
Visual Abstraction: A Plug-and-Play Approach for Text-Visual Retrieval

Guofeng Ding¹ Yiding Lu¹ Peng Hu¹ Mouxing Yang¹ Yijie Lin^{†1} Xi Peng^{†12}

<https://github.com/XLearning-SCU/2025-ICML-VISA>

Abstract

Text-to-visual retrieval often struggles with semantic redundancy and granularity mismatches between textual queries and visual content. Unlike existing methods that address these challenges during training, we propose VISual Abstraction (VISA), a test-time approach that enhances retrieval by transforming visual content into textual descriptions using off-the-shelf large models. The generated text descriptions, with their dense semantics, naturally filter out low-level redundant visual information. To further address granularity issues, VISA incorporates a question-answering process, enhancing the text description with the specific granularity information requested by the user. Extensive experiments demonstrate that VISA brings substantial improvements in text-to-image and text-to-video retrieval for both short- and long-context queries, offering a plug-and-play enhancement to existing retrieval systems.

1. Introduction

The purpose of abstraction is not to be vague, but to create a new semantic level in which one can be absolutely precise.

Edsger Wybe Dijkstra

Text-to-visual retrieval (TVR) attempts to identify relevant images or videos based on a textual query. In this field, vision-language pre-training models (VLMs) (Radford et al., 2021; Jia et al., 2021) have made significant progress by employing contrastive learning on large-scale paired data to construct a cross-modal joint space. Despite their suc-

¹School of Computer Science, Sichuan University, Chengdu, China ²National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, China. [†]Correspondence to: Yijie Lin <linyijie.gm@gmail.com>, Xi Peng <pengx.gm@gmail.com>.

cess, we observe that traditional VLMs struggle to satisfy the precise and diverse retrieval demands of users, due to limitations in both training strategies and data. Specifically,

i) Training Strategy Aspect. Retrieval tasks require the model to focus on core semantic elements of an image based on the query while disregarding superfluous details. However, VLMs like CLIP typically contrast text and images at a global level to emphasize overall relationships. This global strategy inevitably incorporates low-level visual elements such as snow textures or tree branches, which are irrelevant to high-level semantics. These redundant features can hinder the model’s ability to understand key concepts, causing it to incorrectly attend to unimportant patches, such as the tire tracks (see red rectangle in Query 1 of Fig. 1(a)).

ii) Training Data Aspect. As the saying goes, “an image is worth a thousand words”. While images inherently exhibit unlimited granularity (Tang et al., 2023), the web-crawled image descriptions are often brief and coarse (Xiao et al., 2024), failing to capture the multi-granular concepts in images. However, retrieval tasks require models to dynamically adjust their focus based on the granularity of the user query. This mismatch prevents the model from effectively matching user queries of varying granularities with the corresponding image content. As shown in Queries 2 and 3 with different granularities in Fig. 1(a), VLM predominantly focuses on the main action, skiing, but fails to capture fine-grained details such as the hat and jacket.

To address these challenges, early works (Li et al., 2020; Chen et al., 2020; Lee et al., 2018) utilize object detection techniques (Ren et al., 2016) to extract key visual concepts, filtering out low-level details to cover the essential semantics of an image. However, these methods are inherently limited by a finite number of objects (Li et al., 2022a) and cannot explicitly represent relationships, actions, or scenes. Recent works (Zheng et al., 2024) attempt to address the varying granularity demands by leveraging large multimodal models (Liu et al., 2024a; Hurst et al., 2024) to generate fine-grained captions, enriching text data with multi-granularity information (Cao et al., 2024).

However, the above methods require training VLMs from scratch, which is a complex and non-trivial endeavor. Un-

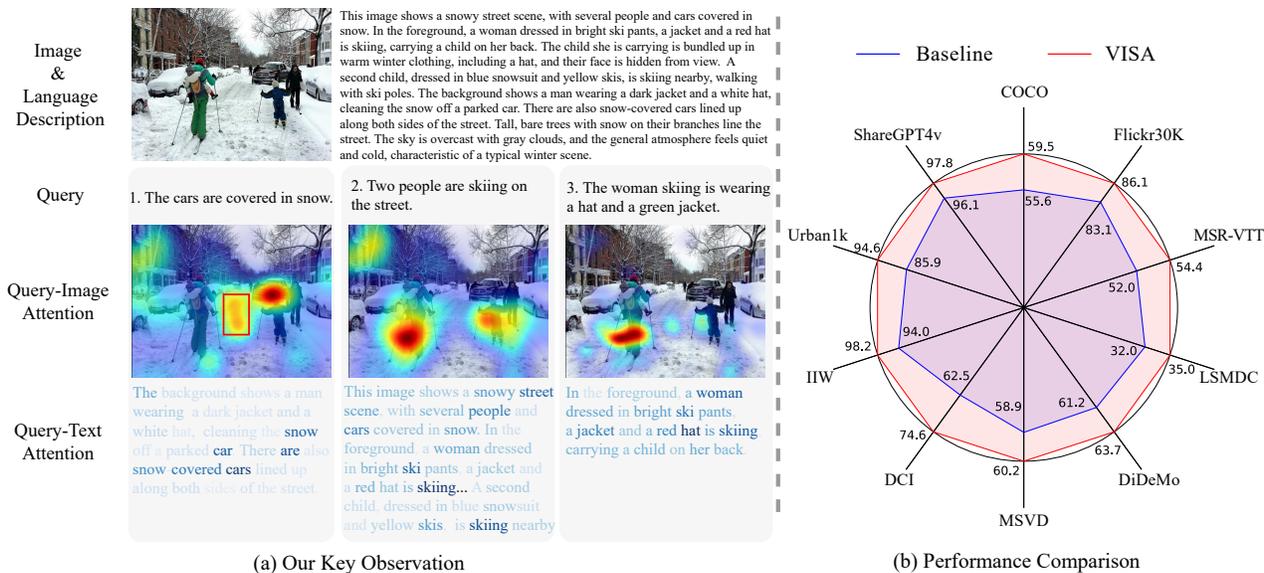


Figure 1. (a) Our key observation. We visualize the attention maps of vision-language pre-training model LoTLIP (Wu et al., 2024) and text retriever gemma2 (Chen et al., 2024a) given diverse text queries. For clarity, only a subset of the query-text attention is shown here; full attention details are available in Fig. 6. (b) Performance comparison (R@1) across ten image and video datasets.

like previous works, we explore a test-time solution that addresses these two challenges and enhances the retrieval capabilities of VLMs. The core idea is to transform the retrieval process from a cross-modal space to a text-only space, inspired by our observation in the last row of Fig. 1(a). Specifically, we observe that text descriptions of images, with dense semantics, naturally omit the low-level redundant information that often present in visual signals and enable a sharper focus on the core concepts (Query 1). Moreover, when textual descriptions encompass comprehensive conceptual details, they can seamlessly match user queries at varying granularities (Queries 2 and 3), effectively satisfying diverse retrieval demands.

Inspired by this observation, we propose VISual Abstraction (VISA), a plug-and-play approach that abstracts sparse visual concepts in images into dense natural language to enhance text-to-visual retrieval. Specifically, VISA leverages off-the-shelf large multimodal model (LMM) and text retrievers to transform visual signals into textual descriptions and perform similarity measurements within the textual space. By abstracting visual content into textual representations, VISA filters out irrelevant low-level details and focuses on the core concepts. To match the required granularity, VISA employs a question-answering strategy that decomposes the user query into targeted questions and uses off-the-shelf LMMs to answer these for each candidate. The resulting answers better align the visual abstraction with the user query. The contribution of this paper is summarized as,

- We propose to enhance text-to-visual retrieval by abstract-

ing the visual concepts into natural language through a question-answering process, possibly providing a new perspective towards advancing visual retrieval during test-time.

- By employing off-the-shelf large models, our method works in a plug-and-play manner, achieving significant improvements for both short- and long-context queries.

