HI-CLIP: CHARACTERIZING FUNCTIONAL INTERACTIONS AMONG ATTENTION HEADS IN CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

As CLIP increasingly serves as the visual backbone of multimodal models, understanding its internal visual representations is important for building reliable and controllable models. Existing interpretability studies primarily focus on what concepts individual attention heads encode or what functional roles they play, while the functional relationships among heads remain underexplored, limiting the reliability of head-level interventions such as pruning and editing. Therefore, we propose **HI-CLIP**, a framework for interpreting inter-head functional relationships in CLIP-ViT. HI-CLIP identifies criterion-specific expert and collaborative heads, while characterizing inter-head relationships as supportive, redundant, or interfering. Experiments on ViT-B/16 and ViT-L/14 show that HI-CLIP reveals how criterion-specific visual representations are organized and formed through head interactions. Moreover, the discovered head roles and interactions improve customized retrieval without additional training and guide parameter-efficient adaptation. Overall, HI-CLIP provides a relational perspective for understanding how attention heads interact in CLIP, while demonstrating practical utility for customized retrieval and targeted fine-tuning.

1 INTRODUCTION

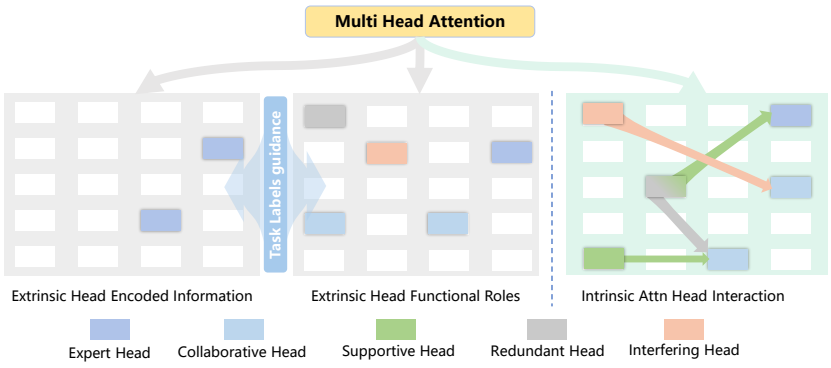

Figure 1: Comparison of different paradigms for attention-head analysis in CLIP. Left: prior methods identify task-related individual functions heads Zhou et al. (2025); Yeo et al. (2026); Zheng et al. (2026). Mid: recent approaches further characterize multiple functional head types in the task space (Lu et al., 2026; Nam et al., 2025). Right: our HI-CLIP directly models representation-based interactions among heads, revealing expert, supportive, redundant, and interfering relationships.

CLIP (Radford et al., 2021) has become a key component of vision-language models, and advances in CLIP have continued to improve the capabilities of multimodal large language models (MLLMs). Since its introduction in 2021, CLIP has established an influential paradigm for vision-language learning. Blip-2 (Li et al., 2023) further enabled large language models to perceive visual signals through CLIP-based visual representations. Since then, a series of MLLMs, including MiniGPT-4, LLaVA, CogVLM, Pixtral, Llama3.2-Vision, and Molmo (Zhu et al., 2023; Liu et al., 2024; Wang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

