Unified Generative and Discriminative Training for Multi-modal Large Language Models

Wei Chow[1]   **Juncheng Li**[1,†]   **Qifan Yu**[1]   **Kaihang Pan**[1]   **Hao Fei**[2]

Zhiqi Ge[1]   **Shuai Yang**[1]   **Siliang Tang**[1,†]   **Hanwang Zhang**[3]   **Qianru Sun**[4]

[1]Zhejiang University   [2]National University of Singapore

[3]Nanyang Technological University   [4]Singapore Management University

{xieqiao, junchengli, yuqifan, kaihangpan}@zju.edu.cn

{zhiqige, syang, siliang}@zju.edu.cn

haofei37@nus.edu.sg, hanwangzhang@ntu.edu.sg, qianrusun@smu.edu.sg

Abstract

In recent times, Vision-Language Models (VLMs) have been trained under two predominant paradigms. Generative training has enabled Multimodal Large Language Models (MLLMs) to tackle various complex tasks, yet issues such as hallucinations and weak object discrimination persist. Discriminative training, exemplified by models like CLIP, excels in zero-shot image-text classification and retrieval, yet struggles with complex scenarios requiring fine-grained semantic differentiation. This paper addresses these challenges by proposing a unified approach that integrates the strengths of both paradigms. Considering interleaved image-text sequences as the general format of input samples, we introduce a structure-induced training strategy that imposes semantic relationships between input samples and the MLLM's hidden state. This approach enhances the MLLM's ability to capture global semantics and distinguish fine-grained semantics. By leveraging dynamic sequence alignment within the Dynamic Time Warping framework and integrating a novel kernel for fine-grained semantic differentiation, our method effectively balances generative and discriminative tasks. Extensive experiments demonstrate the effectiveness of our approach, achieving state-of-the-art results in multiple generative tasks, especially those requiring cognitive and discrimination abilities. Additionally, our method surpasses discriminative benchmarks in interleaved and fine-grained retrieval tasks. By employing a retrieval-augmented generation strategy, our approach further enhances performance in some generative tasks within one model, offering a promising direction for future research in vision-language modeling. The project repository is here.

1   Introduction

In recent times, Vision-Language Models (VLMs) have been trained under two predominant paradigms: generative training and discriminative training. **Generative Training** has achieved remarkable success in enabling Multimodal Large Language Models (MLLMs) [1, 55, 86] to develop a wide range of powerful capabilities that can handle various complex tasks (*e.g.,* open-world visual question-answering, image caption generation, etc.) within a single model. However, challenges such as hallucinations and weak image object discrimination abilities [7, 89] persist. **Discriminative Training**, exemplified by CLIP [73], exhibits remarkable representation capabilities for zero-shot image-text classification and retrieval. Nonetheless, it encounters difficulties in processing complex scenarios (*i.e.,* , retrieving multi-modal documents with interleaved images and texts) [53, 54] and exhibits a limited ability to discern detailed semantic differences [79, 85].

[†] Corresponding Author.

The disparity between these two paradigms has sparked recent studies aimed at imparting discriminative ability to generative pre-trained MLLMs. However, certain aspects of performance still pose limitations (*e.g.*, singular discriminative tasks [89], weak discriminative task performance [40], weak generalization [59], etc.), while others entail compromising the model's original generative capabilities [8].

Overall, the reason generative paradigms struggle with performing discriminative tasks like retrieval is due to overlooking two crucial abilities:

(i) **Comprehensively capturing the global semantics**. Recent studies have revealed that causal LLMs tend to exhibit a bias towards capturing global information from the input samples, often resulting in a tendency to overlook information located in the middle, especially for long sequences [15, 57]. As illustrated in Figure 1(a), we chose 500 samples from WebQA [10], where the task is to find and reason about the right image-text pair among five distractors to produce a yes or no answer. We conducted experiments using VILA [52], a MLLM with state-of-the-art interleaved image-text comprehension ability, alongside our model. When placing the relevant pair in different positions, the performance of MLLMs followed a 'U' shape, indicating a bias in capturing global semantic information. Consequently, MLLMs encounter difficulties in forming comprehensive representations that encompass global semantics for retrieval tasks.

(ii) **Keenly differentiating the detailed semantics**. Some research [47, 82] has found that the existing generative training framework cannot fully distinguish input semantics in certain contexts, causing MLLMs to struggle with tasks requiring fine-grained semantics [46, 98]. As depicted in Figure 1(b), we noticed that MLLMs face challenges in choosing the right description for two similar images in the MMVP-VLM benchmark [81]. This indicates that MLLMs struggle to effectively differentiate the detailed semantics of input samples, naturally leading to difficulties in forming effective queries for retrieval.
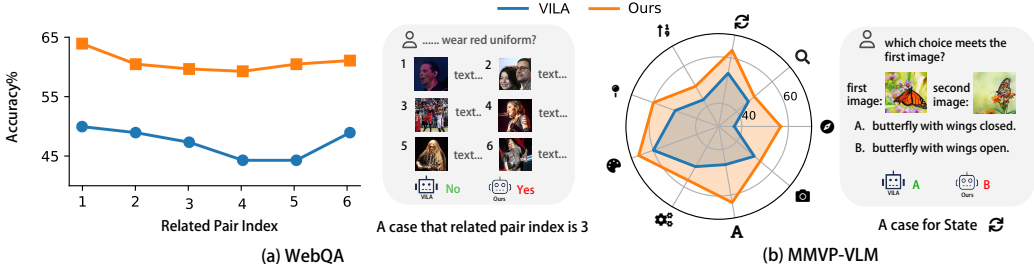

Figure 1: (a) In WebQA [10], the accuracy roughly forms a "U" shape curve when the relevant image-text pair for a question appears at different positions. While our model also shows similar trends, it tends to be more stable overall. (b) The accuracy of various types of questions in MMVP-VLM [81], it can be observed that our model's performance improves on such tasks after introducing the discriminative training. Details can be seen in Appendix E.3

In this paper, we argue that the current separated paradigms possess the potential for achieving synergistic gains. We propose **Sugar**: **S**tructure-induced approach to **u**nify **g**enerative **a**nd disc**r**iminative paradigms (shown in Figure 2), leveraging discriminative training to acquire the two abilities above while harnessing the potential of generative training in complex discriminative tasks like image-text interleaved retrieval and fine-grained retrieval. Specifically, we explicitly impose the semantic relationships between different input samples as an induced structural constraint on the hidden state of MLLMs. We consider the interleaved image-text sequence as the general format of input samples, and then formulate the relationship between any two samples as a dynamic sequence alignment problem within the Dynamic Time Warping framework [67, 33]. In this way, we can explicitly modulate the hidden states of the MLLM by leveraging the semantic relationships between interleaved input sequences, thereby encouraging the MLLM to fully **capture the global semantics** of the inputs.

