Supplementary materials of CAST: Context-Aware Dynamic Latent Space Transformation for Interactive Text-to-Image Retrieval

Xuanzuo Lin[1,2], Min Zhang[3], Daizong Liu[4†], Zhiwen Zuo[1,2]
Xun Yang[5], Changting Lin[6], Xun Wang[1,2], Jianfeng Dong[1,2†]
[1]Zhejiang Gongshang University,
[2]Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology,
[3]East China Normal University, [4]Wuhan University,
[5]University of Science and Technology of China, [6]Zhejiang University

We report more experimental results and additional technical details that are not included in the main paper due to space constraints:

- More clarification of our main motivation (Section A).
- Analyze the effectiveness and potential limitations of the Hits@K metric (Section B).
- Evaluation of different LLMs for dialogue context generation during inference (Section C).
- Ablation study on the impact of the rank r in the CLP module (Section D).
- Performance comparison across different backbone scales (Section E).
- Complete per-round Recall@10 results for all dialogue sources and Hits@K on VisDial [2] (Section F).
- Performance comparison using BRI metric (Section G).
- Implementation details of our CAST framework, including module configurations, training settings and prompt(Section H).
- More qualitative visualizations demonstrating context-aware space transformation (Section I).

A. Motivation Clarification

Our motivation lies in the fact that user intent in multi-round I-TIR is constantly evolving during the multi-round interaction. For instance, the initial query may first focus on the object's appearance, while later turns may shift towards its surrounding context or other attributes. Under such evolving intent, we argue that performing retrieval in a *static* context-agnostic feature space is suboptimal, since it cannot adjust the matching according to the user's dynamic focus. We therefore propose to dynamically transform the feature space conditioned on the dialogue context, so that features relevant to the current intent are emphasized. This enables finer-grained, intent-specific alignment between text and images across rounds, leading to improved multi-round cross-modal retrieval.

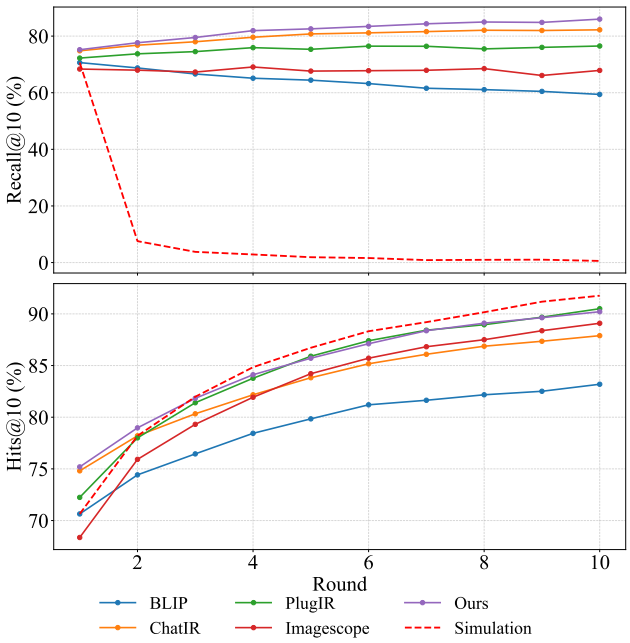


Figure 1. Performance in terms of Recall@10 (up) and Hits@10 (down) on the VisDial dataset. The red dashed line denotes our *Simulation* baseline, which simulates significant query variations by removing the top-10 most similar samples retrieved in the previous round at each round. This experiment shows artificially increased Hits@10 but a rapid decline in Recall@10, illustrating that high Hits may not correspond to true retrieval coverage. In contrast, our method maintains improvement in both metrics.

B. Analysis of Evaluation Metrics

Previous interactive retrieval works [3, 4, 7] typically use Hits@K as the main evaluation metric. However, Hits@K only measures whether the target image appears in the top-K results in the current or any previous round, and thus fails to capture retrieval consistency across dialogue rounds.

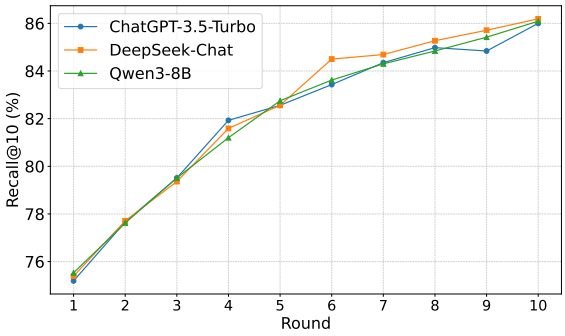


Figure 2. Performance comparison across three different models for dialogue context generation on the VisDial dataset.