2. Related Work

2.1. Vision-language Pre-training Model in Text-visual Retrieval

Vision-language pre-training models (Radford et al., 2021; Lin et al., 2024; Huang et al., 2024) have emerged as the cornerstone for text-visual retrieval in recent years. Based on the differences in architectural design, existing approaches can be roughly categorized into three groups: i) Single-stream architectures like UNITER (Chen et al., 2020) and OSCAR (Li et al., 2020) use object detectors for semantic feature extraction and fine-grained fusion of visual and textual data. However, the high computational overhead and limited concepts of object detector hinder their scalability to large-scale datasets. ii) Dual-stream architectures, such as CLIP (Radford et al., 2021) and Frozen (Bain et al., 2021), employ separated encoders for each modality, achieving high-level semantic alignment but struggling with fine-grained cross-modal interactions. iii) Hybrid architectures, such as ALBEF (Li et al., 2021) and InternVideo (Wang et al., 2022; 2024b), take the best of both worlds to balance

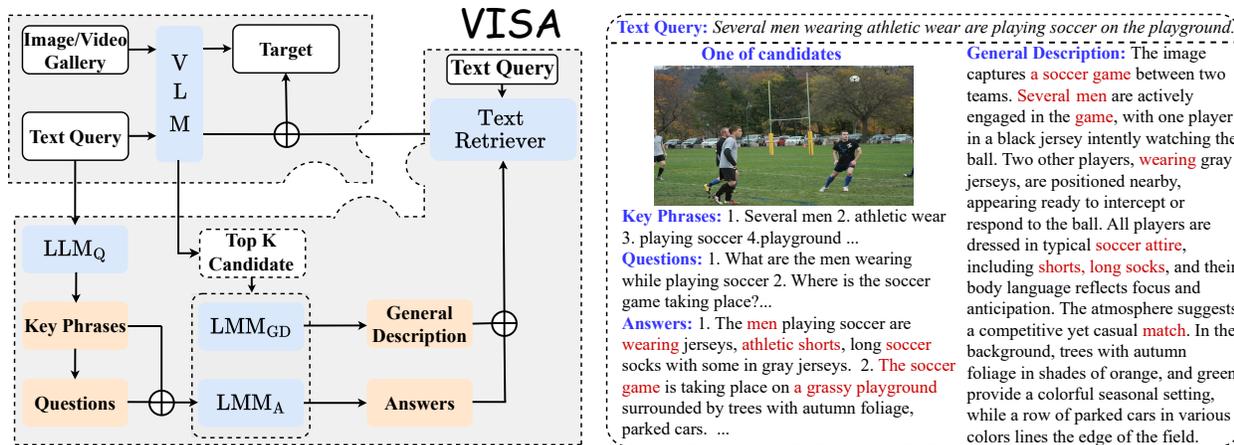


Figure 2. The pipeline of our method VISA. (Left) VISA operates as a plug-and-play enhancement for existing VLMs, transforming visual content into text descriptions to improve text-to-visual retrieval within the text space. The visual abstraction consists of a general description and the answers from QA-based description refinement. (Right) An illustrated example where query-related information are highlighted in red. Additional examples are provided in Appendix F.

fine-grained alignment with computational efficiency.

Although existing VLMs achieve significant processes in TVR, they often fail to capture multi-granularity semantics inherent in the visual modality, resulting in suboptimal retrieval performance. To address these limitations, we abstract visual content into textual descriptions to provide comprehensive semantic coverage across different levels of granularity. Notably, the proposed VISA framework is orthogonal to existing VLM-based methods and serves as a plug-and-play module, boosting retrieval performance across various architectures as demonstrated in our experiments.

2.2. Large Models in Text-to-visual Retrieval

In very recent, large language models (LLMs) and LMMs have opened new avenues for cross modal retrieval due to their strong understanding capabilities. These approaches can be categorized into three groups: i) Augmentation with LMM-generated captions (Zheng et al., 2024; Xiao et al., 2024; Shi et al., 2024; Lu et al., 2025), which enhance existing VLMs by leveraging enriched datasets containing fine-grained captions generated by LMMs. ii) Augmentation with frozen LLMs (Cao et al., 2024), which extends the input length and contextual understanding of LLMs to VLMs, improving retrieval performance for complex queries or long text inputs. iii) Augmentation by fine-tuning LMMs (Chen et al., 2024c; Liu et al., 2024c), which transforms LMMs into sophisticated visual-language re-rankers, leveraging their advanced cross-modal reasoning capabilities to refine retrieval results.

The aforementioned methods rely on resource-intensive training to improve retrieval performance, requiring sub-

stantial computational resources to retrain or fine-tune models. In contrast, this work introduces a test-time computing approach to enhance the retrieval performance of existing VLMs. Specifically, our method leverages off-the-shelf LMMs and text retriever to transform visual content into abstract textual descriptions, providing a plug-and-play solution that improves retrieval performance without additional training. Furthermore, by utilizing LLM-based text retriever, our approach benefits from their advanced contextual understanding to handle long-context queries.

3. Method

In this section, we introduce VISA, a plug-and-play approach designed to enhance text-to-visual retrieval during test-time. First, we explain how VISA integrates seamlessly into existing VLMs for text-to-visual retrieval in Sec. 3.1. Then, we detail the process of abstracting visual content into dense natural language in Sec. 3.2.

3.1. Enhancing Text-to-visual Retrieval with VISA

Text-to-visual retrieval aims to identify the most relevant image or video based on a textual user query. Formally, given a user query q and a visual gallery $\mathcal{G} = \{I_1, I_2, \dots, I_m\}$, where I represents an image or video, VLMs typically encode both modalities into a shared cross-modal space. This space is designed to capture the semantic content of both the textual query and the candidate visual samples. The matching score for the i -th sample is computed as:

$$s(I_i|q) = \text{normalization}(\text{VLM}(I_i, q)), \quad (1)$$

where $\text{VLM}(I_i, q)$ denotes the similarity between candidate I_i and query q computed by a specific VLM. The function

normalization rescales similarity scores to $[0, 1]$ using min-max scaling based on the scores among all candidates.

Despite the effectiveness of VLMs in text-to-visual retrieval, they often struggle with granularity mismatch and semantic inconsistency between text and visual content. In brief, the general embeddings learned by VLMs may overlook fine-grained details, leading to suboptimal ranking, particularly when user queries require various granularity reasoning beyond the high-level semantic alignment captured by VLMs. To address this limitation, we introduce VISA, a plug-and-play approach that enhances text-to-visual retrieval by reranking the top- k retrieved candidates $\mathcal{G}_{\text{top-}k}$. Specifically, VISA abstracts visual content into textual descriptions and performs retrieval within the text space, allowing for better alignment with user queries.

Let T_i represent the visual abstraction of I_i . The matching score of the i -th sample in the text space is computed as:

$$s(T_i|q) = \text{normalization}(\text{Text-Retri}(T_i, q)), \quad (2)$$

where $\text{Text-Retri}(T, q)$ denotes the similarity score between the query q and the candidate text T_i , calculated by an off-the-shelf text retriever within a unified text space. The final retrieval result I^* is obtained by combining the matching scores from both the VLM and the text retriever:

$$I^* = \underset{I_i \in \mathcal{G}_{\text{top-}k}}{\text{argmax}} (s(I_i|q) + s(T_i|q)), \quad (3)$$

3.2. Visual Abstraction

As illustrated in Fig. 2, VISA consists of two key processes: i) General Description Generation: we employ off-the-shelf LMMs to generate long-form descriptions for each of the top- k retrieved candidates. This process filters out low-level redundancies in visual signals, producing concise yet information-rich descriptions that better align with textual queries. ii) QA-Based Description Refinement: while general descriptions provide a broad contextual overview, they may lack the specific granularity details required by the query. To refine these descriptions, we introduce a question-answering (QA) process guided by a chain-of-thought (CoT) strategy (Wei et al., 2023).

3.2.1. GENERAL DESCRIPTION GENERATION

To construct a semantically rich textual representation of each candidate, we employ off-the-shelf LMMs to generate a long-form general description of the visual content. Formally, the general description T^{GD} is derived as:

$$T_i^{\text{GD}} = \text{LMM}_{\text{GD}}(\text{Prompt}_{\text{GD}}, I_i), \quad \forall I_i \in \mathcal{G}_{\text{top-}k} \quad (4)$$

where the prompt for LMM is provided in Table 8, utilizing the default prompt settings of these models to generate detailed descriptions.

This step offers two key advantages: i) Eliminating low-level redundancies. Unlike raw visual signals, textual descriptions naturally filter out irrelevant details such as background noise and texture redundancies, leading to a more semantically compact representation. ii) Aligning visual representations with text queries. By converting sparse visual signals into dense language descriptions, we create representations that are more directly comparable to textual queries, enhancing retrieval consistency within a unified text space.

3.2.2. QA-BASED DESCRIPTION REFINEMENT

While the general description effectively removes low-level redundancies, it may lack the necessary granularity to fully align with the user query. To address this, we introduce a question-answering process that refines the description by extracting query-relevant details, ensuring a more precise alignment between the visual abstraction and the user query.

This process mitigates the semantic granularity discrepancy by ensuring that descriptions focus on the most relevant aspects of the query. To achieve this, we employ a chain-of-thought approach, where an LLM first analyzes the user query to extract key phrases and generates targeted questions based on these phrases. An LMM then examines each visual candidate and provides context-aware answers, determining whether the extracted key phrases are present in the image or video.

Question generation with key phrases. As shown in chain-of-thought prompt Prompt_Q in Table 10, we first identify and extract key elements from the user query, focusing on objects, attributes, actions, locations, and interactions to accurately capture the core visual concepts. Once these key phrases are identified, the system generates query-specific questions based on them:

$$\text{Questions} = \text{LLM}_Q(\text{Prompt}_Q, \text{key-phrases}, q), \quad (5)$$

where the Prompt_Q is carefully designed to ensure the questions meet the following criteria:

- **Clearly Answerable.** The generated questions are designed so that the LMM can provide definitive answers based on the visual content. This ensures that the presence or absence of queried objects, attributes, or actions can be explicitly confirmed or denied, reducing ambiguity and preventing speculative or overly broad responses.
- **Consistent Granularity.** The questions are strictly derived from the key phrases extracted from the user query, ensuring that each question targets a specific semantic aspect of the visual content. By instructing LLM to avoid direct repetition or close paraphrasing of the original query in Prompt_Q , the questions guides the answering process toward distinct details, enabling the description of visual

Table 1. Image retrieval results on COCO and Flickr30K. * indicates results are re-evaluated using official checkpoints from HuggingFace.

