variations, particularly for plain-colored garments or logo-based differences requiring pixel-level precision. These limitations suggest the need for spatial grounding or attention mechanisms to better localize fine-grained attributes.

This analysis highlights the model's strength in processing **composite attributes with distinctive visual signatures**, while revealing limitations in handling **generic color/shape descriptions** that dominate fashion datasets. The absence of spatial grounding mechanisms again exacerbates challenges in distinguishing near-identical plain-colored items.

5 Discussion and Future Work

5.1 Discussion

Although the proposed framework demonstrates encouraging qualitative behaviour and improved handling of complex, fine-grained queries, the overall quantitative performance still falls short of fully realizing the potential of fine-grained fashion retrieval. We attribute this observation to several interrelated factors:

- **Caption Complexity and Alignment Noise.** LLaVA-generated captions, while rich in attribute-level detail, often exhibit higher linguistic variability and verbosity compared to human-annotated descriptions. This increases the semantic gap between textual and visual embeddings.
- **Absence of Spatial Grounding.** The model encodes images globally without explicit mechanisms to focus on attribute-relevant regions (e.g., sleeves, neckline, logos). As a result, retrieval performance degrades in cases where subtle, localized differences determine correctness.
- **Limited Fine-Tuning Scale.** Current results are derived from a partial dataset due to time and resource constraints. Training on limited samples restricts the diversity of negative pairs and constrains the model's exposure to visual variations, limiting its generalization across broader retrieval scenarios.
- **Imbalance in Attribute Distribution.** Fashion datasets are skewed toward dominant features such as colour or pattern, while fine-grained attributes like fabric texture or subtle styling cues remain underrepresented. This imbalance can bias the model toward coarser visual signals, limiting precision in detailed queries.

Despite these limitations, the qualitative improvements observed in compositional reasoning and attribute-level retrieval underscore the potential of integrating large multimodal language models into composed image retrieval. These findings suggest that, with appropriate alignment strategies, the proposed framework can leverage the power of image captioning to bridge the gap between visual and linguistic semantics.

5.2 Potential and Future Directions

The findings reveal several promising directions for further development:

- **Caption Refinement and Filtering.** Future work will explore structured prompting and linguistic post-processing (e.g., attribute extraction and simplification) to mitigate noise introduced by free-form LLaVA captions. A

filtering pipeline that retains only dataset-aligned attribute terms may enhance embedding consistency.
– **Spatial Grounding and Attention.** Integrating spatial attention mechanisms or region-level feature pooling could enable the model to localize modifications more accurately. Such grounding is critical for resolving subtle distinctions, such as sleeve types or small logos, that current global embeddings may overlook.

Overall, while the current quantitative metrics remain close to baseline, the observed qualitative gains in compositional reasoning indicate significant potential. With continued optimization in caption alignment, grounding, and sampling strategies, the proposed framework has the capacity to advance fine-grained fashion retrieval beyond existing contrastive learning paradigms.

6 Conclusion

This study demonstrates the feasibility of enhancing composed image retrieval through refined caption generation and contrastive learning. By integrating a large multimodal language model (LLaVA) into the retrieval pipeline, we introduce a novel mechanism for generating attribute-aware captions that capture fine-grained visual details often overlooked by traditional annotation methods. Preliminary experiments indicate that, while overall recall remains slightly below baseline levels, the proposed framework exhibits stronger performance in complex, multi-attribute queries, highlighting its potential for improved compositional reasoning and interpretability. The observed trade-off between detailed caption expressiveness and alignment stability underscores the importance of future research into structured prompting, linguistic filtering, and attribute-level supervision. It will serve a promising work to bridge the gap between synthetic caption-based training and real-world search behaviour.