To further enhance the ability to **differentiate fine-grained semantics**, we integrate a novel kernel into the Dynamic Time Warping framework. Leveraging the strengths of various discriminative pre-trained models, it performs dynamic sequence alignment for diverse embeddings tailored to specific contexts, thus addressing the inherent limitations in fully utilizing input semantics. Through this explicit structure-induced constraint, our framework enables MLLMs to capture the global semantics and fine-grained details of the input multimodal sequence more effectively, thus bridging the gap between generative and discriminative training paradigms.
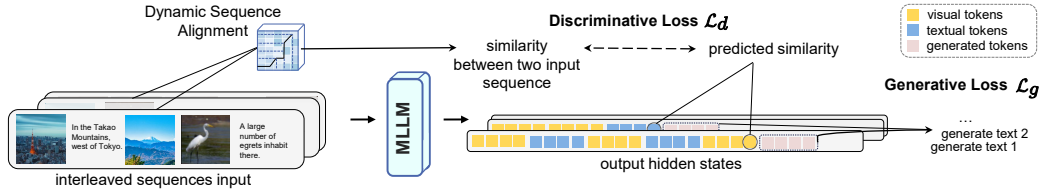
2

Figure 2: Our structure-induced generative and discriminative training joint training strategy.

Our method effectively balances both discriminative and generative tasks, demonstrating synergistic benefits. *(i)* Large-scale generative pre-trained models possess semantic-rich hidden states [41, 91, 23], which facilitate **discriminative tasks** like retrieval. Moreover, harnessing the capabilities of MLLM is crucial for complex discriminative tasks, such as interleaved image-text retrieval and fine-grained retrieval. *(ii)* By integrating discriminative tasks, the model's effectiveness in **generative tasks**, particularly within tasks requiring cognitive and discrimination abilities, is enhanced, thereby mitigating certain occurrences of hallucinations. *(iii)* We can employ Sugar to realize **retrieval-augmented generation** [2], eliminating the need for an off-the-shelf retrieval module [75], thereby amplifying the performance of various generative tasks. The usage of off-the-shelf retrieval presents a challenge wherein the retriever's performance affects the generator's final output [62]. This necessitates independent optimization of both components, posing a dilemma in selecting optimal configurations. However, our approach circumvents such optimization challenges.

Through extensive experimentation, we have demonstrated the effectiveness of our approach. For generative tasks, Sugar establishes new state-of-the-art results on the tasks for complicated multi-modal comprehension tasks (*i.e.,* DEMON [47]), fine-grained semantic distinctions (*i.e.,* VizWiz [28], MME [95]), object hallucinations detection (*i.e.,* POPE [51]) (Section 4.2 and Section 4.3). For discriminative tasks, we achieved competitive results in image-text retrieval compared, and significantly surpassed CLIP in interleaved retrieval and fine-grained retrieval (Section 4.4). Furthermore, employing the retrieval-augmented generation (RAG) strategy led to further improvements in a series of generative tasks (Section 4.5).

2 Related Work

Multi-modal Large Language Models. Flamingo [3] and BLIP-2 [49] integrate LLMs with visual encoders, showcasing impressive zero-shot capabilities by aligning visual features with language representations. Building upon the advancements of LLaVA-1.5 [55], subsequent studies [103, 19, 94, 6, 42, 72, 95, 98, 45] propose fine-tuning MLLMs with multimodal instruction tuning data [102]. Recently, there has been a surge in research [52, 80, 22, 21, 48] dedicated to enhancing the capacity of MLLMs to process interleaved image-text inputs effectively. However, these models primarily focus on generative tasks, overlooking the importance of introducing discriminative constraints. In this paper, we propose a structure-induced joint training strategy for unifying generative and discriminative tasks, further enhancing the capabilities of MLLMs, especially those requiring cognitive and discriminative abilities.

Vision-Language Pre-training. Vision-Language Pre-training primarily come in two forms: single-stream and dual-stream. In single-stream models, the embeddings for the image and text modalities are concatenated and jointly encoded [39, 50], while in dual-stream models, they are encoded by separate modality-specific encoders with optional cross-modality fusion [73, 31, 5]. These models have shown effectiveness in tasks such as classification and retrieval. However, they face challenges including difficulty in processing complex composed sequences [53, 54] and limited ability to discern detailed semantic differences [81, 79]. Recent attempts to utilize generative MLLMs for discriminative tasks have faced limitations, such as singular discriminative tasks [89], weak discriminative task performance [40], poor generalization [59], and compromised generative capabilities [8].

LLMs for Retrieval. Early models for retrieval primarily focused on word representations [16, 64, 74], with minimal generative capabilities. Some recent works have endeavored to fine-tune generative pre-trained LLMs to generate discriminative embeddings, albeit at the expense of compromising the model's original generative capabilities [44, 70, 65, 63, 24, 71]. GRIT [66] integrates generative and discriminative tasks in NLP and demonstrates mutual benefits between them. However, its training cost is prohibitively high compared to individual tasks. Moreover, due to its specialized attention mechanism, the model can only be trained from scratch.