	COCO			Flickr30K		
	R@1	R@5	R@10	R@1	R@5	R@10
CoCa (Yu et al., 2022)	51.2	74.2	82.0	80.4	95.7	97.7
FLAME (Cao et al., 2024)	43.9	70.4	79.7	73.3	91.7	95.5
DreamLIP (Zheng et al., 2024)	41.1	67.0	76.6	66.4	88.3	93.3
FLAIR (Xiao et al., 2024)	53.3	77.5	-	81.1	94.9	-
UMG-CLIP (Shi et al., 2024)	57.5	80.4	87.3	81.2	95.7	97.7
RAGVL (Chen et al., 2024c)	-	-	-	84.4	95.2	96.3
InternVL-G (Chen et al., 2024b)	58.6	81.3	88.0	85.0	97.0	98.6
SigLIP* + EVA-CLIP*	56.0	79.1	85.9	84.1	96.7	98.3
BLIP-2 (Li et al., 2023)	-	-	-	89.7	98.1	98.9
SigLIP* (Zhai et al., 2023)	54.2	76.8	84.2	83.0	96.1	98.0
SigLIP* + VISA (Ours)	57.2	80.3	86.9	85.1	97.1	98.6
Δ	+3.0	+3.5	+2.7	+2.1	+1.0	+0.6
EVA-CLIP* (Sun et al., 2023)	55.6	77.9	85.2	83.1	95.8	97.9
EVA-CLIP* + VISA (Ours)	59.5	81.2	87.5	86.1	97.3	98.6
Δ	+3.9	+3.3	+2.3	+3.0	+1.5	+0.7

content across varying levels of granularity, from high-level scenes to fine-grained attributes.

Answering based on key phrases. To supplement the details required by the user, we use an LMM to answer questions for each top- k candidate sample. Formally, the corresponding answer is obtained as:

$$T_i^A = \text{LMM}_A(\text{Prompt}_A, \text{key-phrases}, \text{questions}, I_i), \quad (6)$$

where the key phrases are also provided to guide the LMM in generating context-aware answers. As shown in Table 9, the prompt Prompt_A is carefully designed to ensure:

- Encouraging Detailed Responses. The model is explicitly instructed to avoid short or binary answers (e.g., “Yes” or “No”) and instead provide complete, context-rich responses that enhance text retrieval.
- Ensuring Relevance. The model is guided to consider the relationship between the candidate and key phrases, avoiding out-of-context or irrelevant answers.
- Handling Uncertainty. If visual content does not provide a clear answer, the model is explicitly instructed to respond with “Uncertain” and such responses are subsequently discarded. This is crucial in preventing hallucinated or misleading information, ensuring responses remain the actual content rather than introducing irrelevant details.

By employing this CoT-based QA approach, VISA ensures that varying granularity details are effectively captured. The final visual abstraction T_i is constructed by concatenating the general description T_i^{GD} with the QA-generated answers T_i^A for the i -th candidate, ensuring query-consistent granularity in retrieval. The final retrieval results are then obtained through Eqs. 2 and 3.

4. Experiments

In this section, we first describe the implementation details of VISA. Next, we integrate VISA into various VLMs and compare its performance with state-of-the-art approaches across ten datasets. Finally, we conduct detailed analyses and ablation studies to evaluate the effectiveness and robustness of our method. Due to space limitations, additional experimental details and results, including visualization examples, are provided in Appendices A to F.

4.1. Experimental Settings

Unless stated otherwise, we utilize LMMs LLaVA-v1.6-34B (Liu et al., 2024a) and LLaVA-Video-32B (Zhang et al., 2024b) to obtain the general descriptions of images and videos, respectively. For the question-answering process, we employ LLM Qwen2.5-32B (Team, 2024) as the question generator, LMM Qwen2-VL-7B (Wang et al., 2024a) as the answer generator, and gemma2 (Chen et al., 2024a) as the text retriever. For video data, the frames per second (FPS) input to LMMs is set to 3. The size of the reranking gallery (k) is set to 20 and the number of questions is 3 for all datasets. For all evaluated models listed in Appendix E, we use the default hyper-parameters provided on HuggingFace. All experiments are conducted on Ubuntu 20.04 with NVIDIA 4090 GPUs.

4.2. Main Results

We evaluate the effectiveness of our proposed method, VISA, on both short- and long-context retrieval tasks for images and videos.

Table 2. Video retrieval results on MSR-VTT, LSMDC, DiDeMo and MSVD.

	MSR-VTT			LSMDC			DiDeMo			MSVD		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP4Clip (Luo et al., 2021)	32.0	57.0	66.9	15.1	28.5	36.4	-	-	-	38.5	66.9	76.8
InternVideo (Wang et al., 2022)	40.7	-	-	17.6	32.4	40.2	31.5	57.6	68.2	43.4	-	-
BT-Adapter (Liu et al., 2024b)	40.9	64.7	73.5	19.5	35.9	45.0	35.6	61.9	72.6	-	-	-
vid-TLDR (Choi et al., 2024)	42.1	63.9	72.4	-	-	-	52.0	74.0	81.0	50.0	77.6	85.5
UMT-L (Li et al., 2024)	42.6	64.4	73.1	25.2	43.0	50.5	48.6	72.9	79.0	49.0	76.9	84.7
LanguageBind (Zhu et al., 2024)	42.8	67.5	76.0	-	-	-	39.7	65.5	73.8	54.1	81.1	88.1
mPLUG-2 (Xu et al., 2023)	47.1	69.7	79.0	24.1	43.8	52.0	45.7	71.1	79.2	-	-	-
VAST (Chen et al., 2023b)	49.3	68.3	73.9	-	-	-	55.5	74.3	79.6	-	-	-
InternVideo2-C* (Wang et al., 2024b)	46.0	70.3	79.3	24.3	42.8	50.5	45.9	71.8	79.8	55.1	81.2	87.9
InternVideo2-C* + VISA (Ours)	48.8	73.7	80.9	28.3	47.3	55.1	54.8	78.8	84.0	57.8	83.2	89.1
Δ	+2.8	+3.4	+1.6	+4.0	+4.5	+4.6	+8.9	+7.0	+4.2	+2.7	+2.0	+1.2
InternVideo2-G* (Wang et al., 2024b)	52.0	74.6	81.8	32.0	52.4	59.4	61.2	82.4	87.3	58.9	83.0	88.7
InternVideo2-G* + VISA (Ours)	54.4	75.3	82.9	35.0	53.1	61.0	63.7	86.0	89.7	60.2	84.5	89.7
Δ	+2.4	+0.7	+1.1	+3.0	+0.7	+1.6	+2.5	+3.6	+2.4	+1.3	+1.5	+1.0

4.2.1. SHORT-CONTEXT IMAGE RETRIEVAL

We evaluate our approach on the MS-COCO (Chen et al., 2015) and Flickr30K (Plummer et al., 2015) datasets. MS-COCO includes 5,000 test images, each annotated with five manual annotated captions describing a diverse range of objects. Flickr30K emphasizes real-world scenarios with complex interactions, providing five detailed annotations for each of its 1,000 test images, capturing object relationships and actions.

We first evaluate the effectiveness of VISA based on widely-used VLMs SigLIP (Zhai et al., 2023) and EVA-CLIP (Sun et al., 2023). As shown in Table 1, our approach achieves substantial improvements in image retrieval, particularly in terms of R@1. For instance, applying VISA to SigLIP (approximately 1B parameters) boosts R@1 from 54.2% to 57.2% on MS-COCO and from 83.0% to 85.1% on Flickr30K. Similarly, integrating VISA with the much larger 18B-parameter EVA-CLIP raises R@1 from 55.6% to 59.5% on MS-COCO and from 83.1% to 86.1% on Flickr30K. Note that BLIP-2 is finetuned on COCO and zero-shot transferred to Flickr30K, contributing to the enhanced zero-shot performance on Flickr30K (Chen et al., 2024b).

Notably, even combining SigLIP and EVA-CLIP, which together contain a massive number of parameters, underperforms compared to our approach. This suggests that larger-scale VLMs do not inherently guarantee better performance, highlighting the efficacy of our carefully designed visual abstraction strategy.

Our approach also achieves superior performance compared to existing methods that leverage large models for retrieval (as discussed in Sec. 2.2). Specifically, VISA out-

perform methods like DreamLIP and FLAIR, which generate long textual descriptions of images to enrich semantics. Additionally, our method surpasses reranking approach RAGVL (Chen et al., 2024c) that fine-tune LMMs, as well as FLAME (Cao et al., 2024) that distill knowledge from frozen LLMs. This illustrates that employing large models at test time can effectively enhance retrieval performance without requiring extensive retraining or fine-tuning.