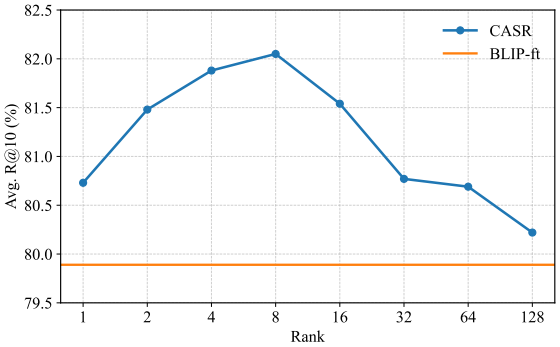


Figure 3. Impact of different rank settings in the CLP module.

Many methods artificially inflate Hits@K by allowing the query embedding to drift, increasing the chance of occasional hits but reducing overall recall. To investigate this, we conduct an experiment simulating significant query variations, where the top-10 most similar samples retrieved in the previous round are removed at each round. As shown in Figure 1, removing the top-10 retrieved samples from each subsequent round raises Hits@10 from 70.64% to 91.76%, yet Recall@10 drops from 70.64% to 0.58% after ten rounds, revealing that high Hits may coexist with poor retrieval coverage. In contrast, our method achieves the best overall Recall performance while maintaining Hits scores comparable to the current state-of-the-art methods. This indicates that the improvements stem from genuine semantic refinement rather than superficial hit accumulation. Therefore, we adopt Recall@K as the principal evaluation metric, as it better reflects feature alignment and retrieval robustness throughout multi-round interactions.

C. Impact of Dialogue Context Generation Models

In our framework, the generation of dialogue context plays a crucial role in the retrieval performance. To evaluate the impact of different large language models (LLMs) on retrieval performance through context generation, we conducted experiments comparing the performance of the retrieval system using various LLMs for dialogue context reconstruction. During training, we used ChatGPT-3.5-Turbo [9] to generate the dialogue context dataset and ensured consistency in the reconstruction prompt. However, for inference, we experimented with three different models for dialogue context generation: ChatGPT-3.5-Turbo [9], DeepSeek-Chat [6], and Qwen3-8B [1]. Although the context dataset used for training was generated by ChatGPT-3.5-Turbo, our experimental results demonstrate in Table 1 that performance remains high even when other large models are used for context generation during inference. In fact, in certain rounds, the contexts generated by DeepSeek-Chat and Qwen3-8B led to even better retrieval performance compared to ChatGPT-3.5-Turbo. This shows that, within our framework, although the capabilities of different large models may vary, their impact on retrieval performance is still minimal, and overall performance remains stable.

This flexibility highlights a core strength of our framework: it can adapt to contexts generated by different large language models and allows the most suitable model to be chosen based on task needs. As a result, the framework maintains stable and strong retrieval performance even when switching between various LLMs, further enhancing its overall versatility and reliability.

D. Analysis of Rank in the CLP Module

The rank r in the CLP module plays a crucial role in determining the capacity of the model to perform spatial transformations. We have analyzed the impact of different values of r on the performance of the CAST framework. As shown in Figure 3, increasing the rank r initially enhances the model's representational capacity, allowing CLP to capture richer contextual variations. The performance peaks at a rank of 8, where the model demonstrates the most effective balance between expressiveness and stability. This suggests that at this point, the low-rank constraint is sufficiently robust to prevent overfitting while still maintaining flexibility for semantic transformations. However, when r is too large, the low-rank constraint is weakened, and feature transformation diffuses into redundant dimensions, leading to unstable semantics and degraded generalization. In the case of higher ranks, as demonstrated by the gradual decrease in performance after $r = 8$, the model starts to incorporate less relevant features that hinder its ability to generalize well across different tasks. This indicates that performance gains arise not from higher dimensionality but from selecting semantically relevant subspaces. Therefore, we set $r = 8$ as the default for a balanced trade-off between expressiveness and stability.