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG)[25, 69], which harnesses the advanced inference capabilities of LLMs along with external knowledge, has the potential to significantly mitigate issues related to long-tail entities and reduce the occurrence of hallucina-

tory responses [29, 36, 101, 77, 90, 92, 97]. Recently, there have also been related studies in the multimodal domain attempting to utilize retrieval augmentation [93, 96]. These methods typically require an additional retrieval module (*e.g.,* CLIP), leading to component optimization challenges where the overall model performance is affected by the performance of the retrieval model, as well as concerns regarding the compatibility between the retrieval model and the MLLMs. Furthermore, retrieval modules like CLIP struggle to handle compositional or fine-grained scenarios, posing certain challenges for retrieval.

3   Method

As illustrated in Figure 3, we initially introduce the problem formulation and offer an overview of our structure-induced joint training strategy in Section 3.1. Subsequently, we delve into the specifics of dynamic sequence alignment algorithm in Section 3.2. Finally, we further introduce the Triple Kernel to aid in discriminating detailed semantics in Section 3.3.

3.1   Problem Formulation and Architecture Overview
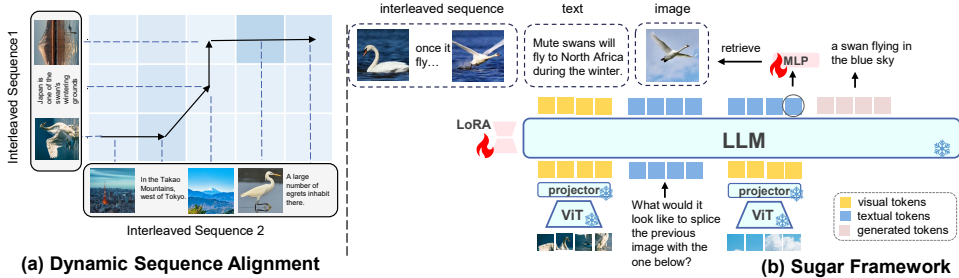
Figure 3: (a) **Dynamic Sequence Alignment**. Semantically matched slices are connected with a blue dashed line. The arrows indicate the direction of the ordered temporal alignment path. With these alignments, we can obtain the similarity between two interleaved inputs for training. (b) **Sugar Framework**. Sugar supports both multi-modal generation and retrieval simultaneously.

We view the interleaved image-text sequence as the general format for input samples, where images and textual data are alternately arranged. Typically, Multimodal Large Language Models (MLLMs) [55, 52, 11, 6] are tailored to generate text based on such input sequences, and it is conventionally optimized using self-regressive loss $\mathcal{L}_g$. A special scenario arises when the input comprises only one image and a question, prompting the MLLMs to generate an answer accordingly.

While intuitive, this optimization objective solely supervises text generation and lacks constraints on the hidden states of the entire interleaved sequence input. Additionally, the existing generative training framework struggles to fully distinguish input semantics in certain contexts, such as discerning fine-grained object details. Consequently, it fails to adequately capture the global information or distinguish detailed semantics of the input samples.

Hence, we introduce a structure-induced constraint $\mathcal{L}_d$ (see in Figure 2), which explicitly imposes the semantic relationships between different input samples as an induced structural constraint on the hidden states of MLLMs, facilitating the model in **capturing global semantics**. We conceptualize the derivation of semantic relationships between input samples as a Dynamic Sequence Alignment problem [67]. Additionally, we straightforwardly select a token in the hidden state of the MLLM to encompass all preceding input information, eliminating the need for training any specialized tokens.

To further effectively **distinguish detailed semantics**, we integrate a novel kernel into the Dynamic Time Warping framework. Leveraging the strengths of various discriminative pre-trained models. Combined with this newly proposed loss with a hyperparameter α, the training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_g + \alpha \mathcal{L}_d \tag{1}$$

3.2   Dynamic Sequence Alignment

We formulate the computation of relationships within input interleaved sequences as a dynamic sequence alignment problem, and solve it by global alignment kernel. For two interleaved image-text sequence, each consisting of n and m images/sentences in total respectively (which we'll refer to as slices later on). We encode and normalize each slice, resulting in two sequences $\mathbf{x} = (x_1, \ldots, x_n)$ and

$\mathbf{y} = (y_1, \ldots, y_m)$ all of which take values in a state space $\mathcal{X}$, that is two elements of $\mathcal{X}^\star \overset{\text{def}}{=} \bigcup_{i=1}^\infty \mathcal{X}^i$. In our setting, $\mathcal{X}$ is simply $\mathbb{R}^d$, d refers to the feature dimension. We define the global alignment kernel as follows, and it has been proved to be positive-definite under mild conditions and may prove more robust to quantify the similarity of two sequences [73, 68]:

$$K(\mathbf{x}, \mathbf{y}) = \sum_{\pi \in \mathcal{A}(\mathbf{x},\mathbf{y})} \prod_{i=1}^{|\pi|} e^{-\phi_\sigma} \in (0, 1] \tag{2}$$

Following the suggestion by [18], we let $\varphi_\sigma = \frac{1}{2\sigma^2} \varphi \left(x_{\pi_1(i)}, y_{\pi_2(i)} \right) + \log(2 - e^{-\frac{\varphi\left(x_{\pi_1(i)}, y_{\pi_2(i)} \right)}{2\sigma^2}})$, σ is standard deviation, and it can be calculated by $\sigma = \delta \sqrt{\frac{M+N}{2}}$ for x_i, y_i in $\mathbf{x}, \mathbf{y}$. δ is a fixed pre-defined hyperparameter and $\varphi \left(x_{\pi_1(i)}, y_{\pi_2(i)} \right)$ is the distance between slice $x_{\pi_1(i)}$ and $y_{\pi_2(i)}$ for an alignment (details for the definition of alignment can be seen in Appendix D.2).