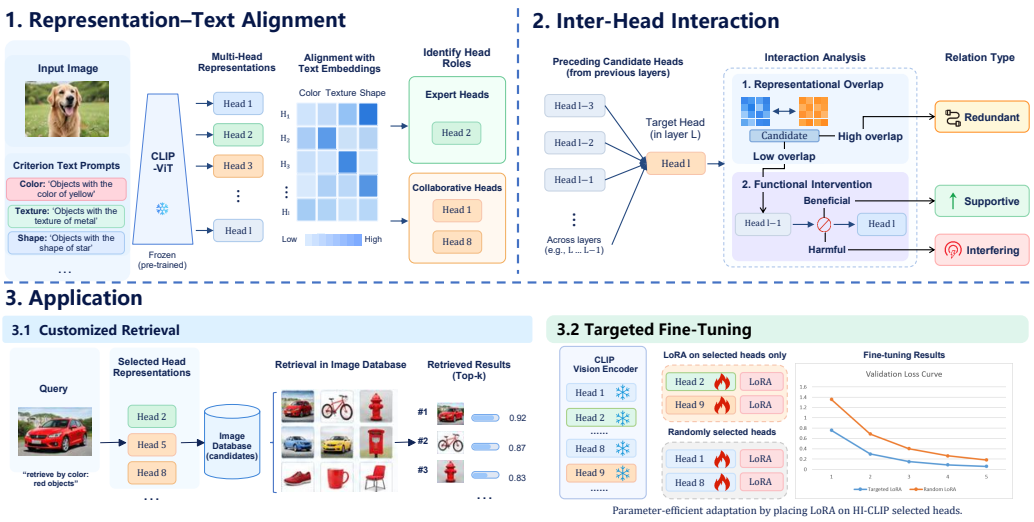

Figure 2: Overview of HI-CLIP. The framework consists of three stages: criterion-specific head-role identification via representation-text alignment, inter-head interaction analysis based on representational overlap and functional dependence, and downstream application to customized retrieval and targeted model intervention.

et al., 2024; Agrawal et al., 2024; Grattafiori et al., 2024; Deitke et al., 2024), have adopted or built upon the CLIP paradigm, demonstrating increasingly strong visual understanding. Despite its widespread adoption, it remains unclear how CLIP's internal components encode visual information and how such information is distributed, reused, or suppressed across attention heads. Without this knowledge, selecting which components to preserve, adapt, or remove remains largely heuristic, making head-level intervention and parameter-efficient adaptation less reliable.

Existing studies on CLIP interpretability have primarily focused on individual attention heads or neurons, characterizing the visual concepts and semantic information encoded by these components (Gandelsman et al., 2024a; Gandelsman et al., 2024b; Lin et al., 2025). Subsequent works further identify functionally important heads for specific tasks and use them to improve targeted model behaviors. For example, Yeo et al. (2026) distinguishes bias-related heads from those encoding target-category information, while Zhou et al. (2025) and Zheng et al. (2026) identify heads critical to safety-related behavior through intervention. More recently, DAHIVT (Lu et al., 2026) and CHG (Nam et al., 2025) further categorize heads as task-specific expert, collaborative, redundant, interfering, or facilitating, relying respectively on LLM-based semantic interpretation and learned head-gating objectives. However, such approaches may entangle the inferred head roles with external language priors or task- and dataset-specific training signals, **making the resulting relationships task-conditioned rather than direct measurements of the intrinsic interactions among the original CLIP heads.** This limits their generalization to customized representation construction and targeted parameter-efficient fine-tuning. A faithful characterization of inter-head relationships should therefore rely on the heads' own representations and model-internal causal effects, without external semantic models or additional task-specific training.

However, directly characterizing inter-head functional relationships is challenging. For an expert or collaborative head that represents a particular visual criterion, other heads may influence its representation in different ways: supportive heads provide complementary or lower-level visual evidence required by the target head; redundant heads encode substantially overlapping information whose contribution can be compensated for by alternative pathways; and interfering heads introduce irrelevant or competing information that degrades the target representation. Distinguishing these relationships is difficult because individual head representations are often polysemantic. As a result, similarity between full representations may not accurately indicate what information is shared between heads. Moreover, representational similarity does not necessarily imply functional dependence because attention heads are coupled through shared residual streams and cross-layer informa-

tion flow. Interpreting inter-head relationships therefore requires isolating dominant representational subspaces and disentangling representational overlap from functional dependence.