4.2.2. SHORT-CONTEXT VIDEO RETRIEVAL

For the video retrieval task, we evaluate our approach on four widely-used datasets: MSR-VTT (Jun et al., 2016), DiDeMo (Hendricks et al., 2017), LSMDC (Rohrbach et al., 2015), and MSVD (Chen & Dolan, 2011). MSR-VTT spans a broad range of scenarios, including sports, music, cooking, and social activities. DiDeMo provides sentence-level annotations describing different time segments of a video, which are merged into a single query for video-paragraph retrieval. LSMDC consists of clips from movies across various genres, annotated by a concise description of key events and interactions. MSVD includes YouTube videos capturing everyday activities and events. These datasets contain diverse scenarios and annotation styles, offering a comprehensive benchmark for evaluating video retrieval tasks.

We employ InternVideo2-C and InternVideo2-G (Wang et al., 2024b) as the base VLM model. InternVideo2-C is a dual-stream architecture, while InternVideo2-G is a single-dual hybrid model built upon InternVideo2-C. As shown in Table 2, our method consistently achieves higher R@1 scores across all datasets compared to competing approaches. For the dual-stream InternVideo2-C, our method improves R@1 from 46.0% to 48.8% on MSR-VTT and

Table 3. Long-context retrieval results on DCI, IIW, Urban1k and ShareGPT4v.

	DCI			IIW			Urban1k			ShareGPT4v		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP* (Radford et al., 2021)	46.0	67.8	75.5	91.3	98.5	99.3	62.1	82.7	89.2	82.2	96.2	98.3
LiT (Fan et al., 2023)	40.9	-	-	82.7	-	-	-	-	-	80.0	-	-
ALIGN (Jia et al., 2021)	57.4	-	-	90.7	-	-	-	-	-	85.3	-	-
SigLIP* (Zhai et al., 2023)	61.3	79.0	84.3	93.3	99.5	99.7	74.8	88.8	92.8	88.9	98.2	99.2
FLAME* (Cao et al., 2024)	-	-	-	-	-	-	87.9	-	-	93.2	-	-
FLAIR (Xiao et al., 2024)	66.2	-	-	-	-	-	87.7	-	-	98.0	-	-
Long-CLIP* (Zhang et al., 2024a)	61.0	79.6	85.1	94.6	99.3	99.5	80.5	95.4	96.8	95.4	99.9	100.0
LoTLIP* (Wu et al., 2024)	62.5	82.9	88.1	94.0	99.0	100.0	85.9	97.1	98.4	96.1	99.5	99.8
LoTLIP* + VISA (Ours)	74.6	88.7	91.1	98.2	100.0	100.0	94.6	99.4	99.8	97.8	99.8	99.9
Δ	+12.1	+5.8	+3.0	+4.2	+1.0	+0.0	+8.7	+2.3	+1.4	+1.7	+0.3	+0.1

24.3% to 28.3% on LSMDC. Notably, it achieves a significant R@1 gain on DiDeMo, increasing from 45.9% to 54.8%, a remarkable improvement of 8.9%. Similarly, VISA benefits the single-stream InternVideo2-G significantly, with R@1 improvements from 52.0% to 54.4% on MSR-VTT, 32.0% to 35.0% on LSMDC, and 61.2% to 63.7% on DiDeMo. These results demonstrate that VISA could boost retrieval across various architectures, showcasing its versatility and robustness.

4.2.3. LONG-CONTEXT RETRIEVAL

The ability to effectively process and retrieve long-context texts plays a critical role in meeting fine-grained and multi-granularity retrieval demands. To evaluate our approach, we utilize four datasets: DCI (Urbanek et al., 2024), IIW (Garg et al., 2024), Urban-1k (Zhang et al., 2024a), and ShareGPT4v (Chen et al., 2023a). DCI and Urban-1k both include urban scenes, with DCI featuring human-annotated descriptions averaging over 170 tokens, and Urban-1k highlighting object types and spatial relationships. IIW provides multi-granularity annotations capturing spatial arrangements and nuanced attributes. ShareGPT4v emphasizes detailed visual information.

We use LoTLIP (Wu et al., 2024) as the base text-to-image model. As shown in Table 3, our method achieves significant improvements across all datasets, with consistent gains across all metrics. Specifically, our approach boosts R@1 on DCI from 62.5% to 74.6%, a remarkable gain of 12.1%, demonstrating its capability to handle complex fine-grained details. Similarly, R@1 increases by 8.7% on Urban1k, 4.2% on IIW, and 1.7% on ShareGPT4v, further highlighting the robustness and generalizability of our method across diverse datasets.

Notably, the substantial improvements on DCI and Urban1k emphasize VISA’s ability to process dense and multi-granularity user queries. Compared to existing methods specifically designed for long-context queries (Zhang et al.,

2024a; Xiao et al., 2024; Cao et al., 2024), our method consistently outperforms them, demonstrating its effectiveness in handling long-context retrieval tasks.

4.3. Analysis Experiment

To evaluate the robustness of our method, we conduct a series of analytical experiments.

General Description. Table 4(a) presents the results of our method using different LMMs to generate long textual descriptions. Here the results are obtained without applying the question-answering refinement. As shown, small models such as LLaVA-7B and Qwen2VL-2B already achieve significant improvements. Additionally, as the model size increases, the quality of generated captions improves accordingly, leading to better retrieval performance.

Number of Questions. We analyze the impact of varying the number of questions on retrieval performance. As shown in Table 4(b), increasing the number of questions generally improves performance, with 3–5 questions providing an optimal balance between effectiveness and computational cost. However, using more than five questions leads to a performance drop on MSR-VTT, likely because the user queries in this dataset are relatively short, resulting in redundant questions that repetitively asks the common concepts across different candidates.

Answer Generator. We evaluate different answer generators in Table 4(c). The results demonstrate that large-scale answer generators significantly enhance performance. Interestingly, the Qwen 7B model achieves performance comparable to the much larger LLaVA-34B model. Based on our observations of the generated answers, we attribute Qwen’s superior capabilities to its excellence in both answering questions and rejecting irrelevant ones (*i.e.*, response with “Uncertain” as discussed in Sec. 3.2.2).

Text Retriever. We evaluate different text retriever in Table 4(d). Our findings indicate that while single-stream

Table 4. Analysis experiments on different modules of VISA framework. We report R@1 and R@5 on Flickr30K (based on EVA-CLIP), MSR-VTT (based on InternVideo2-G) and Urban1k (based on LoTLIP). Default setting are marked in gray .

(a) **General Description (GD)**. Higher quality descriptions lead to better retrieval performance.

GD	Flickr30K		MSR-VTT		Urban1k	
	R@1	R@5	R@1	R@5	R@1	R@5
w/o GD	83.1	95.8	52.0	74.6	85.9	97.1
LLaVA-7B	84.8	96.8	52.4	73.6	93.3	99.4
LLaVA-34B	85.6	96.8	53.3	75.0	94.2	99.2
Qwen2VL-2B	84.5	96.6	52.5	73.9	95.1	99.5
Qwen2VL-7B	85.1	96.9	53.5	74.5	95.6	99.6
Qwen2VL-72B	85.4	96.8	53.6	74.9	95.9	99.7

(b) **Number of Questions (NQ)**. A medium number of questions balances effectiveness and computational cost.

NQ	Flickr30K		MSR-VTT		Urban1k	
	R@1	R@5	R@1	R@5	R@1	R@5
0	85.6	96.8	53.3	75.0	94.2	99.2
1	85.7	97.0	53.9	74.3	94.3	99.1
3	86.1	97.3	54.4	75.3	94.6	99.4
5	86.2	97.1	54.6	75.8	94.6	99.6
7	86.1	97.1	53.8	75.2	94.6	99.4
9	86.2	97.2	53.9	75.3	94.9	99.4

(c) **Answer Generator (AG)**. Stronger answer generators lead to improved retrieval performance.

AG	Flickr30K		MSR-VTT		Urban1k	
	R@1	R@5	R@1	R@5	R@1	R@5
w/o AG	85.6	96.8	53.3	75.0	94.2	99.2
LLaVA-7B	83.4	96.4	52.6	73.7	92.1	99.2
LLaVA-34B	86.4	97.2	54.2	75.4	93.9	99.3
Qwen2VL-2B	85.5	96.7	53.2	75.0	94.2	99.3
Qwen2VL-7B	86.1	97.3	54.4	75.3	94.6	99.4

(d) **Text Retriever (TR)**. Whether single-, dual-stream, or lightweight models can effectively enhance retrieval performance.

TR	type	Flickr30K		MSR-VTT		Urban1k	
		R@1	R@5	R@1	R@5	R@1	R@5
w/o TR	-	83.1	95.8	52.0	74.6	85.9	97.1
stella-435M	dual	85.2	97.2	53.4	75.0	93.3	98.9
gte-Qwen2-1.5B	dual	85.3	97.0	53.4	75.3	93.0	99.0
bge-en-icl-7B	dual	85.9	97.2	53.6	75.4	94.1	99.2
reranker-568M	single	85.5	97.4	53.5	74.0	91.0	98.7
layerwise-2.7B	single	86.2	97.4	54.2	74.8	94.1	99.7
gemma2-9B	single	86.1	97.3	54.4	75.3	94.6	99.4

Table 5. Ablation study. Here, “GD”, “QA” and “CoT” refer to general description, question-answering, and chain-of-thought, respectively. Default setting are marked in gray .

method	Flickr30K		MSR-VTT		Urban1k	
	R@1	R@5	R@1	R@5	R@1	R@5
VISA (Ours)	86.1	97.3	54.4	75.3	94.6	99.4
w/o GD	84.4	96.9	53.2	74.1	90.3	98.9
w/o QA	85.6	96.8	53.3	75.0	94.2	99.2
w/o GD&QA	83.1	95.8	52.0	74.6	85.9	97.1
w/o CoT	85.6	97.1	54.0	73.0	94.2	99.4

models exhibit explicit sentence fusion and achieve better performance than dual-stream models, dual-stream models still demonstrate substantial performance improvements. Moreover, performance improvements are consistently observed across models of varying sizes, including both large-scale and lightweight retrievers.