Due to the causal attention mechanism, the token in hidden state of MLLM can encapsulate information from preceding tokens in the sequence. Therefore, we directly utilize the last token d_i of a sequence from the MLLM's hidden state and map it to the r_i using an MLP to represent the entire in-context sequence. During training, we obtain a set of $(r_1, r_2, \ldots, r_n)$ and their corresponding input sequence embedding set $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$. It is noteworthy that r_i and r_j ($\mathbf{x}_i$ and $\mathbf{x}_j$) may originate from the same sequence but occupy different positions, thus enabling our method to utilize samples more efficiently.

Leveraging the GAK, we can derive the similarity matrix of $(r_1, r_2, \ldots, r_n)$ and $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ distinctively, denoted as $\mathcal{M}^r, \mathcal{M}^l \in \mathbb{R}^{n \times n}$. For imposing the semantic relationships between different input samples as an induced structural constraint on the hidden state of MLLMs, we employ Mean Squared Error (MSE) loss aligned $\mathcal{M}^r$ with the label matrix $\mathcal{M}^l$. This approach eliminates the need for pre-defined label (*i.e.,* positive and negative candidates) during training, allowing seamless integration into the aforementioned training framework (for specific training templates, please refer to Appendix E.1). Thus, we have the discriminative loss $\mathcal{L}_d$:

$$\mathcal{L}_d = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \left(m_{ij}^r - m_{ij}^l \right)^2 \tag{3}$$

Additionally, when both $\mathbf{x}$ and $\mathbf{y}$ contain only one slice, the computed result of the formula is monotonically increasing with the directly calculated cosine similarity (proof can be seen in Appendix 3). Therefore, in such cases, we simplify the computation by directly using cosine similarity. If r_i and r_j comes from the same input interleaved sample, we manually set $m_{ij}^l = 1$.

3.3 Detailed Semantics Modeling

To further effectively distinguish detailed semantics, we further propose the Triple Kernel (TK), a positive definite kernel compatible with the previous framework. The TK leverages representations from diverse pre-trained discriminative models across uni-modal and cross-modal settings, harnessing their respective strengths. The definition is as below:

For two slice $a, b \in \mathbb{R}^d$, meets (i) $|a| = |b| = 2$, $d = d_1 + d_2$, $a = \text{concat}(a_1, a_2), b = \text{concat}(b_1, b_2)$, $a_1, b_1 \in \mathbb{R}^{d_1}, a_2, b_2 \in \mathbb{R}^{d_2}$ and $|a_1| = |a_2| = |b_1| = |b_2| = 1$; or (ii) $|a| = |b| = 1$. We define tripe kernel as follows:

$$\varphi(a, b) = \begin{cases} ||a_1 - b_1||^2 & |a| = |b| = 2 \text{ and } a, b \text{ in uni-modal} \\ ||a_2 - b_2||^2 & |a| = |b| = 2 \text{ and } a, b \text{ in cross-modal} \\ ||a - b||^2 & \text{else} \end{cases} \tag{4}$$

We prove triple kernel φ is a conditionally positive-definite kernel defined on $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Appendix 2), aligning with the kernel definition in [18], thereby possessing its properties.

In practice, we let the feature dimension $d = d_1 + d_2$. For images, we employ DINOv2-base [68] and CLIP ViT-L/14 [73] for encoding, then concatenate the embeddings after normalization. For sentences, we utilize BGE-base [87] and CLIP ViT-L/14, keeping the dimension unchanged. By utilizing the Triple Kernel, we can fully leverage the strengths of these three models, effectively distinguishing detailed semantics.

5

Figure 4: Selected examples for various image-text tasks. The `pink background` indicates retrieval results, while the `blue background` indicates generated results. More examples are provided in the Appendix F.2.

Method	LLM	Res.	VQA^{v2}	GQA	VizWiz	SQA^I	VQA^T	POPE	MME^P	MME^C	MMB	LLaVA^{Wd}	MM-Vet
BLIP-2 [49]	Vicuna-13B	224	41.0	41	19.6	61	42.5	85.3	1293.8	290.0	–	29.1	22.4
InstructBLIP [19]	Vicuna-7B	224	–	49.2	34.5	60.5	50.1	–	–	–	36	–	26.2
InstructBLIP [19]	Vicuna-13B	224	–	49.5	33.4	63.1	50.7	78.9	1212.8	291.9	–	–	25.6
Shikra [12]	Vicuna-13B	224	77.4*	–	–	–	–	–	–	–	58.8	–	–
IDEFICS-9B [42]	LLaMA-7B	224	50.9	38.4	35.5	–	25.9	–	–	–	48.2	–	–
IDEFICS-80B [42]	LLaMA-65B	224	60.0	45.2	36.0	–	30.9	–	–	–	54.5	–	–
Qwen-VL [6]	Qwen-7B	448	78.8*	59.3*	35.2	67.1	63.8	–	–	–	38.2	–	–
Qwen-VL-Chat [6]	Qwen-7B	448	78.2*	57.5*	38.9	<u>68.2</u>	61.5	–	1487.5	**360.7**	60.6	–	–
LLaVA-1.5 [55]	Vicuna-7B	336	78.5*	<u>62.0*</u>	50.0	66.8	58.2	<u>85.9</u>	1510.7	–	64.3	49.0	30.5
VILA-7B [52]	Llama-2-7B	336	79.9*	**62.3***	<u>57.8</u>	<u>68.2</u>	64.4	85.5	<u>1533.0</u>	296.1	**68.9**	70.0	**34.9**
Sugar	Vicuna-7B	336	76.0*	58.7*	**60.4**	**69.4**	57.5	**86.6**	**1550.8**	<u>300.0</u>	64.9	**75.6**	<u>31.3</u>

Table 1: Comparison with state-of-the-art methods on 11 visual-language benchmarks. We mark the best performance **bold** and the second-best <u>underlined</u>. Benchmark names are abbreviated due to space limits. VQA-v2 [27]; GQA [35]; VizWiz [28]; SQA^I: ScienceQA-IMG [61]; VQA^T: TextVQA [76]; POPE [51]; MME^P, MME^C: MME Perception, MME Cognition [95]; MMB: MMBench [58]; LLaVA^{Wd}: LLaVA-Bench(In-the-Wild)-Detail [56]; MM-Vet [99]. * indicates the training images of the datasets are observed during training.