Acknowledgments. This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

1. Bai, Y., Xu, X., Liu, Y., Khan, S., Khan, F., Zuo, W., Goh, R.S.M., Feng, C.M.: Sentence-level prompts benefit composed image retrieval (2023), https://arxiv.org/abs/2310.05473
2. Bai, Y., Xu, X., Liu, Y., Khan, S., Khan, F., Zuo, W., Goh, R.S.M., Feng, C.M.: Sentence-level prompts benefit composed image retrieval. arXiv preprint (2023)
3. Baldrati, A., Bertini, M., Uricchio, T., del Bimbo, A.: Composed image retrieval using contrastive learning and task-oriented clip-based features (2023), https://arxiv.org/abs/2308.11485
4. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4959–4968 (Jun 2022)

5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020), https://arxiv.org/abs/2002.05709
6. Feng, Z., Zhang, R., Nie, Z.: Improving composed image retrieval via contrastive learning with scaling positives and negatives (2024), https://arxiv.org/abs/2404.11317
7. Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., Yun, S.: Compodiff: Versatile composed image retrieval with latent diffusion (2024), https://arxiv.org/abs/2303.11916
8. Hare, J.S., Lewis, P.H., Enser, P.G., Sandom, C.J.: Mind the gap: Another look at the problem of the semantic gap in image retrieval. In: Multimedia Content Analysis, Management, and Retrieval 2006. vol. 6073, pp. 75–86. SPIE (2006)
9. Karmokar, P.R., Parekh, R.: Recognition of semantic content in image and video. International Journal of Computer Applications **73**(15) (2013)
10. Liu, A.A., Zhang, T., Song, D., Li, W., Zhou, M.: Frsfn: A semantic fusion network for practical fashion retrieval. Multimedia Tools and Applications **80**, 17169–17181 (2021)
11. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
12. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2125–2134 (Oct 2021)
13. Pal, A., Wadhwa, S., Jaiswal, A., Zhang, X., Wu, Y., Chada, R., Natarajan, P., Christensen, H.I.: Fashionntm: Multi-turn fashion image retrieval via cascaded memory. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11323–11334 (October 2023)
14. Patel, Y., Tolias, G., Matas, J.: Recall@k surrogate loss with large batches and similarity mixup (2022), https://arxiv.org/abs/2108.11179
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020, arXiv preprint arXiv:2103.00020
16. Ren, X., Zheng, X., Zhou, H., Liu, W., Dong, X.: Contrastive hashing with vision transformer for image retrieval. International Journal of Intelligent Systems **37**(12), 12192–12211 (2022). https://doi.org/https://doi.org/10.1002/int.23082, https://onlinelibrary.wiley.com/doi/abs/10.1002/int.23082
17. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19305–19314 (Jun 2023)
18. Shi, J., Yin, X., Chen, Y., Zhang, Y., Zhang, Z., Xie, Y., Qu, Y.: Multi-schema proximity network for composed image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2025 (2025)
19. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Hu, Y., Wu, Q.: Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. Proceedings of the AAAI Conference on Artificial Intelligence **38**(6), 5180–5188 (Mar 2024). https://doi.org/10.1609/aaai.v38i6.28324, https://ojs.aaai.org/index.php/AAAI/article/view/28324

20. Valle, D., Ziviani, N., Veloso, A.: Effective fashion retrieval based on semantic compositional networks. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2018). https://doi.org/10.1109/IJCNN.2018.8489494
21. Ventura, L., Yang, A., Schmid, C., Varol, G.: Covr-2: Automatic data construction for composed video retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 11409–11421 (Dec 2024). https://doi.org/10.1109/tpami.2024.3463799, http://dx.doi.org/10.1109/TPAMI.2024.3463799
22. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: Fashion iq: A new dataset towards retrieving images by natural language feedback (2020), https://arxiv.org/abs/1905.12794
23. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination (2018), https://arxiv.org/abs/1805.01978
24. Xu, Y., Bin, Y., Wei, J., Yang, Y., Wang, G., Shen, H.T.: Multi-modal transformer with global-local alignment for composed query image retrieval. Trans. Multi. **25**(1), 8346–8357 (Jan 2023). https://doi.org/10.1109/TMM.2023.3235495, https://doi.org/10.1109/TMM.2023.3235495
25. Zhao, Y., Song, Y., Jin, Q.: Progressive learning for image retrieval with hybrid-modality queries (2022), https://arxiv.org/abs/2204.11212
26. Zhou, L., Li, Y.: Coarse-to-fine alignment makes better speech-image retrieval (2024), https://arxiv.org/abs/2408.13119