4.4. Ablation Study

In this section, we conduct ablation studies to evaluate the impact of different modules in VISA. As shown in Table 5, omitting the general descriptions (w/o GD) results in a significant performance drop, highlighting the importance of

providing a comprehensive visual abstraction for capturing the image’s semantics. Similarly, removing the question-answering process (w/o QA) also degrades performance, although to a lesser extent. Notably, eliminating both components together (w/o GD&QA) results in the largest decline, highlighting their complementary roles in the retrieval process. Additionally, we observe a consistent performance decline when the key phrases in chain-of-thought is removed (w/o CoT), confirming the critical role of deep reasoning in generating questions. These findings strongly support our hypothesis that granularly consistent visual abstraction significantly enhances visual retrieval performance.

5. Conclusion

In this paper, we explore a new perspective for text-to-visual retrieval by transforming visual content into textual representations and conducting retrieval within a unified text space. To achieve this, we propose VISA, a plug-and-play approach that seamlessly integrates with existing VLMs, leveraging off-the-shelf large models to generate concise and query-aligned descriptions, thereby enhancing retrieval performance at test time. By reformulating traditional cross-modal tasks into a pure text space, visual abstraction has the potential to benefit other multi-modal applications, such as video grounding and composed image retrieval, provid-

ing a foundational framework for improving cross-modal alignment.

Impact Statement

VISA inherits potential challenges associated with LMM-generated descriptions, such as biases and hallucinations. Since descriptions are automatically generated, the retrieval process might be influenced by inaccurate or biased representations, leading to unintended consequences, particularly in sensitive domains like news media analysis or legal decision-making. Additionally, transforming images into textual descriptions could raise privacy concerns, especially when applied to personal or surveillance-related datasets.

Acknowledgments

This work was supported in part by NSFC under Grant 62176171, 624B2099, 62472295, U24B20174; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303, CJ202403; in part by Sichuan Science and Technology Planning Project under Grant 24NSFTD0130; and in part by Baidu Scholarship.

References

- Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.
- Cao, A., Wei, X., and Ma, Z. Flame: Frozen large language models enable data-efficient language-image pre-training, 2024. URL <https://arxiv.org/abs/2411.11927>.
- Chen, D. and Dolan, W. Collecting highly parallel data for paraphrase evaluation. *Meeting of the Association for Computational Linguistics*, Jun 2011.
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024a.
- Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., and Lin, D. Sharegpt4v: Improving large multi-modal models with better captions, 2023a. URL <https://arxiv.org/abs/2311.12793>.
- Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M., Zhu, X., and Liu, J. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset, 2023b. URL <https://arxiv.org/abs/2305.18500>.
- Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.
- Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., and Dai, J. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024b. URL <https://arxiv.org/abs/2312.14238>.
- Chen, Z., Xu, C., Qi, Y., and Guo, J. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training, 2024c. URL <https://arxiv.org/abs/2407.21439>.
- Choi, J., Lee, S., Chu, J., Choi, M., and Kim, H. J. vid-tldr: Training free token merging for light-weight video transformer, 2024. URL <https://arxiv.org/abs/2403.13347>.
- Fan, L., Krishnan, D., Isola, P., Katabi, D., and Tian, Y. Improving clip training with language rewrites, 2023. URL <https://arxiv.org/abs/2305.20088>.
- Garg, R., Burns, A., Ayan, B. K., Bitton, Y., Montgomery, C., Onoe, Y., Bunner, A., Krishna, R., Baldridge, J., and Soricut, R. Imageinwords: Unlocking hyper-detailed image descriptions, 2024.
- Hendricks, L. A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. Localizing moments in video with natural language. In *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.618. URL <http://dx.doi.org/10.1109/iccv.2017.618>.
- Huang, Z., Yang, M., Xiao, X., Hu, P., and Peng, X. Noise-robust vision-language pre-training with positive-negative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL <https://arxiv.org/abs/2102.05918>.

- Jun, X., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. *IEEE Conference Proceedings, IEEE Conference Proceedings*, Jan 2016.
- Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 201–216, 2018.
- Li, D., Li, J., Li, H., Niebles, J. C., and Hoi, S. C. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4953–4963, 2022a.
- Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.
- Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022b. URL <https://arxiv.org/abs/2201.12086>.
- Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL <https://arxiv.org/abs/2301.12597>.
- Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., and Qiao, Y. Unmasked teacher: Towards training-efficient video foundation models, 2024. URL <https://arxiv.org/abs/2303.16058>.
- Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. Oscar: Object-semantic aligned pre-training for vision-language tasks. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.
- Lin, Y., Zhang, J., Huang, Z., Liu, J., Wen, Z., and Peng, X. Multi-granularity correspondence learning from long-term noisy videos, 2024. URL <https://arxiv.org/abs/2401.16702>.
- Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Liu, R., Li, C., Ge, Y., Shan, Y., Li, T. H., and Li, G. Bt-adapter: Video conversation is feasible without video instruction tuning, 2024b. URL <https://arxiv.org/abs/2309.15785>.
- Liu, Z., Sun, W., Teney, D., and Gould, S. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *Transactions on Machine Learning Research*, 2024c. ISSN 2835-8856. URL <https://openreview.net/forum?id=fJAwemcVpL>.
- Lu, Y., Yang, M., Peng, D., Hu, P., Lin, Y., and Peng, X. Llava-reid: Selective multi-image questioner for interactive person re-identification, 2025. URL <https://arxiv.org/abs/2504.10174>.
- Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval, 2021. URL <https://arxiv.org/abs/2104.08860>.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015. URL <http://arxiv.org/abs/1505.04870>.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.
- Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- Rohrbach, A., Rohrbach, M., Tandon, N., and Schiele, B. A dataset for movie description. *arXiv: Computer Vision and Pattern Recognition, arXiv: Computer Vision and Pattern Recognition*, Jan 2015.
- Shi, B., Zhao, P., Wang, Z., Zhang, Y., Wang, Y., Li, J., Dai, W., Zou, J., Xiong, H., Tian, Q., and Zhang, X. Umg-clip: A unified multi-granularity vision generalist for open-world understanding, 2024. URL <https://arxiv.org/abs/2401.06397>.
- Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., and Wang, X. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2023.
- Tang, C., Xie, L., Zhang, X., Hu, X., and Tian, Q. Visual recognition by request. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15265–15274, 2023.
- Team, Q. Qwen2.5: A party of foundation models, September 2024. URL <https://qwenlm.github.io/blog/qwen2.5/>.

- Urbanek, J., Bordes, F., Astolfi, P., Williamson, M., Sharma, V., and Romero-Soriano, A. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26700–26709, June 2024.
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.
- Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Yu, J., Wang, Y., Wang, L., and Qiao, Y. Internvideo: General video foundation models via generative and discriminative learning, 2022. URL <https://arxiv.org/abs/2212.03191>.
- Wang, Y., Li, K., Li, X., Yu, J., He, Y., Wang, C., Chen, G., Pei, B., Yan, Z., Zheng, R., Xu, J., Wang, Z., Shi, Y., Jiang, T., Li, S., Zhang, H., Huang, Y., Qiao, Y., Wang, Y., and Wang, L. Internvideo2: Scaling foundation models for multimodal video understanding, 2024b. URL <https://arxiv.org/abs/2403.15377>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.
- Wu, W., Zheng, K., Ma, S., Lu, F., Guo, Y., Zhang, Y., Chen, W., Guo, Q., Shen, Y., and Zha, Z.-J. Lotlip: Improving language-image pre-training for long text understanding, 2024. URL <https://arxiv.org/abs/2410.05249>.
- Xiao, R., Kim, S., Georgescu, M.-I., Akata, Z., and Alaniz, S. Flair: Vlm with fine-grained language-informed image representations, 2024. URL <https://arxiv.org/abs/2412.03561>.
- Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., and Zhou, J. mplug-2: A modularized multi-modal foundation model across text, image and video, 2023. URL <https://arxiv.org/abs/2302.00402>.
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models, 2022. URL <https://arxiv.org/abs/2205.01917>.
- Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training, 2023.
- Zhang, B., Zhang, P., Dong, X., Zang, Y., and Wang, J. Long-clip: Unlocking the long-text capability of clip, 2024a. URL <https://arxiv.org/abs/2403.15378>.
- Zhang, D., Li, J., Zeng, Z., and Wang, F. Jasper and stella: distillation of sota embedding models, 2025. URL <https://arxiv.org/abs/2412.19048>.
- Zhang, Y., Li, B., Liu, h., Lee, Y. j., Gui, L., Fu, D., Feng, J., Liu, Z., and Li, C. Llava-next: A strong zero-shot video understanding model, April 2024b. URL <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.
- Zheng, K., Zhang, Y., Wu, W., Lu, F., Ma, S., Jin, X., Chen, W., and Shen, Y. Dreamlip: Language-image pre-training with long captions, 2024. URL <https://arxiv.org/abs/2403.17007>.
- Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, W., Li, Z., Liu, W., and Yuan, L. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2024. URL <https://arxiv.org/abs/2310.01852>.