4 Experiments

To assess Sugar's **generative** ability, we conduct a comprehensive comparison with state-of-the-art models on 11 commonly used visual-language benchmarks in Section 4.2. Furthermore, we evaluate more complicated multimodal comprehension tasks on DEMON with 29 datasets in Section 4.3. For **discriminative** tasks, we compare performance across three different retrieval tasks: image-text retrieval, interleaved retrieval, and fine-grained retrieval in Section 4.4. Subsequently, we leverage Sugar's discriminative ability for **retrieval-augmented generation** compared with common used retrieve module in Section 4.5. Finally, we conduct ablation experiments to analyze the effectiveness of our method in Section 4.6.

4.1 Setup

We apply our method to VILA [52], a recent state-of-the-art MLLM supporting interleaved input. We further fine-tune VILA using LoRA [30]. Details about the experiments setting, datasets and the instruction examples, please check in Appendix E.

4.2 Multimodal Comprehension on 11 Benchmarks

We conduct a comprehensive comparison with state-of-the-art models on 11 commonly used benchmarks, as shown in Table 1. Compared to existing models, Sugar achieves remarkable improvements over the second-best performing model on tasks requiring fine-grained semantics (*i.e.,* LLaVA^{Wd} [56], VizWiz [28], SQA [61] improve by 8%, 4.5%, 1.8% respectively) and benchmarks for detecting hallucinations (*i.e.,* POPE [51]), while maintaining competitive results in other tasks. Notably, Sugar excels in discriminative tasks and still achieves 5 state-of-the-art results and 3 second-best results on 11 benchmarks for generative tasks, even outperforming some models larger than 7B. Our results demonstrate the benefits of incorporating the discriminative loss, aiding in fine-grained semantic tasks and reducing hallucinations.

4.3 Complicated Multimodal Comprehension on DEMON

Table 2 demonstrates the superior performance of Sugar on the DEMON benchmark, which comprises 7 categories and a total of 29 sub-tasks. These tasks are considerably more complex than the previously used 11 common benchmarks. DEMON is tailored to evaluate the capacity of models and systems to understand demonstrative instructions that include multiple, interleaved, and multimodal contexts, presenting the essential information needed to complete a task. Sugar surpasses the previous state-of-the-art model on the DEMON benchmark, VPG-C [47], across 6 of 7 categories. For example, we achieve performance improvements of 36.1% in Text-Rich Images QA (TRQA) tasks and 17.2% in Visual Relation Inference (VRI) tasks, both of which require detailed semantics, compared to the second-best performing model. This underscores our advanced ability to associate interleaved

	LLM	MMD	VST	VRI	MMC	KGQA	TRQA	MMR
OpenFlamingo [4]	MPT-7B	16.9	24.2	13.9	21.7	32.0	30.6	41.6
BLIP-2 [49]	Vicuna-13B	26.1	21.3	10.7	17.9	39.2	33.5	39.7
InstructBLIP [19]	Vicuna-7B	33.6	24.4	11.5	21.2	47.4	44.4	48.6
MiniGPT-4 [103]	Vicuna-7B	13.7	17.1	8.0	16.6	30.3	26.4	43.5
LLaVA [56]	Vicuna-7B	7.8	10.7	8.3	15.9	36.2	28.3	41.5
mPlug-Owl [94]	LLaMA-7B	12.7	19.3	5.4	16.3	33.3	32.5	42.5
VPG-C [47]	Vicuna-7B	37.5	25.2	25.9	**22.2**	48.6	44.9	50.3
VILA-7B [52]	Vicuna-7B	47.8	25.8	13.2	17.2	60.1	42.1	50.5
Sugar	Vicuna-7B	**51.8**	**34.3**	**32.3**	16.8	**64.4**	**65.9**	**51.7**

Table 2: Comparision with state-of-the-art method on DEMON [47] benchmark.

(a) MSCOCO

Model	R@1	R@5	R@10
Text → Image			
FROMAGe(d)	**23.4**	47.3	59.0
FROMAGe(g+d)	23.4	47.2	58.0
Sugar	22.0	**49.1**	**63.1**
Image → Text			
FROMAGe(d)	**26.8**	52.4	63.6
FROMAGe(g+d)	26.4	52.3	63.4
Sugar	25.6	**53.6**	**66.7**

(b) VIST

Model	Inputs	R@1	R@5	R@10
CLIP ViT-L/14	5c	5.9	19.5	28.0
FROMAGe	5c	**11.9**	23.8	31.7
Sugar	5c	10.1	**26.3**	**36.2**
BLIP†	5c	6.2	16.8	23.4
CLIP ViT-L/14†	5c	8.8	22.3	29.8
FROMAGe†	5c	**13.2**	28.5	36.7
Sugar†	5c	11.0	**27.3**	**37.0**
CLIP ViT-L/14	5c+4i	2.4	21.3	34.0
FROMAGe†	5c+4i	18.2	42.7	51.8
Sugar†	5c+4i	**21.9**	**46.7**	**59.2**

(c) Winoground

Model	Text	Image	Group
VinVL	37.8	17.8	14.5
UNITER$_{large}$	38.0	14.0	10.5
VisualBERT$_{base}$	15.5	2.5	1.5
ViLLA$_{large}$	37.0	13.25	10.0
ViLT ViT-B/32	34.8	14.0	9.3
LXMERT	19.3	7.0	4.0
ViLBERT$_{base}$	23.8	7.3	4.8
FLAVA$_{ITM}$	32.3	20.5	14.3
FLAVA$_{contrastive}$	25.3	13.5	9.0
CLIP ViT-B/32	30.8	10.5	8.0
Sugar	**40.0**	**36.3**	**27.0**

Table 3: Retrieval results compared with previous models, reported by Recall@k for (a)(b) and Accuracy (%) for (c). **(a) MSCOCO** for image-text retrieval: FROMAGe(d) indicates the FROMAGe model pre-trained only with discriminative loss, and FROMAGe(g+d) indicates joint training with both discriminative and generative losses. **(b) VIST** for interleaved retrieval: † indicates retrieval over images not previously seen in the story sequence. "5c+4i" is shorthand for 5 captions and 4 images, and "5c" is shorthand for 5 captions. **(c) Winoground** for fine-grained retrieval.

text-image inputs for stronger in-context understanding, and Sugar's strong capability to capture global semantics in interleaved sequences, facilitated by joint training with discriminative loss.