To address this challenge, we propose **HI-CLIP**, **a framework for characterizing functional relationships among attention heads in CLIP.** As shown in Figure 2, we first identify criterion-specific expert and collaborative heads through representation-text alignment, and further analyze the distribution of semantic categories across attention heads. Taking the identified heads as targets, we use SVD to extract the dominant representational subspace of each head and measure subspace similarity between candidate and target heads. High subspace overlap indicates that a candidate may encode information redundant with the target, whereas low overlap may correspond to either complementary supportive information or competing interfering information. We therefore further perform causal ablation to determine how each candidate functionally affects the target representation. Experiments on ViT-B/16 and ViT-L/14 validate that the identified heads capture substantially stronger criterion-specific representations and that the characterized relationships correspond to distinct functional effects under intervention. Beyond interpretation, we further show that these head roles and interactions can be directly exploited for training-free customized retrieval and can provide interpretable guidance for targeted parameter-efficient fine-tuning. **In conclusion, HI-CLIP not only identifies where criterion-specific visual information is represented, but also explains how such representations are formed through supportive, redundant, and interfering interactions among attention heads, while making these interpretations practically actionable.**

Our contributions can be summarized as follows:

1. We formulate the problem of inter-head functional relationship interpretation in CLIP, shifting the focus of interpretability from isolated head roles and head-task associations toward understanding how attention heads functionally interact to form criterion-specific visual representations.
2. We propose HI-CLIP, a quantitative and training-free framework for disentangling representational overlap from functional dependence under cross-layer head coupling. HI-CLIP first identifies criterion-specific expert and collaborative heads through representation-text alignment, and then combines subspace similarity with causal intervention to characterize inter-head relationships as supportive, redundant, or interfering.
3. We validate HI-CLIP on ViT-B/16 and ViT-L/14 and demonstrate its practical utility. The discovered head roles and functional relationships reveal how criterion-specific information is organized across attention heads, improve criterion-aware representations for customized retrieval, and provide interpretable guidance for placing LoRA adapters during parameter-efficient fine-tuning.

2 Related Works

2.1 Semantic Information Encoded by Attention Heads

Existing studies have sought to interpret the information encoded by attention heads or neurons of CLIP. These methods interpret the information encoded by CLIP's internal components in natural language using texts generated by language models. Oikarinen & Weng (2022) describes the semantic concepts encoded by individual neurons by matching their activations with natural-language concepts. Gandelsman et al. (2024a) further characterizes the information represented by CLIP's attention heads, using sparse text representations, enabling fine-grained semantic descriptions of internal representations. By analyzing the second-order effects of MLP neurons in CLIP, Gandelsman et al. (2024b) reveals how information represented by individual neurons is affected by other components. These methods provide descriptions of the visual concepts and semantic information encoded by internal components. However, the resulting interpretations can be difficult to quantify and compare systematically across attention heads.

2.2 Functional Roles of Attention Heads

Recent studies have moved beyond describing encoded information and instead characterized attention heads according to the functional roles they play in model behavior or downstream tasks. Zheng et al. (2026) identifies a set of heads closely associated with safety judgments. Lin et al.

(2025) evaluates the contributions of different attention heads to visual representations and downstream performance. Li et al. (2021), Yun et al. (2021), and Zhang (2025) analyze diversity among attention heads and its relationship with model performance, while Butler et al. (2026) quantifies the aggregate effects of interactions among multiple heads. These studies demonstrate that attention heads can make substantially different contributions to model computation, motivating more fine-grained characterization of their functional roles.

More recently, DAHIVT (Lu et al., 2026) and CHG (Nam et al., 2025) have introduced explicit role-based frameworks for attention-head interpretation. DAHIVT builds on Gandelsman et al. (2024a) and leverages large language models to interpret head semantics before identifying expert, collaborative, interfering, and redundant attention heads. CHG employs a gating-based training framework to distinguish task-relevant, task-irrelevant, and interfering attention heads under different optimization settings. However, these approaches still primarily characterize the relationship between individual heads and a task or behavior—they explain what functional role a head plays for a particular task, rather than directly modeling how one attention head functionally affects another.

3 METHODOLOGY

Fig. 2 presents an overview of HI-CLIP. The framework first identifies criterion-specific expert and collaborative heads through representation-text alignment, then characterizes inter-head relationships by combining representational overlap with functional dependence, and finally uses the resulting head roles and relationships to support customized retrieval and targeted model intervention.

3.1 CRITERION-SPECIFIC HEAD IDENTIFICATION

Different attention heads may encode a visual criterion through either a single dominant head or multiple category-specialized heads. We distinguish these expert and collaborative organizations using representation–text alignment.