Appendix Overview

This supplementary document is organized as follows:

- Sec. A shows the performance of VISA based on BLIP-2.
- Sec. B shows the FLOPs and latency of VISA.
- Sec. C places the prompts for large models used in VISA.
- Sec. D shows the details of the evaluation datasets.
- Sec. E shows the details of the evaluation models
- Sec. F presents visualization results.

A. VISA based on BLIP-2

We compare our method with BLIP-2, referencing Table 5 from the original BLIP-2 paper (Li et al., 2023). As shown in Table 6, VISA consistently improves performance when applied on top of BLIP-2. This demonstrates the compatibility and effectiveness of our method even when applied to finetuned VLMs.

Table 6. Results of VISA based on BLIP-2.

Method	COCO Fine-tuned			Flickr30K Zero-shot		
	R@1	R@5	R@10	R@1	R@5	R@10
ALBEF (Li et al., 2021)	60.7	84.3	90.5	82.8	96.3	98.1
BLIP (Li et al., 2022b)	65.1	86.3	91.8	86.7	97.3	98.7
BLIP-2 ViT-L (Li et al., 2023)	66.3	86.5	91.8	88.6	97.6	98.9
BLIP-2 ViT-g (Li et al., 2023)	68.3	87.7	92.6	89.7	98.1	98.9
BLIP-2 ViT-g + VISA (Ours)	68.9	88.0	92.9	90.2	98.4	99.2
Δ	+0.6	+0.3	+0.3	+0.5	+0.3	+0.3

B. Efficiency of VISA

In this section, we analyze the efficiency of VISA on the Flickr30K dataset. When deployed in practice, the VISA-based retrieval process can be divided into two stages:

- **Offline stage** precomputes visual features (via VLM) and generates general descriptions for all gallery candidates.
- **Online stage** performs the text encoding (via VLM), QA-based description refinement and text-level re-ranking during inference.

Table 7. FLOPs and latency of VISA.

Type	Model	GFLOPs	Latency (second / dataset)
offline	VLM (SigLIP / EVA-CLIP / BLIP-2)	335 / 4560 / -	4 / 18 / 23
	General Description (LLaVA-v1.6-34B)	138000	437.5
Type	Model	GFLOPs	Latency (second / per query)
online	VLM (SigLIP / EVA-CLIP / BLIP-2)	26 / 9 / 1460	0.0002 / 0.0029 / 0.0421
	Question Generator (Qwen2.5-32B)	26870	0.02
	Answer Generator (Qwen2VL-7B)	15450	1.00
	Text Retriever (gemma-9B / stella-435M)	4160 / 1.6	0.13 / 0.0005

We report the FLOPs and latency of VISA in Table 7. The offline stage of VISA includes generating general descriptions for gallery candidates using a large multi-modal model (e.g., LLaVA-v1.6-34B), with a total computational cost of 138K

GFLOPs and 437.5 seconds on the Flickr30K dataset. This stage is executed only once and does not affect online inference. For fair comparison, we report only the dual-stream latency of BLIP-2 in the offline stage, given its hybrid-stream architecture.

The online stage includes query encoding, QA-based refinement, and text-level reranking with an average latency of approximately 1 second per query. This stage supports multiple configurations that enable trade-offs between performance and efficiency: i) QA-based refinement can be treated as an optional step. Using only general descriptions still yields strong performance, as shown in Table 4(a); ii) the default text retriever can be replaced with a lightweight model such as stella-435M (Zhang et al., 2025), reducing latency to 0.0005 seconds per query while still providing substantial performance gains (as demonstrated in Table 4(d)). These configurable options allow VISA to adapt flexibly to different deployment scenarios, balancing retrieval quality with computational efficiency based on application demands.

C. Prompts of Large Models

The prompts utilized in VISA are provided in Tables 8 to 10.

Table 8. The prompt of general description generation.

```
prompt of image data=  
f" ""  
Please generate detailed descriptions of the given image.  
""  
  
prompt of video data=  
f" ""  
Please provide a detailed description of the video, focusing on the main subjects, their actions, the background scenes.  
""
```

Table 9. The prompt of answer generator.

```
prompt = f" ""  
You need to answer the following question about the given image/video in English: {question}.  
  
Avoid giving answers that are just 'Yes', 'No', or a single word. Each response should provide a complete sentence.  
  
If you cannot determine the answer or there are no objects that are asked by the question, just answer a single word 'Uncertain'.  
  
Note that if the given video is unrelated or only loosely related to these phrases:{key phrases}, just response a single word  
'Uncertain'.  
  
""
```

Table 10. The prompt for the question generator, including an example to specify the output format which is used for all datasets.

```
prompt = f" " "
You are observing a image/video. A single sentence is provided to describe what you are seeing in the image/video.
Based on this sentence, perform the following tasks:

1. Extract Key Phrases: Identify and list key phrases from the sentence that represent the main elements of the scene. Focus
on capturing: object types, object attributes, object actions, object locations, interactions between objects or other dynamic
elements.

2. Generate Three Questions: Based on the extracted key phrases, create three natural language questions that a visual AI
assistant might ask about the image/video. The questions should:
    2.1 Be specific and focused on the visual details described in the sentence.
    2.2 Ensure that the answer to each question can be determined confidently:
        - Either the content is present and can be answered confidently from the image/video.
        - Or the content is absent, and its non-presence can be confidently verified.
    2.3 Only ask questions about quantities (e.g., "How many...") if the description explicitly mentions a number or quantity
(e.g., "three," "several").
    2.4 Avoid directly repeating or closely paraphrasing most of the sentence.
    2.5 Ensure the answers to the questions can be provided using the extracted key phrases.

Output Requirements:
Key phrases: Provide a numbered list of key phrases extracted from the input sentence.
Questions: Provide a numbered list of three relevant questions based on the key phrases.

Example:
Input:
A green and yellow tennis sweater is hanging on the back of the sofa.

Output:
key phrases:
1. green and yellow
2. tennis sweater
3. a green and yellow tennis sweater
4. hanging on
5. the back of
6. sofa
7. the back of the sofa
8. hanging on the back of the sofa
questions:
1. What colors are present on the tennis sweater?
2. Where is the tennis sweater located?
3. What type of object is the sweater hanging on?

Now, perform these tasks for the following input:
{query}
" " "
```

D. Evaluation Datasets

In this section, we place more details about the evaluation datasets involving text-to-image retrieval (COCO and Flickr30k), text-to-video (MST-VTT, LSMDC, DiDeMo, and MSVD), and long-context retrieval (DCI, IIW, Urban1k and ShareGPT4v).

- COCO (Chen et al., 2015) includes images featuring a wide variety of objects such as people, animals, vehicles, and household items. The test set of COCO contains 5,000 images with 25,010 manual annotations.
- Flickr30k (Plummer et al., 2015) focuses on real-world scenes with complex interactions and relationships. Each image is annotated with five descriptive sentences, capturing detailed information about objects, actions, and their interactions. The test set of Flickr30k contains 1,000 images with 5,000 manual annotations.
- MSR-VTT (Jun et al., 2016) spans a broad spectrum of everyday scenarios and events, including sports, music performances, cooking demonstrations, gaming sessions, and various social and cultural activities. The test set of MSR-VTT contains 1000 videos.
- DiDeMo (Hendricks et al., 2017) encompasses a wide variety of real-world scenarios, including cooking, sports, travel, and general daily activities. The test set of DiDeMo contains 1,003 videos, each trimmed to a maximum of 30 seconds and paired with several descriptive sentences that highlight distinct moments. We perform video-paragraph retrieval by merging all sentence descriptions of a video into a single query.
- LSMDC (Rohrbach et al., 2015) contains short video clips from movies, spanning various genres such as drama, comedy, and action. The test set of LSMDC contains 1,000 clips and each clip is annotated with one concise English descriptions focusing on key events, dialogues, and character interactions.
- MSVD (Chen & Dolan, 2011) covers short YouTube videos capturing diverse everyday activities, objects, and events. The test set of MSVD contains 670 videos, each annotated with around 40 English descriptions that highlight various actions and contexts. We perform video-paragraph retrieval by merging all sentence descriptions of a video into a single query.
- DCI (Urbanek et al., 2024) includes a diverse range of real-world scenes, covering urban landscapes, architectural structures, natural environments, and everyday objects. It consists of 7,805 natural images, each accompanied by detailed human-annotated descriptions averaging over 170 tokens per image.
- IIW (Garg et al., 2024) includes object-level annotations and image-level descriptions that capture spatial arrangements, interactions, and nuanced attributes. It contains 612 images with hyper-detailed descriptions, averaging 217.2 tokens per image.
- Urban-1k (Zhang et al., 2024a) consists of 1,000 urban images sourced from the Visual Genome dataset, paired with captions generated by GPT-4V. Each caption averages approximately 107 words and offers detailed descriptions, including object types, colors, and spatial relationships.
- ShareGPT4v (Chen et al., 2023a) emphasize a wide range of visual information, including object properties, spatial relationships, aesthetic evaluations, and contextual knowledge. It contains 1,000 images with 1,000 detailed descriptions.