4.4 Zero-shot Cross-modal Information Retrieval

Image-text Retrieval. We evaluated the performance of Sugar on the widely adopted MSCOCO [38] dataset in the context of a standard image-text retrieval task. Sugar demonstrated comparable performance to FROMAGe [40] in R@1 and surpassed it in R@5 and R@10, highlighting Sugar's superiority in normal retrieval tasks.

Interleaved Retrieval. To assess the proficiency of Sugar in processing multimodal contextual information, we evaluated its performance in retrieving relevant images conditioned on sequences of interleaved image-text inputs from the Visual Storytelling (VIST) dataset [32]. We conducted evaluations across several experimental configurations, following the same setup as FROMAGe [40] (see Appendix F.1). Our results show that Sugar outperforms FROMAGe in most settings, particularly achieving a 20.3% improvement in the 5c+4i configuration, significantly surpassing both CLIP and BLIP-2. This demonstrates that our method effectively leverages MLLMs' ability to handle complex interleaved sequence inputs, thereby achieving superior retrieval performance.

Fine-grained Retrieval. We tested fine-grained retrieval using the Winoground dataset [79], which evaluates the ability to perform vision-linguistic compositional reasoning. Surprisingly, Sugar outperformed all discriminative pre-trained models (both single-stream and dual-stream encoder architectures), achieving improvements of 5.3%, 77.1%, and 86.2% over the second-best model in the Text, Image, and Group dimensions, respectively. This demonstrates Sugar's strong capability to distinguish detailed semantics and performing compositional reasoning.

4.5 Retrieval-Augmented Generation

Due to Sugar's dual capabilities in both discrimination and generation, we can achieve retrieval augmentation without the need for an additional retrieval module. For performing retrieval-augmented generation (RAG), we selected two tasks, namely VizWiz and SQA^I, as they offer held-in data that were not seen during model training. We utilized a mixed set comprising the widely-used LLaVA-1.5 SFT subset and the held-in datasets of the two tasks as the knowledge base and employed different

(a) Retrieval-Augmented Generation		
	VizWiz	SQA^I
LLaVA-1.5	50.0	66.8
+CLIP	42.6(−14.8%)	62.0(−8.6%)
+BLIP2	43.0(−14.0%)	62.5(−6.4%)
VILA	57.8	64.4
+CLIP	49.3(−14.7%)	65.7(+0.6%)
+BLIP2	49.6(−14.2%)	66.1(+2.6%)
Sugar	60.4	69.4
+RAG	**61.9**(+2.5%)	**71.9**(+3.6%)

(b) Ablation Study						
	Generative Tasks			Discriminative Tasks		
	SQA^I	POPE	KGQA	MSCOCO	VIST	Winoground
Sugar	72.6	**86.6**	64.4	49.1	**46.7**	**36.3**
w/o data$_d$	**72.8**	85.3	61.7	**49.6**	42.0	34.8
w/o data$_g$	68.0	86.4	62.5	46.0	40.7	33.5
w/o GAK	72.1	86.0	61.1	48.2	33.5	29.3
w/o TK	71.6	85.3	63.9	49.0	44.1	20.5
w/o AvgPool	72.4	86.2	**64.6**	39.7	38.1	31.5

Table 4: (a) **Retrieval-Augmented Generation.** (b) **Ablation Study.** For MSCOCO, we report the R@5 in text-to-image retrieval. For VIST, we report the R@5 of retrieving an image given 5 captions and 4 images. For Winoground, we report the Image score. For other tasks, we report Accuracy (%).

retrieval modules to retrieve relevant knowledge for the MLLM. The results are as follows: *(i)* We observed a drop in performance for LLaVA-1.5 with RAG in all tasks. This may be because LLaVA is designed solely for single-image input, without the ability to utilize in-context external knowledge. *(ii)* Compared to VILA, Sugar's performance improved in both tasks, whereas VILA improved in SQA^I but decreased in VizWiz. These findings suggest that Sugar's retrieved knowledge is more beneficial, while the knowledge retrieved by CLIP and BLIP-2 may hinder performance.

4.6 Ablations

Importance of Both Tasks. (1) w/o data$_g$: As shown in Table 4, when we reduce the amount of data for discriminative tasks (Row 2), there are performance drops of 1.5% in hallucination detection tasks (*i.e.,* POPE) and 4.2% in interleaved multi-modal comprehension tasks (*i.e.,* KGQA). **(2) w/o data$_d$**: Similarly, reducing the data for generative tasks (Row 3), the performance on generative tasks declines with a 10.1% decrease in VIST, which requires global semantics capturing, and a 4.1% decrease in Winoground, which necessitates fine-grained semantic understanding. This indicates that **generative and discriminative training can mutually benefit each other**.

Effectiveness of Individual Components. (1) w/o GAK: When we exclude the Global Alignment Kernel (GAK) (Row 4) and resort to using the average similarity for the slices, a notable decrease in performance is observed across several interleaved image-text tasks (*i.e.,* a 5.1% decrease in KQGA and a 28.3% decrease in VIST). This underscores the fundamental role of GAK in aiding Sugar to capture global semantics effectively. **(2) w/o TK**: Upon removal of the Triple Kernel (TK) (Row 5) and utilization of CLIP for encoding the input sequence instead, a dramatic performance decline is evident in Winoground, with a 43.5% decrease. This underscores the significant role of TK in facilitating the distinction of detailed semantics. **(3) w/o AvgPool**: When solely using the last token for retrieval, a general decline in performance is observed across discriminative tasks, with decreases of 19.1% for MSCOCO, 22.6% for VIST, and 13.2% for Winoground. This phenomenon may be attributed to the last token of an image often corresponding to a pooling token, containing relatively weaker semantic information. Utilizing all image tokens and performing AvgPooling tends to yield greater improvements in retrieval tasks.