Expert and Collaborative Heads. An expert head dominates most semantic categories within a criterion. Collaborative heads instead specialize in complementary category subsets when no single head dominates the criterion. We identify these organizations by measuring the alignment between each head and category.

Representation-Text Alignment. Consider a visual criterion q with semantic categories

$$C_q = \{c_1, c_2, \ldots, c_M\}. \tag{1}$$

For an image x_i, let $z_h(x_i)$ be the representation produced by head h, and let t_c be the CLIP text embedding of category c. Their cosine similarity is

$$a_{i,h,c} = \frac{z_h(x_i)^\top t_c}{\|z_h(x_i)\|_2 \|t_c\|_2}. \tag{2}$$

We average over samples in category c:

$$s(h,c) = \frac{1}{|D_c|} \sum_{x_i \in D_c} a_{i,h,c}, \tag{3}$$

where D_c is the set of samples in category c. Each category is assigned to its highest-scoring head:

$$h^*(c) = \arg\max_h s(h,c). \tag{4}$$

The coverage of head h is

$$\text{Coverage}(h) = \frac{|\{c \in C_q : h^*(c) = h\}|}{|C_q|}. \tag{5}$$

We identify an expert organization when

$$\max_h \text{Coverage}(h) \geq \tau_e. \tag{6}$$

The highest-coverage head is selected as the expert.

Otherwise, letting $\mathcal{H}_K$ denote the top-K heads by coverage, we identify a collaborative organization when

$$\max_h \text{Coverage}(h) < \tau_e \quad \text{and} \quad \sum_{h \in \mathcal{H}_K} \text{Coverage}(h) \geq \tau_c. \tag{7}$$

We set $\tau_e = 0.7$ and $\tau_c = 0.8$ in all experiments.

Thus, the category-wise assignment distinguishes expert from collaborative organization without additional training or learned gates.

3.2 Inter-head Functional Relationship Characterization

Identifying criterion-specific expert and collaborative heads reveals where visual information is strongly represented, but does not explain how such representations are formed through interactions with other attention heads. Since the outputs of earlier heads are written into the shared residual stream, they may subsequently be reused, duplicated, or counteracted by deeper heads. We therefore characterize the directional functional relationship $h \to t$, where t denotes an identified criterion-specific target head and h denotes a preceding candidate head.

Representational overlap. We first characterize the representational overlap between h and t. Given their representations over N samples,

$$Z_h, Z_t \in \mathbb{R}^{N \times d}, \tag{8}$$

we perform singular value decomposition and retain the top-k right singular vectors, $V_h^{(k)}, V_t^{(k)}$. Their subspace similarity is measured using the principal angles between the two subspaces:

$$S(h,t) = \frac{1}{k} \sum_{i=1}^{k} \cos^2 \theta_i, \tag{9}$$

where $S(h,t) \in [0,1]$. A higher similarity indicates stronger overlap between the dominant representational subspaces of the two heads. Importantly, we use this similarity only as a structural prior, since representational overlap does not necessarily imply functional dependence.

Functional intervention. To determine how candidate head h functionally affects target head t, we perform mean ablation by replacing the contribution of h with its calibration-set mean representation and recomputing the downstream representation of t. Let Q_t denote the criterion-specific representation quality of t under normal inference and $Q_t^{(-h)}$ denote its quality after ablating h. We define the intervention effect as

$$\Delta(h \to t) = Q_t^{(-h)} - Q_t. \tag{10}$$

A negative intervention effect indicates that removing h degrades the target representation, whereas a positive effect indicates that removing h improves it. We then characterize the functional relationship as

$$R(h \to t) = \begin{cases} \text{Supportive}, & \Delta(h \to t) < -\tau, \\ \text{Interfering}, & \Delta(h \to t) > \tau, \\ \text{Redundant}, & |\Delta(h \to t)| \leq \tau \wedge S(h,t) > \tau_s, \\ \text{Non-interacting}, & \text{otherwise}. \end{cases} \tag{11}$$

A supportive head provides information that is functionally required for forming the criterion-specific representation of the target head, whereas an interfering head introduces information whose removal improves that representation. A redundant head exhibits substantial representational overlap with the target head but produces little functional effect when individually removed, indicating that its information can be compensated for by alternative pathways. In this way, subspace similarity characterizes what information is shared, while causal intervention determines whether and how that information is functionally used.