Table 1. Performance comparison across different backbone scales and fine-tuning settings on the VisDial dataset. BLIP-ft refers to a backbone that is additionally fine-tuned on the VisDial. The results show that our method consistently improves Recall@1, Recall@5, and Recall@10 across all dialogue rounds, regardless of backbone capacity or whether the backbone is fine-tuned, demonstrating the plug-and-play nature of our framework.

Metric	Backbone	Method	1	2	3	4	5	6	7	8	9	10
Recall@1	BLIP-Base	BLIP	37.31	34.45	30.96	28.00	27.03	24.27	22.14	20.49	19.62	17.49
		+ours	**39.78**	**40.46**	**40.41**	**40.46**	**41.38**	**41.42**	**42.54**	**42.93**	**42.44**	**42.54**
		BLIP-ft	40.94	43.85	45.25	46.8	47.67	48.06	49.47	50.39	51.50	51.84
		+ours	**41.72**	**44.09**	**46.12**	**47.77**	**49.08**	**50.39**	**52.42**	**53.05**	**54.46**	**55.43**
	BLIP-Large	BLIP	40.31	38.86	36.68	34.59	33.28	32.22	32.56	30.91	30.23	30.09
		+ours	**41.04**	**43.27**	**44.53**	**44.53**	**45.49**	**45.88**	**46.32**	**45.93**	**46.51**	**46.41**
		BLIP-ft	42.78	45.16	46.85	48.40	49.90	50.78	51.60	52.33	52.52	53.10
		+ours	**43.60**	**45.59**	**48.35**	**49.76**	**51.89**	**53.39**	**55.43**	**56.10**	**56.78**	**57.56**
Recall@5	BLIP-Base	BLIP	58.77	56.49	54.22	49.52	47.00	43.85	43.07	40.84	38.42	35.76
		+ours	**62.98**	**64.49**	**64.53**	**65.41**	**65.12**	**66.76**	**66.47**	**65.79**	**66.04**	**66.96**
		BLIP-ft	64.63	66.81	68.31	70.45	71.27	72.19	72.48	73.16	73.06	73.59
		+ours	**65.55**	**68.27**	**70.20**	**72.43**	**73.79**	**75.05**	**76.55**	**77.13**	**77.47**	**77.23**
	BLIP-Large	BLIP	61.82	59.54	57.95	57.03	55.43	53.59	52.81	52.33	51.41	50.39
		+ours	**64.49**	**64.87**	**66.13**	**67.20**	**66.72**	**67.97**	**68.31**	**68.75**	**68.80**	**69.23**
		BLIP-ft	65.02	67.30	68.70	70.59	72.38	73.30	74.32	75.05	74.90	75.53
		+ours	**66.47**	**68.99**	**71.03**	**73.06**	**74.61**	**76.02**	**76.79**	**78.15**	**78.10**	**79.07**
Recall@10	BLIP-Base	BLIP	68.36	66.28	62.50	59.16	56.88	54.89	52.91	49.81	47.24	44.14
		+ours	**71.37**	**72.19**	**73.89**	**74.90**	**74.13**	**75.92**	**75.78**	**76.16**	**75.68**	**75.53**
		BLIP-ft	73.06	74.76	75.82	78.05	78.92	79.60	79.99	80.52	80.38	81.01
		+ours	**74.18**	**76.11**	**78.05**	**80.52**	**80.77**	**82.17**	**82.17**	**83.43**	**83.62**	**84.40**
	BLIP-Large	BLIP	70.64	68.75	66.62	65.12	64.44	63.23	61.58	61.09	60.47	59.40
		+ours	**73.16**	**73.93**	**75.29**	**76.50**	**74.90**	**77.13**	**76.50**	**77.33**	**76.41**	**77.13**
		BLIP-ft	73.98	76.26	77.28	79.60	80.38	81.54	82.22	82.32	82.46	82.99
		+ours	**75.19**	**77.66**	**79.51**	**81.93**	**82.56**	**83.43**	**84.35**	**84.98**	**84.84**	**86.00**

Table 2. Detailed performance comparison using different dialogue sources on the Recall@10 metric. Our proposed method consistently outperforms previous works, demonstrating its strong adaptability to various dialogue styles.