5 Conclusion

Vision-Language Models (VLMs) have been trained using both generative and discriminative paradigms, each with distinct advantages and limitations. To bridge this gap, we introduce **S**tructure-induced approach to **u**nify **g**enerative **a**nd disc**r**iminative paradigms, which imposes semantic relationships between input samples, thereby enhancing the MLLM's ability to capture global semantics and distinguish fine-grained details. This approach effectively balances generative and discriminative tasks, yielding synergistic benefits. Extensive experiments demonstrate the effectiveness of our approach, achieving state-of-the-art results in multiple generative tasks, particularly those requiring cognitive and discrimination abilities, while also demonstrating competitive performance in discriminative tasks such as image-text retrieval and achieving state-of-the-art results in interleaved and fine-grained retrieval. Furthermore, employing a retrieval-augmented generation strategy within a single model leads to additional improvements, offering a promising direction for future research.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022. 3

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 8

[5] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. *arXiv preprint arXiv:2410.08261*, 2024. 3

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*, 2023. 3, 4, 7, 19

[7] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1

[8] Oriol Barbany, Michael Huang, Xinliang Zhu, and Arnab Dhua. Leveraging large language models for multimodal search. *arXiv preprint arXiv:2404.15790*, 2024. 2, 3

[9] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. *Coyo-700m: Image-text pair dataset*, 2022. 20

[10] Yinghsan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. WebQA: Multihop and Multimodal QA. 2021. 2, 20, 30

[11] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*, 2023. 4, 19

[12] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 7

[13] Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z Pan, Zonggang Yuan, and Huajun Chen. Zero-shot visual question answering using knowledge graph. In *The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20*, pages 146–162. Springer, 2021. 30

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 19

[15] João Coelho, Bruno Martins, João Magalhães, Jamie Callan, and Chenyan Xiong. Dwell in the beginning: How language models embed long documents for dense retrieval. *arXiv preprint arXiv:2404.04163*, 2024. 2

[16] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017. 3

[17] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR, 2017. 17

[18] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–413. IEEE, 2007. 5, 17

[19] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7, 8

[20] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 22

[21] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024. 3

[22] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024. 3

[23] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[24] Minghe Gao, Juncheng Li, Hao Fei, Wei Ji, Guoming Wang, Wenqiao Zhang, Siliang Tang, and Yueting Zhuang. De-fine: Decomposing and refining visual programs with auto-feedback. *arXiv preprint arXiv:2311.12890*, 2023. 3

[25] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 3

[26] Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. Worldgpt: Empowering llm as multimodal world model. *arXiv preprint arXiv:2404.18202*, 2024. 24

[27] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 7, 20

[28] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 3, 7, 20

[29] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. 4

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 7, 20

[31] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021. 3

[32] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016. 8, 20

[33] Xiaohua Huang, Abhinav Dhall, Roland Goecke, Matti Pietikainen, and Guoying Zhao. A global alignment kernel based approach for group-level happiness intensity estimation. *arXiv preprint arXiv:1809.03313*, 2018. 2

[34] Xuanwen Huang, Wei Chow, Yang Wang, Ziwei Chai, Chunping Wang, Lei Chen, and Yang Yang. One graph model for cross-domain dynamic link prediction. *arXiv preprint arXiv:2402.02168*, 2024. 19

[35] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 7, 20

[36] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-seng Chua. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23013–23022, 2023. 4

[37] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 23

[38] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 8, 20, 23

[39] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 3

[40] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2, 2023. 2, 3, 8, 23, 24

[41] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

[42] Hugo Laurençon, Daniel van Strien, Stas Bekman, Leo Tronchon, Lucile Saulnier, Thomas Wang, Siddharth Karamcheti, Amanpreet Singh, Giada Pistilli, Yacine Jernite, et al. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023. *URL https://huggingface. co/blog/idefics. Accessed*, pages 09–18, 2023. 3, 7

[43] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012. 29

[44] Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*, 2023. 3

[45] Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, and Yueting Zhuang. Gradient-regulated meta-prompt learning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2551–2562, 2023. 3

[46] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 2

[47] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 3, 7, 8, 20, 26

[48] Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, Fei Wu, and Yueting Zhuang. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12601–12617, 2023. 3

[49] 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*, pages 19730–19742. PMLR, 2023. 3, 7, 8

[50] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[51] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3, 7, 20

[52] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023. 2, 3, 4, 7, 8, 19, 20, 22

[53] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 3

[54] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint*, (arXiv:2402.08327), 2024. 1, 3

[55] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3, 4, 7, 19, 20

[56] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 7, 8, 20

[57] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. 2

[58] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 7, 20

[59] Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Rar: Retrieving and ranking augmented mllms for visual recognition. *arXiv preprint arXiv:2403.13805*, 2024. 2, 3

[60] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 20

[61] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 7, 20

[62] Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning? *arXiv preprint arXiv:2404.12866*, 2024. 3

[63] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*, 2023. 3

[64] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[65] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022. 3

[66] Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*, 2024. 3

[67] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007. 2, 4, 17

[68] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5

[69] Kaihang Pan, Zhaoyu Fan, Juncheng Li, Qifan Yu, Hao Fei, Siliang Tang, Richang Hong, Hanwang Zhang, and Qianru Sun. Towards unified multimodal editing with enhanced knowledge collaboration. *arXiv preprint arXiv:2409.19872*, 2024. 3