E. Details of Evaluated Models

Table 11 provides references for all the models used in our paper.

Table 11. Model links. The table is categorized based on model types, including VLMs, LLMs, LMMs, and text retrievers. “HF” is the abbreviation of HuggingFace.

Type	Model Name	Official Name	Model Link
VLM	SigLIP	SigLIP	HF:timm/ViT-SO400M-14-SigLIP-384
	EVA-CLIP	EVA-CLIP-18B	HF:BAAI/EVA-CLIP-18B
	LoTLIP	LoTLIP	Github:wuw2019/LoTLIP/tree/main
	InternVideo2	InternVideo2	Github:OpenGVLab/InternVideo
LLM	Qwen2.5-32B	Qwen2.5-32B-Instruct	HF:Qwen/Qwen2.5-32B-Instruct
LMM	LLaVA-7B	llava-v1.6-mistral-7b	HF:liuhaotian/llava-v1.6-mistral-7b
	LLaVA-v1.6-34B	llava-v1.6-34b	HF:liuhaotian/llava-v1.6-34b
	LLaVA-Video-32B	LLaVA-NeXT-Video-32B-Qwen	HF:Imms-lab/LLaVA-NeXT-Video-32B-Qwen
	Qwen2-VL-2B	Qwen2-VL-2B-Instruct	HF:Qwen/Qwen2-VL-2B-Instruct
	Qwen2-VL-7B	Qwen2-VL-7B-Instruct	HF:Qwen/Qwen2-VL-7B-Instruct
	Qwen2-VL-72B	Qwen2-VL-72B-Instruct	HF:Qwen/Qwen2-VL-72B-Instruct
Text Retriever	stella-435M	stella-en-400M-v5	HF:NovaSearch/stella-en-400M-v5
	gte-Qwen2-1.5B	gte-Qwen2-1.5B-instruct	HF:Alibaba-NLP/gte-Qwen2-1.5B-instruct
	bge-en-icl-7B	bge-en-icl	HF:BAAI/bge-en-icl
	reranker-568M	bge-reranker-v2-m3	HF:BAAI/bge-reranker-v2-m3
	layerwise-2.7B	bge-reranker-v2-minicpm-layerwise	HF:BAAI/bge-reranker-v2-minicpm-layerwise
	gemma2-9B	bge-reranker-v2.5-gemma2-lightweight	HF:BAAI/bge-reranker-v2.5-gemma2-lightweight

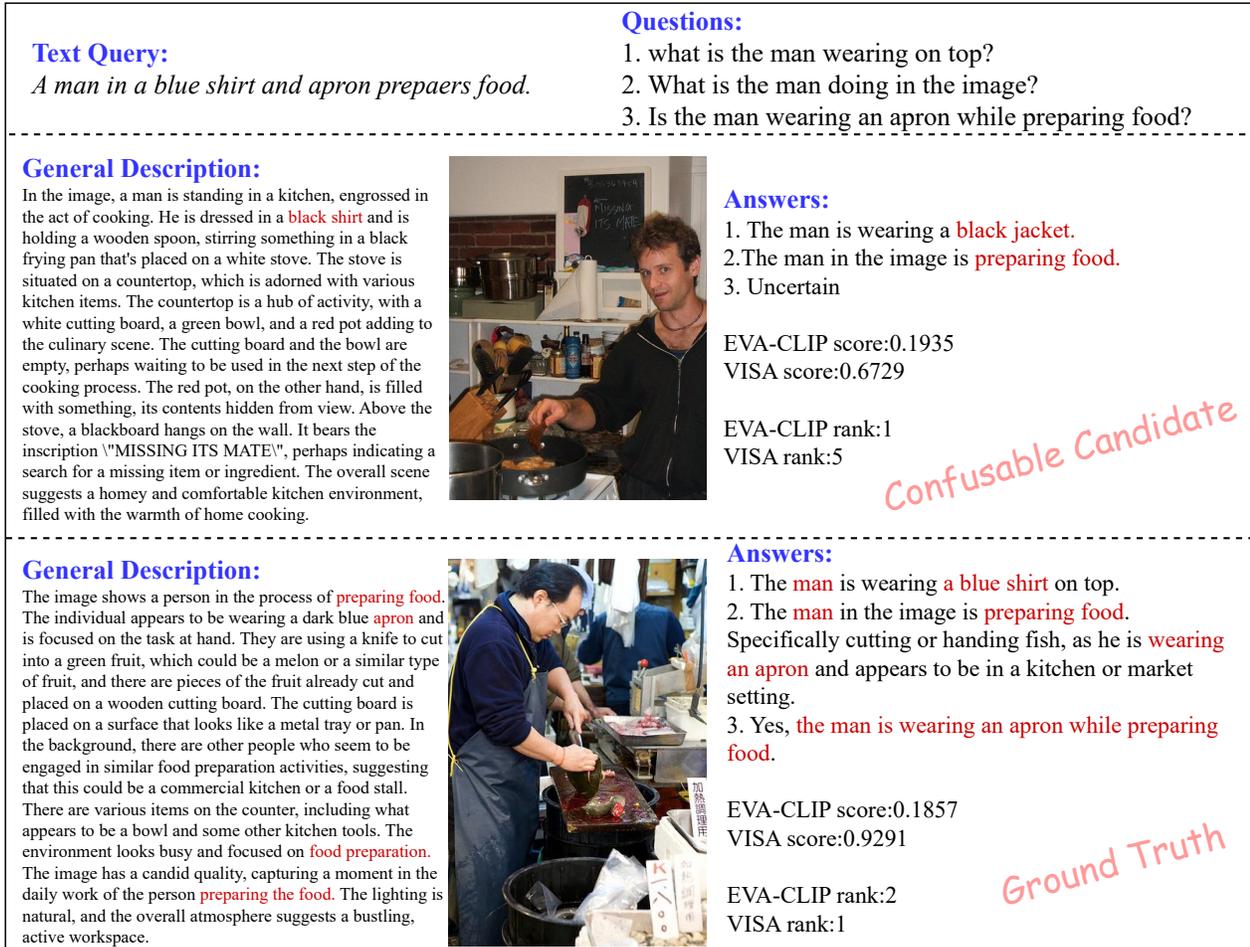
F. Visualization Results

The visualization results of our method are presented in Fig. 3-5. These examples highlight how VISA enhances retrieval by refining visual abstraction through question-answering.

In Fig. 3, our method generates questions related to clothing and an apron, refining fine-grained details. Specifically, in the top case, the general description incorrectly identifies a black shirt, whereas the person is actually wearing a black jacket. The question-answering process corrects this mistake, improving the accuracy of the visual abstraction. In the bottom case, our method correctly identifies a blue shirt and an apron, ensuring better alignment with the user query.

In Fig. 4, the QA process successfully incorporates the countdown timer, aligning the retrieved content more closely with the user query.

In Fig. 5, the question generator identifies that “windows” in the query refers to the operating system and formulates relevant questions accordingly. Although the query includes “man” in the first video sample, it does not contain the “Windows” concept. InternVideo fails to capture this, whereas VISA successfully recognizes and incorporates the missing context, leading to more accurate retrieval. This improvement is attributed to the contextual understanding capabilities of the questioner LLM, which enables more precise alignment between the query and the retrieved content.



Confusable Candidate

Ground Truth

Figure 3. A visualization example from Flickr30K with two candidate images.

Text Query:

This image captures an urban street scene dominated by a traffic signal pole at the center. Affixed to the pole are various signs: a digital crosswalk signal indicating a red hand with a countdown of "3" seconds, a street name sign for "2nd Ave" with a green background and white lettering, and a "ONE WAY" sign with a directional arrow pointing to the right. In the background, there's a blue and yellow sign that advertises "Breakfast All Day" above a restaurant entrance. In the foreground, pedestrians appear to be walking along the sidewalk, and a set of traffic lights is partially visible at the top right corner. The location looks like a bustling downtown area with ample sunshine.

Questions:

1. What is the countdown number displayed on the digital crosswalk signal?
2. What is the name of the street indicated by the sign with a green background and white lettering?
3. What type of sign is visible above the restaurant entrance, and what does it advertise?

General Description:

The image captures a vibrant city scene at an intersection. Dominating the frame is a traffic light, its red light glowing brightly, signaling vehicles to halt. The traffic light is affixed to a sturdy metal pole, which also supports a street sign. The sign, painted in a striking combination of green and white, reads "F.D.R. Drive", indicating the name of the street. Just below the traffic light, a blue and white street sign announces "East 42nd St", marking the location of the intersection. The signs are clear and easily visible, guiding both pedestrians and drivers. In the background, a building stands tall. Its facade is adorned with a sign that reads "Innovate", perhaps indicating the nature of the business or the spirit of the neighborhood. The building's presence adds a sense of depth and scale to the image. The image is a snapshot of urban life, capturing the rhythm of the city in its everyday hustle and bustle.