Dialogue Source	Method	1	2	3	4	5	6	7	8	9	10	Avg
ChatIR (ChatGPT-BLIP2)	BLIP [5]	73.84	71.32	70.83	69.43	68.51	67.15	66.76	65.94	63.95	62.84	68.06
	ChatIR [4]	76.79	78.25	79.17	79.41	79.99	80.23	80.43	81.01	80.86	80.86	79.70
	PlugIR [3]	73.64	74.66	74.66	75.24	75.97	76.84	75.63	75.87	74.76	74.32	75.16
	Imagescope [7]	69.96	69.28	70.00	69.77	68.27	67.54	67.20	68.56	66.81	66.28	68.37
	Ours	**76.94**	**79.07**	**80.38**	**80.81**	**81.20**	**82.95**	**83.09**	**84.01**	**83.53**	**83.72**	**81.57**
ChatIR (FLAN-BLIP2)	BLIP	73.98	73.35	69.91	65.16	61.05	56.98	51.79	47.09	43.70	40.21	58.32
	ChatIR	75.29	77.23	77.33	77.23	76.70	76.36	75.82	75.48	75.10	74.03	76.06
	PlugIR	74.18	74.66	74.32	73.69	73.45	73.59	73.59	73.64	73.45	72.92	73.75
	Imagescope	69.57	69.96	70.01	68.07	67.15	66.57	65.50	65.46	65.36	65.07	67.27
	Ours	**76.60**	**78.44**	**78.78**	**78.34**	**77.76**	**77.47**	**77.42**	**77.08**	**76.89**	**76.79**	**77.56**
ChatIR (Human-BLIP2)	BLIP	69.91	69.04	66.28	65.16	62.69	61.82	61.14	60.42	59.06	57.90	63.34
	ChatIR	73.59	75.29	76.60	76.84	77.13	77.62	78.29	78.44	78.68	79.22	77.17
	PlugIR	70.06	71.12	70.93	72.53	71.66	71.41	71.95	71.90	70.78	71.37	71.37
	Imagescope	65.75	65.60	66.13	65.55	63.91	62.74	61.14	62.21	60.56	61.14	63.47
	Ours	**73.89**	**75.97**	**76.60**	**77.42**	**78.88**	**78.54**	**79.55**	**80.52**	**80.91**	**81.49**	**78.38**
PlugIR (ChatGPT-BLIP2)	BLIP	74.56	74.52	74.08	74.13	73.35	72.53	70.54	68.70	67.64	66.62	71.67
	ChatIR	76.50	78.29	80.18	80.52	81.73	81.88	82.07	82.27	82.75	82.80	80.90
	PlugIR	74.27	75.48	76.26	76.11	76.79	76.11	76.02	75.05	74.37	74.32	75.48
	Imagescope	69.57	70.45	70.11	70.16	69.33	69.14	67.54	68.31	67.05	66.52	68.82
	Ours	**77.71**	**79.26**	**80.23**	**81.30**	**82.46**	**83.04**	**83.24**	**83.28**	**83.48**	**83.58**	**81.76**

Method	BLIP	ChatIR	PlugIR	ImageScope	Ours
BRI	1.034	1.003	0.887	0.942	**0.879**

Table 3. Performance comparison in terms of BRI. Lower scores indicate better performance.

E. Performance across Backbone Scales

To further assess the stability and scalability of the proposed framework under different backbone capacities and fine-tuning settings, we conducted systematic experiments on both the BLIP-Base and BLIP-Large [5] variants, while also considering whether the backbone was fine-tuned on VisDial [2]. Here, BLIP-ft refers to a backbone that is additionally fine-tuned on the VisDial dataset prior to being incorporated into our framework. The complete round-wise results are presented in Table 1. The results show that, regardless of whether a smaller Base backbone or a more powerful Large backbone is used, our method consistently yields stable and significant improvements across all metrics. More importantly, these performance gains remain consistent across different backbone scales, indicating that the proposed method does not rely on a specific backbone capacity, it can serve as a general enhancement module applicable to encoders of various sizes.

F. Complete Results

In this section, we provide the complete results for all dialogue variants used in the main paper. Specifically, we report the per-round Recall@10 performance across the full 10-round interactive retrieval process for the four dialogue sources [3, 4]. These detailed results correspond to the averaged values presented in the main paper and offer a comprehensive view of the round-wise behavior of different methods. Show in Table 2 across all dialogue sources and all rounds, our method consistently achieves the best overall Recall@10 performance, demonstrating that the proposed context-aware dynamic latent space transformation enables stable improvements throughout the entire interaction process. Notably, the advantages become more pronounced in later rounds, where the evolving dialogue context plays a more critical role in guiding the retrieval model.