Table 1: Customized clustering performance on ViT-B/16 and ViT-L/14.

Backbone	Method	Color			Texture			Shape		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
ViT-B/16	Base	11.63	0.63	0.09	12.33	1.59	0.40	95.23	93.62	92.44
	Random	20.96	18.12	8.15	14.88	3.72	1.55	72.84	71.95	64.30
	DAHIVT Expert	93.32	91.41	86.99	17.71	8.78	3.78	42.09	35.43	25.38
	DAHIVT Collaborative	91.20	90.19	84.95	12.41	1.49	0.46	94.20	90.75	88.62
	HI-CLIP (Ours)	**95.68**	**93.38**	**90.95**	**68.63**	**50.81**	**44.66**	**98.53**	**96.59**	**96.80**
ViT-L/14	Base	11.78	0.97	0.15	12.30	1.38	0.37	98.07	96.55	96.41
	Random	16.73	5.72	3.83	14.17	2.84	1.12	71.74	68.10	61.01
	DAHIVT Expert	37.91	31.41	21.22	24.33	13.81	7.84	92.33	91.47	89.54
	DAHIVT Collaborative	43.73	41.51	29.27	34.16	28.87	20.15	96.23	95.54	94.86
	HI-CLIP (Ours)	**97.58**	**95.47**	**94.81**	**63.22**	**47.89**	**39.82**	**98.65**	**97.05**	**97.01**

4 EXPERIMENTS

4.1 CRITERION-SPECIFIC HEAD VALIDATION VIA CUSTOMIZED CLUSTERING

Since no ground-truth annotations are available for criterion-specific expert or collaborative heads, we evaluate the identified heads through customized clustering. The intuition is that if an attention head truly captures a specific visual criterion, such as color, texture, or shape, its representation should better preserve the category structure under that criterion and therefore yield stronger clustering performance.

We conduct experiments on Clevr-4-10k (Vaze et al., 2023) using CLIP ViT-B/16 and ViT-L/14. For each criterion, we use the representations of the heads identified by HI-CLIP as clustering features and evaluate them with ACC, NMI, and ARI. We compare the HI-CLIP-selected head representations with three alternatives: the final CLIP representation, representations from randomly selected heads, and representations from the expert and collaborative heads identified by DAHIVT.

As shown in Tab. 1, HI-CLIP achieves the best clustering performance across the three criteria and both CLIP architectures, with particularly large improvements on color and texture. For example, on ViT-B/16, HI-CLIP improves texture ACC from 17.71% with DAHIVT to 68.63%, while on ViT-L/14 it improves color ACC from 43.73% to 97.58%. These results indicate that HI-CLIP more accurately identifies attention heads that encode criterion-specific visual information than existing head-selection methods.

4.2 VISUALIZATION OF CRITERION-SPECIFIC REPRESENTATIONS

To examine whether the heads identified by HI-CLIP encode criterion-specific visual information, we visualize their representations using t-SNE. For each criterion, we compare the final CLIP image representation with the representations from the expert or collaborative heads identified by HI-CLIP.

As shown in Fig. 3, we visualize Color, Shape, and Texture on ViT-B/16. The top row shows the final CLIP representation, and the bottom row shows the representations from the criterion-specific heads. Samples are colored by their ground-truth category under each criterion.

The final CLIP representation jointly contains information from multiple visual criteria, so categories under a specific criterion may remain mixed. In contrast, the heads identified by HI-CLIP show clearer criterion-specific structures: the Color expert head forms separated color clusters, the Shape collaborative heads preserve clear shape categories, and the Texture heads also improve texture separation. These results further support that HI-CLIP more accurately identifies heads encoding criterion-specific visual information.