Answers:

1. Uncertain
2. The name of the street indicated by the sign with a green background and white lettering is **F.D.R. Drive**.
3. Uncertain

LoTLIP score:0.2713

VISA score:0.5778

LoTLIP rank:1

VISA rank:2

Confusable Candidate

General Description:

The image captures a vibrant city scene at the corner of **2nd Avenue**. The street is bustling with life, with people walking on the sidewalk and cars driving by. The buildings lining the street are adorned with various signs, including a **yellow and blue** sign for a restaurant named "**Breakfast All Day**". A green street sign reading "**2nd Avenue**" is prominently displayed, indicating the location. A black and white "**One Way**" sign is also visible, guiding the flow of traffic. The sky above is a clear blue, suggesting a **sunny day**. The image is a snapshot of urban life, capturing the hustle and bustle of city life, the variety of businesses, and the orderly flow of traffic. It's a moment frozen in time, a slice of life in a city that's always on the move.



Answers:

1. The countdown number displayed on the digital crosswalk signal is **3**.
2. The name of the street indicated by the sign with a green background and white lettering is **2nd Avenue**.
3. The visible sign above the restaurant entrance advertises that the **breakfast is available all day**.

LoTLIP score:0.2530

VISA score:0.9120

LoTLIP rank:2

VISA rank:1

Ground Truth

Figure 4. A visualization example from Urban1k with two candidate images.

Text Query:

a man describing how to do something in windows.

Questions:

1. What is the man demonstrating in the video?
2. In which software or operating system is the man giving instructions?
3. Is the man providing verbal instructions in the video?



General Description:

The video begins with a person standing in a room, wearing a dark blue V-neck shirt and holding a blue backpack. The background features a map of the United States on the wall, a clock showing 10:10, and several hats hanging on the left side. A sign on the wall reads 'believe you can.' The person appears to be explaining something, using hand gestures. The scene transitions to a close-up of a table filled with various items, including neatly folded clothes, shoes, and other travel essentials. The person is seen packing these items into the blue backpack. The video then returns to the initial scene with the person continuing to explain, still holding the backpack. The final frame shows a collage of various images with the text 'in the cut' and 'urban explorers' overlaid on it, indicating a possible theme or title related to urban exploration. The video continues with the same person standing in the same room, still wearing the dark blue V-neck shirt and holding the blue backpack. The background remains consistent with the map of the United States, the clock showing 10:10, and the hats hanging on the left side. The sign that reads 'believe you can' is also visible. The person continues to explain something, using hand gestures, similar to the previous scenes. The scene maintains a focus on the person and their actions, with no significant changes in the background or additional elements introduced.

Answers:

1. Uncertain
2. Uncertain
3. Yes, the man is providing verbal instructions in the video.

InternVideo2-G score:0.0206

VISA score:0.8534

InternVideo2-G rank:1
VISA rank:2

Confusable Candidate



General Description:

The video begins with a view of a computer desktop screen featuring the Windows 7 logo and various icons on the left side. The Remote Desktop Connection window is open, displaying the 'Remote Desktop Connection' dialog box with fields for entering the name of the computer to connect to. The user types 'contoso' into the 'Computer' field and clicks the 'Connect' button. A new dialog box appears, prompting the user to enter credentials. The user enters 'contoso\Administrator' in the 'User name' field and clicks the 'Enter' button. The user then enters a password in the 'Password' field and clicks the 'OK' button. The user confirms the password by clicking the 'Yes' button in the 'Remote Desktop Connection' dialog box. The scene transitions to a close-up view of the 'Remote Desktop Connection' dialog box, where the user has entered 'contoso\Administrator' in the 'User name' field and is about to click the 'OK' button. The background shows the Windows 7 desktop with the Windows logo and various icons. The user clicks the 'OK' button, and the 'Remote Desktop Connection' window appears, showing a progress bar that fills up from 0% to 100%. The user clicks the 'Show Options' button, revealing additional options such as 'Connect to another remote desktop' and 'Save As'. The user clicks the 'Connect' button, and the progress bar starts filling up again, indicating the connection process. The video continues with a close-up view of the 'Remote Desktop Connection' window, showing the progress bar filling up from 0% to 100%. The background remains the same, with the Windows 7 desktop and the Windows logo visible. The user clicks the 'Show Options' button, revealing additional options such as 'Connect to another remote desktop' and 'Save As'. The user clicks the 'Connect' button, and the progress bar starts filling up again, indicating the connection process. The video concludes with the progress bar still filling up, suggesting that the connection is being established.

Answers:

1. The man in the video is demonstrating how to use the Remote Desktop Connection feature in Windows.
2. The man is giving instructions in the Windows operating system.
3. Yes, the man is providing verbal instructions in the video.

InternVideo2-G score:0.0200

VISA score:0.9693

InternVideo2-G rank:2
VISA rank:1

Ground Truth

Figure 5. A visualization example from MSR-VTT with two candidate videos.

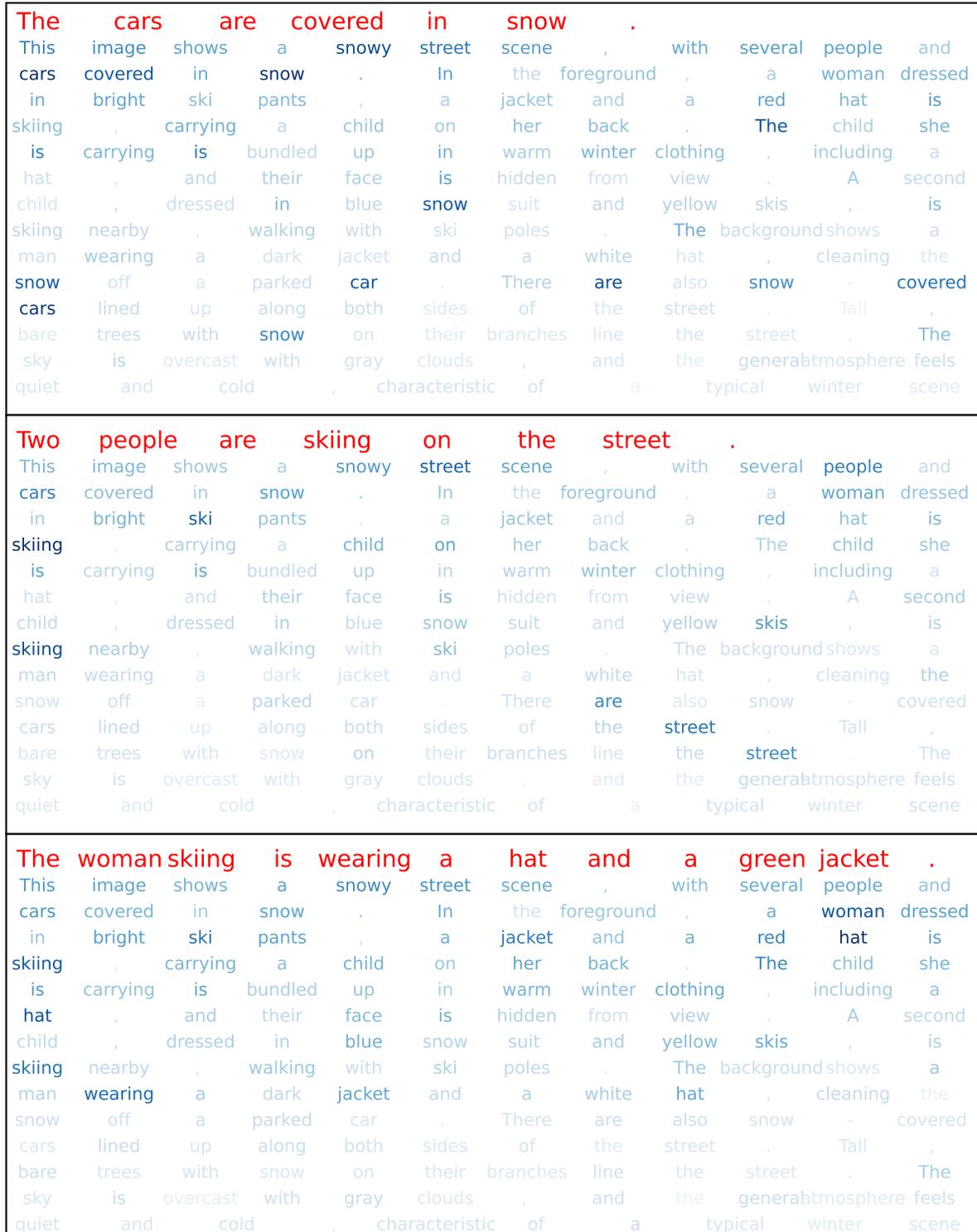


Figure 6. Visualization of the attention heatmaps of Gemma2 (Chen et al., 2024a) given the text queries and descriptions from Fig. 1.