In addition to Recall-based metrics, we also provide the complete per-round Hit@K results on the VisDial dataset. Although Recall@K serves as our primary evaluation metric due to its stronger ability to measure retrieval coverage across rounds, the Hit@K results further confirm the competitiveness of our method. Show in Table 5, our Hit-based performance remains comparable to the best-performing existing methods, indicating that CAST improves Recall without sacrificing hit-oriented retrieval accuracy.

Training Configurations	
epoch	50
total batch size	1024
learning rate schedule	exponential decay
optimizer	AdamW
temperature	0.03
weight decay	1×10^{-4}
learning rate	5×10^{-4}
rank	8

Table 4. Configuration for Training

G. BRI Performance

Table 3 shows the comparison of BRI from round 1 to round 10 on the VisDial dataset across different retrieval methods. CAST maintains its advantage under this evaluation and achieves the best performance.

H. Implementation Details

We use the BLIP-Large version as the backbone model to extract text and image features. During training, only our CASR module is trained. In the CLP module, two separate two-layer MLPs are used to generate the low-rank matrices A and B, respectively, while the CGM is implemented as a three-layer MLP. All reconstructed datasets are generated by ChatGPT-3.5-turbo, consistent with the reconstruction method used previously. The corresponding prompt is also shown in Table 6. During training, we apply a 20% probability to randomly mask the caption part of samples with longer dialogues. More detailed training configurations are shown in Table 4.

I. More Visualization Results

In addition to the visualizations presented in the main paper, we provide two supplementary examples in Figure 4b to further demonstrate the effect of our context-aware latent space transformation. In the first example, we select 100 images most related to the concept "a cat on the grass." When the dialogue context input condition is "a black color cat on the grass," the transformed feature space clearly emphasizes the black color attribute, causing images containing black cats to form a much tighter and more coherent cluster compared to the original space. In the second example, we collect 100 images associated with the broader category "vehicle." When the dialogue input condition is "a red vehicle," the transformed latent space reorganizes the embeddings such that red vehicles, initially dispersed across the original manifold, become significantly more concentrated. These observations demonstrate that our method dynamically transforms the embedding geometry according to the given input conditions, enabling fine-grained semantic separation aligned with the intended dialogue semantics.

Table 5. Performance comparison using Hits@K. Our method achieves performance comparable to the state-of-the-art.

Metric	Method	1	2	3	4	5	6	7	8	9	10	Avg.
Hits@1	BLIP [5]	40.31	44.48	48.01	50.53	52.33	53.73	54.75	55.62	56.40	57.03	51.32
	ChatIR [4]	42.78	45.78	49.22	51.60	53.44	55.14	56.40	57.70	59.06	60.51	53.16
	PlugIR [3]	40.20	46.80	50.90	54.60	58.20	61.00	63.80	65.70	67.20	**68.70**	**57.71**
	ImageScope [7]	32.17	44.62	51.41	**55.33**	**58.96**	**62.11**	**64.29**	**66.09**	**67.30**	68.56	57.08
	Ours	**43.60**	**48.06**	**51.79**	54.55	57.46	59.54	61.34	62.84	64.53	65.79	56.95
Hits@5	BLIP	61.82	65.65	68.51	70.69	72.14	73.21	73.98	74.61	74.90	75.82	71.13
	ChatIR	65.41	69.43	71.37	74.08	76.07	77.81	78.92	80.33	80.67	81.40	75.55
	PlugIR	63.08	69.19	73.40	75.92	78.20	80.14	81.15	82.56	84.11	84.93	77.27
	ImageScope	57.46	67.25	71.61	74.56	77.62	79.75	81.30	82.27	83.04	84.16	75.90
	Ours	**66.47**	**71.12**	**74.27**	**77.18**	**79.36**	**81.25**	**82.61**	**83.77**	**84.54**	**85.56**	**78.61**
Hits@10	BLIP	70.64	74.42	76.45	78.44	79.84	81.20	81.64	82.17	82.51	83.19	79.05
	ChatIR	74.81	78.20	80.33	82.17	83.82	85.17	86.09	86.87	87.35	87.89	83.27
	PlugIR	72.24	78.00	81.40	83.77	**85.90**	**87.40**	**88.42**	88.95	**89.68**	**90.50**	84.63
	ImageScope	68.36	75.92	79.31	81.93	84.21	85.71	86.82	87.50	88.37	89.09	82.72
	Ours	**75.19**	**78.97**	**81.83**	**84.11**	85.71	87.11	88.37	**89.10**	89.63	90.21	**85.02**

Table 6. Prompt template used to reconstruct a new caption from the original caption and dialogue history.