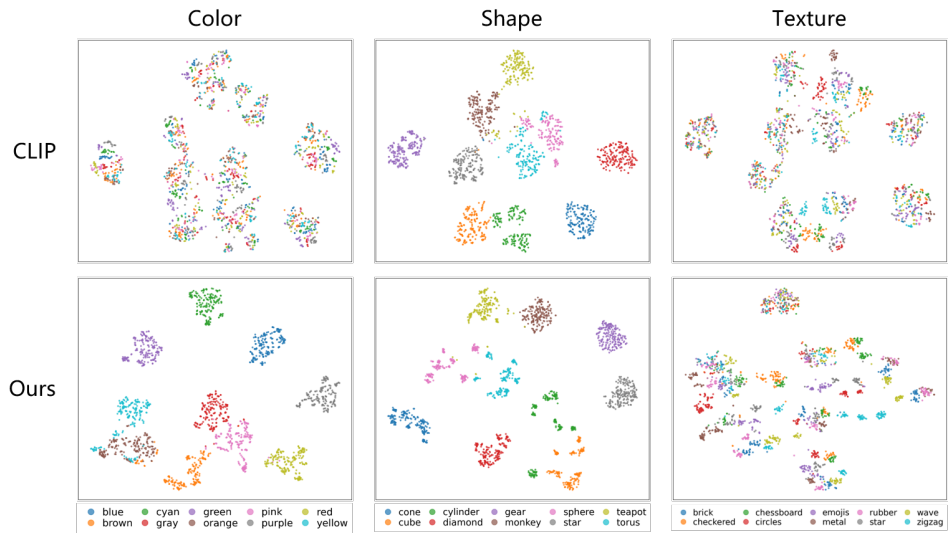

Figure 3: Qualitative analysis of criterion-specific representations identified by HI-CLIP on ViT-B/16. The top row shows the final CLIP representations, while the bottom row shows the corresponding HI-CLIP-selected head representations for Color, Shape, and Texture, which exhibit clearer class separation.

Table 2: Effect of ablating representative candidate heads on the criterion-specific representation quality of their target heads. "Before" and "After" denote target-head performance before and after candidate-head ablation, respectively.

Relationship	**Candidate → Target**	**ACC**		**NMI**		**ARI**		**Effect**
		Before	**After**	**Before**	**After**	**Before**	**After**	
Supportive	L7H2 → L11H10	93.32	91.43	91.41	89.97	86.99	84.53	Decrease
Interfering	L7H5 → L11H10	93.32	94.23	91.41	92.4	86.99	88.78	Increase
Redundant	L8H1 → L11H10	93.32	93.96	91.41	92.1	86.99	88.09	Negligible

4.3 Functional Relationship Characterization

We further validate whether the functional relationships identified by HI-CLIP correspond to distinct effects on criterion-specific representations. For representative candidate-target head pairs, candidate-head ablation produces three characteristic outcomes: performance degradation indicates a supportive relationship, performance improvement indicates an interfering relationship, and little change despite substantial representational overlap indicates redundancy. These effects respectively reflect functional dependence, competing information, and compensable information pathways.

Saliency visualization. We visualize representative head pairs to illustrate the three functional relationships. Supportive heads attend to criterion-relevant regions that align with or complement the target head, whereas interfering heads emphasize irrelevant regions or competing attributes. Redundant heads show saliency patterns similar to the target head, consistent with their representational overlap and functional substitutability.

5 Applications

5.1 Customized Retrieval

Goal and setup. Customized retrieval aims to return images that match a query under a user-specified visual criterion rather than by overall semantic similarity. For color, texture, and shape,



Figure 4: Saliency visualizations of representative Expert, Supportive, Redundant, and Interfering attention heads.

Table 3: Quantitative results for customized retrieval using HI-CLIP-selected heads.