System	Your role is to reconstruct the [Caption] with the additional information given by following [Dialogue]. The reconstructed [New Caption] should be concise and in appropriate form to retrieve a target image from a pool of candidate images.
User	[Caption]: a woman sits on a bench holding a guitar in her lap [Dialogue]: is this in a park? yes, i believe it is, are there others around? no, she is alone, does she have a collection bucket? no, is her hair long? yes, pretty long, is she wearing a dress? i don't think so, hard to tell, does she have shoes on? yes, flip flops, is there grass nearby? yes, everywhere, is it a sunny day? yes, are there trees? in the background there are trees, is the guitar new? i don't think so [New Caption]:
Assistant	a woman with pretty long hair sits alone on a grassy bench in a park on a sunny day, holding a guitar in her lap without a collection bucket, wearing flip flops, with trees in the background, with a slightly worn guitar
User (Query)	[Caption]: {caption} [Dialogue]:{dialogue} [New Caption]:

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[2] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 1, 4

[3] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive textto-image retrieval with large language models: A plugand-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 11–16, 2024. 1, 3, 4, 5

[4] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36:61437–61449, 2023. 1, 3, 4, 5

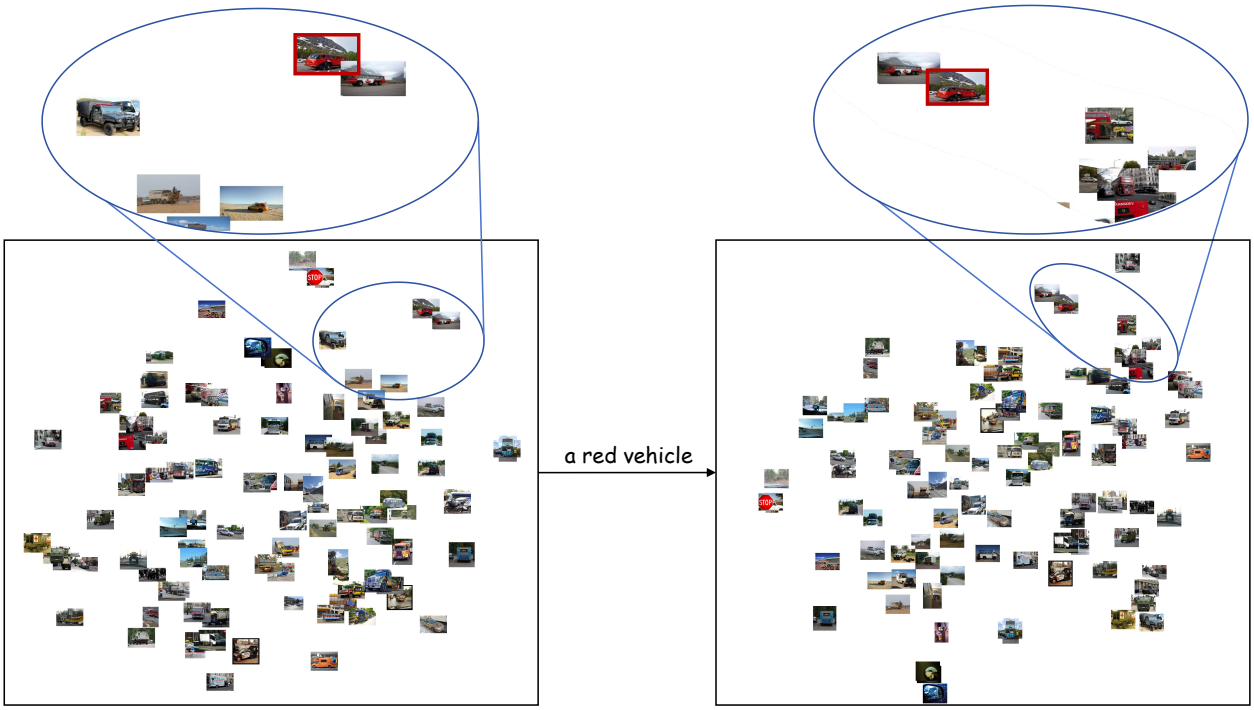
[5] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3, 4, 5

[6] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2

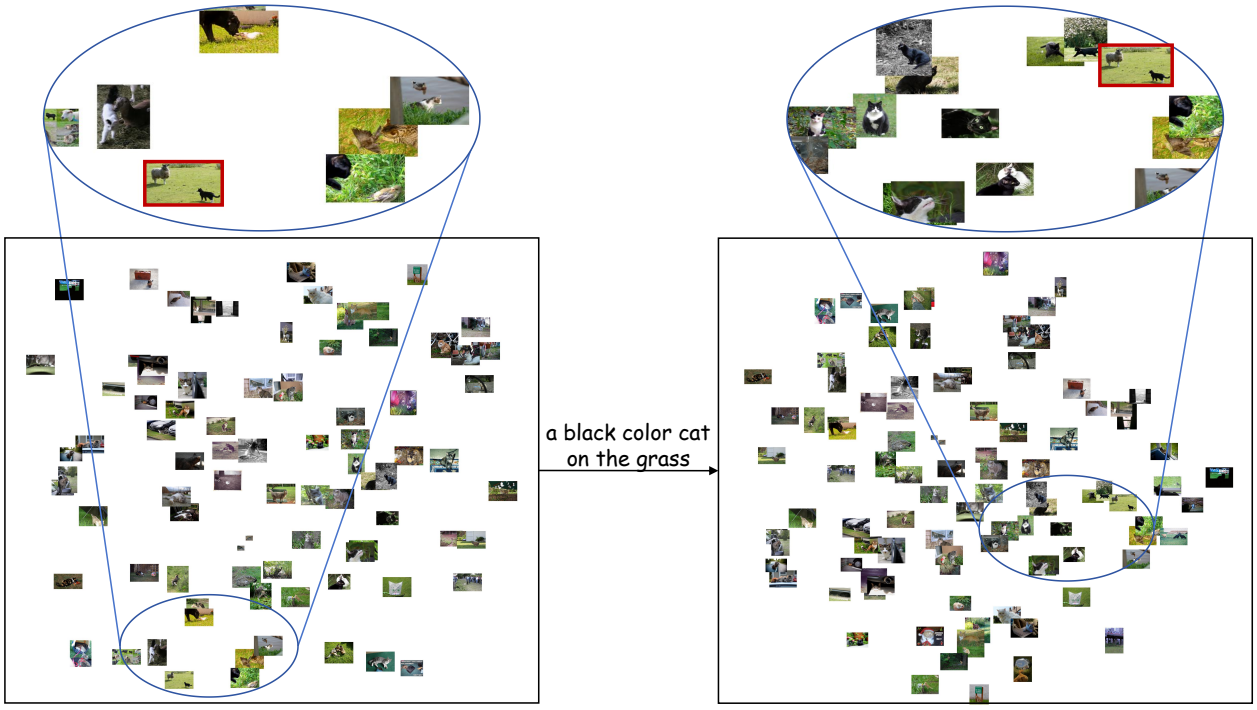
[7] Pengfei Luo, Jingbo Zhou, Tong Xu, Yuan Xia, Linli Xu, and Enhong Chen. Imagescope: Unifying language-guided image retrieval via large multimodal model collective reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 1666–1682, 2025. 1, 3, 5

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 6

[9] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1, 2023. 2



(a)



(b)

Figure 4. More space transformation visualization results, best viewed zoomed in. T-SNE [8] visualizations of 100 images related to (a) "vehicle" and (b) "a cat on the grass." A red bounding box highlights a sample image with a zoomed-in neighborhood to show how the latent space changes under different dialogue input conditions. For the "vehicle" case, when the input condition is "a red vehicle," the feature space reorganizes toward color distinctions, causing red vehicles to cluster more tightly. For the "cat" case, when the input condition is "a black color cat on the grass," the space similarly brings images of black cats closer together. These examples demonstrate how different input conditions dynamically shape the structure of the embedding space.