Model	Criterion	Selected heads	Hit@1	Hit@5
ViT-B/16	Color	L11H10	0.9972	0.9995
ViT-B/16	Texture	L10H5 + L9H7	0.8870	0.9791
ViT-B/16	Shape	L10H5 + L11H5 + L11H7	0.9981	1.0000

we use the corresponding expert or collaborative heads identified by HI-CLIP to represent both the query and gallery images, and rank the gallery by cosine similarity. We evaluate this training-free application on disjoint query and gallery splits of Clevr-4-10k with CLIP ViT-B/16, reporting Hit@1 and Hit@5 according to whether a retrieved image shares the query's label under the requested criterion.

Results. As shown in Tab. 3, the selected heads achieve near-perfect retrieval for color and shape, while also attaining 0.8870 Hit@1 and 0.9791 Hit@5 for the more challenging texture criterion. Fig. 5 further shows that changing the requested criterion changes the returned images for the same query while preserving the relevant visual property. Together, these results demonstrate that HI-CLIP can directly turn its criterion-specific head interpretation into controllable retrieval without additional training.

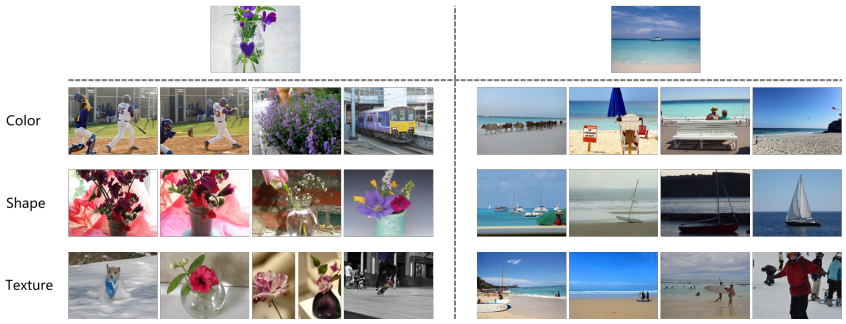

Figure 5: Qualitative customized-retrieval results. The query images are shown at the top, and the rows below show the top-ranked gallery images under different visual criteria. Changing the criterion produces distinct rankings that follow the requested visual property rather than overall image semantics.

5.2 Targeted Parameter-Efficient Fine-Tuning

Goal and setup. We test whether HI-CLIP can identify effective locations for parameter-efficient adaptation. For color, texture, and shape, Targeted LoRA inserts LoRA modules (Hu, 2022) only into the attention projections of the corresponding expert or collaborative heads, with the remaining CLIP parameters frozen. We compare it with Random LoRA, which adapts the same number of randomly selected heads under an identical parameter budget, data split, optimization setting, and five-epoch schedule. Validation loss and accuracy measure adaptation efficiency.

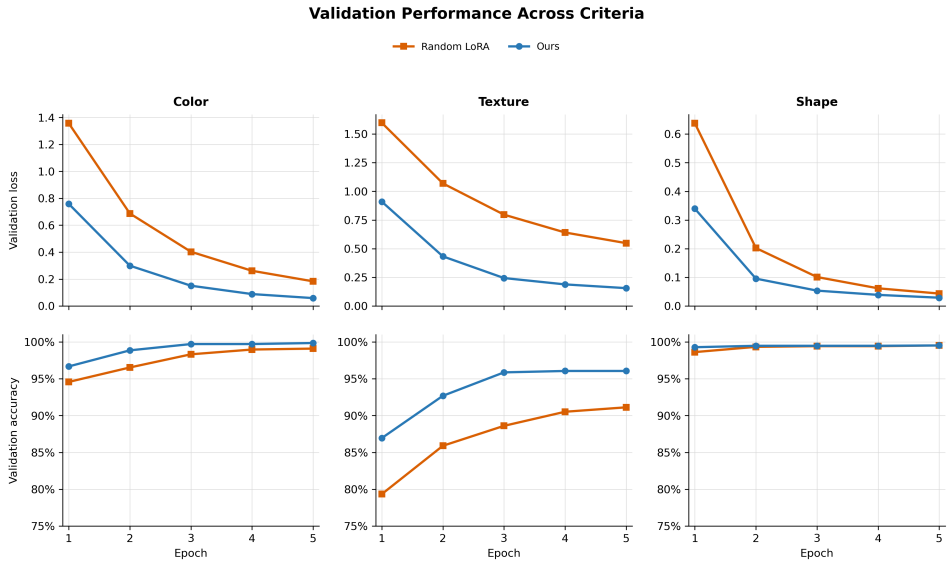

Figure 6: Validation loss and accuracy of Targeted LoRA and Random LoRA over five epochs for color, texture, and shape.

Results. As shown in Fig. 6, Targeted LoRA reduces validation loss faster for all three criteria and improves accuracy more rapidly for color and texture, with the largest advantage on texture. For shape, both methods reach similar final accuracy, but Targeted LoRA converges earlier and maintains a lower loss. Under the same adaptation budget, the HI-CLIP-selected heads therefore provide more effective fine-tuning targets than randomly selected heads.

6 Conclusion

We presented HI-CLIP, a framework for interpreting how attention heads jointly form criterion-specific visual representations in CLIP. HI-CLIP identifies expert and collaborative heads and characterizes inter-head relationships as supportive, redundant, or interfering. Experiments on ViT-B/16 and ViT-L/14 validate these head roles and interactions through clustering, visualization, and intervention. The identified heads also enable criterion-aware retrieval and more efficient LoRA placement. These results show that modeling functional relationships among heads provides both a clearer account of CLIP's internal representations and practical guidance for targeted adaptation.

AI use statement

Generative AI tools were used to assist with language polishing and improving the readability of the manuscript, as well as for literature retrieval and discovery of relevant prior work. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Landon Butler, Abhineet Agarwal, Justin Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. Proxyspex: Inference-efficient interpretability via sparse feature interactions in llms. *Advances in Neural Information Processing Systems*, 38:72306–72340, 2026.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Y Yang, JS Park, M Salehi, N Muennighoff, K Lo, L Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. *URL https://arxiv. org/abs/2409.17146*, 2024.

Yossi Gandelsman, Alexei Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *International Conference on Learning Representations*, volume 2024, pp. 18395–18416, 2024a.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting the second-order effects of neurons in clip. *arXiv preprint arXiv:2406.04341*, 2024b.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Edward J Hu. yelong shen, phillip wallis, zeyuan allen-zhu, yuanzhi li, shean wang, lu wang, and weizhu chen. 2022. lora: Low-rank adaptation of large language models. In *International conference on learning representations*, volume 4, 2022.

Jian Li, Xing Wang, Zhaopeng Tu, and Michael R Lyu. On the diversity of multi-head attention. *Neurocomputing*, 454:14–24, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Feng Lin, Marco Chen, Haokui Zhang, Xiaotian Yu, Guangming Lu, and Rong Xiao. Not all attention heads are what you need: Refining clip's image representation with attention ablation. *arXiv preprint arXiv:2507.00537*, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Zhenyu Lu, Yuheng Jia, Wei You, and Hao Chen. Discovering attention head interactions in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3332–3342, 2026.

Andrew Nam, Henry Conklin, Yukang Yang, Tom Griffiths, Jonathan D Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. *Advances in Neural Information Processing Systems*, 38:115609–115631, 2025.

Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

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 *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. *Advances in Neural Information Processing Systems*, 36:19962–19989, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

Wei Jie Yeo, Rui Mao, Moloud Abdar, Erik Cambria, and Ranjan Satapathy. Locate then correct: Debiasing attention heads in clip. In *Forty-third International Conference on Machine Learning*, 2026.

Hyeongu Yun, Taegwan Kang, and Kyomin Jung. Analyzing and controlling inter-head diversity in multi-head attention. *Applied Sciences*, 11(4):1548, 2021.

Zhen Zhang. Diversity-driven multi-head attention for enhanced trajectory prediction. *IEEE Access*, 13:94970–94981, 2025.

Ziwei Zheng, Lijun He, Junyao Zhao, Le Yang, and Fan Li. Sahs: Unraveling safety attention heads in large vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, volume 2025, pp. 84042–84071, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: enhancing vision-language understanding with advanced large language models. corr abs/2304.10592 (2023). *arXiv preprint arXiv:2304.10592*, 2023.