[70] Kaihang Pan, Juncheng Li, Hongye Song, Hao Fei, Wei Ji, Shuo Zhang, Jun Lin, Xiaozhong Liu, and Siliang Tang. Controlretriever: Harnessing the power of instructions for controllable retrieval. *arXiv preprint arXiv:2308.10025*, 2023. 3

[71] Kaihang Pan, Juncheng Li, Wenjie Wang, Hao Fei, Hongye Song, Wei Ji, Jun Lin, Xiaozhong Liu, Tat-Seng Chua, and Siliang Tang. I3: I ntent-i ntrospective retrieval conditioned on i nstructions. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1839–1849, 2024. 3

[72] Kaihang Pan, Siliang Tang, Juncheng Li, Zhaoyu Fan, Wei Chow, Shuicheng Yan, Tat-Seng Chua, Yueting Zhuang, and Hanwang Zhang. Auto-encoding morph-tokens for multimodal llm. *arXiv preprint arXiv:2405.01926*, 2024. 3

[73] 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*, pages 8748–8763. PMLR, 2021. 1, 3, 5, 19, 22

[74] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[75] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023. 3

[76] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 7, 20

[77] Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Mike Bendersky. Quill: Query intent with large language models using retrieval augmentation and multi-stage distillation. *arXiv preprint arXiv:2210.15718*, 2022. 4

[78] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 22

[79] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 1, 3, 8, 20, 24

[80] Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024. 3

[81] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024. 2, 3, 21

[82] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223*, 2023. 2

[83] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 30

[84] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023. 23

[85] Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024. 1

[86] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*, 2024. 1

[87] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023. 5

[88] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 22

[89] Chenglin Yang, Siyuan Qiao, Yuan Cao, Yu Zhang, Tao Zhu, Alan Yuille, and Jiahui Yu. Ig captioner: Information gain captioners are strong zero-shot classifiers. *arXiv preprint arXiv:2311.17072*, 2023. 1, 2, 3

[90] Nan Yang, Tao Ge, Liang Wang, Binxing Jiao, Daxin Jiang, Linjun Yang, Rangan Majumder, and Furu Wei. Inference with reference: Lossless acceleration of large language models. *arXiv preprint arXiv:2304.04487*, 2023. 4

[91] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 3

[92] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *arXiv preprint arXiv:2302.04858*, 2023. 4

[93] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022. 4

[94] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 3, 8

[95] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 3, 7, 20

[96] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023. 4

[97] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[98] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21560–21571, 2023. 2, 3

[99] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 7, 20

[100] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 22

[101] Wenqiao Zhang, Jiannan Guo, Mengze Li, Haochen Shi, Shengyu Zhang, Juncheng Li, Siliang Tang, and Yueting Zhuang. Boss: Bottom-up cross-modal semantic composition with hybrid counterfactual training for robust content-based image retrieval. *arXiv preprint arXiv:2207.04211*, 2022. 4

[102] Wenqiao Zhang, Tianwei Lin, Jiang Liu, Fangxun Shu, Haoyuan Li, Lei Zhang, He Wanggui, Hao Zhou, Zheqi Lv, Hao Jiang, et al. Hyperllava: Dynamic visual and language expert tuning for multimodal large language models. *arXiv preprint arXiv:2403.13447*, 2024. 3

[103] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 8

[104] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36, 2024. 20

A Broader Impact

The broader impact of Sugar, carries both potential benefits and risks upon deployment and release. Some considerations are unique due to their visual nature, while others mirror existing instruction-following Large Language Models (LLMs). Built upon Vicuna, CLIP, DINOv2, and BGE, Sugar inherits issues associated with LLMs and vision encoders. Below, we outline risks and mitigation strategies for its release.

Hallucination.   Similar to other MLLMs, Sugar may generate outputs detached from facts or input data, posing concerns, especially in critical applications like medicine and the field related to security.

Biases.   Bias from base models can transfer to Sugar, originating from both the vision encoder (CLIP) and the language decoder (Vicuna), potentially leading to biased outcomes or unfair representations.

Ethical Impacts.   This study doesn't raise ethical concerns, as it doesn't involve subjective assessments or private data, only utilizing publicly available datasets.

Expected Societal Implications.   A significant societal concern lies in potential misuse, such as fabricating unauthorized texts leading to misinformation, privacy breaches, and other damaging consequences. Strong ethical standards and ongoing surveillance are essential for mitigation.

These issues aren't unique to our method but are prevalent across different techniques for multi-concept customization. Despite the risks, we believe the benefits outweigh the potential harm, allowing ongoing investigation and improvement of the model while engaging the community in developing better mitigation strategies. Moreover, its release can foster new applications and research directions, contributing to the progress and responsible deployment of foundation models in vision-language tasks.

B Limitations

(i) Our method, while effective, may inherit limitations from the underlying models, such as hallucination in generating outputs detached from facts or input data and potential biases originating from the model we used. *(ii)* The training data might inevitably contain mismatched image and text, which could adversely affect training.

C Mathematical Proof

Theorem 1. *The alignment kernel K can be computed in quadratic complexity, namely in $O(mnd^2)$ iterations. where m, n denotes the length of two sequence and their hidden dimension all is d, $m, n, d \in \mathbb{R}$.*

Proof. Given $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_m)$ two sequences of $\mathcal{X}^\star$, we set the double-subscripted series $M_{i,j}$ as $M_{i,0} = 0$ for $i = 1, ..., n$, $M_{0,j} = 0$ for $j = 1, ..., m$, and $M_{0,0} = 1$. Computing recursively for $(i, j) \in \{1, ..., n\} \times \{1, ..., m\}$ the terms

$$M_{i,j} = (M_{i,j-1} + M_{i-1,j-1} + M_{i-1,j})k(x_i, y_j)$$

we obtain that $K(\mathbf{x}, \mathbf{y}) = M_{n,m}$ The result can be proved by recursion and is intuitively an equivalent of the Dynamic Time Warping(DTW) [67, 17] algorithm where the max-sum algebra is simply replaced by the sum-product one [18].   □

Theorem 2. *triple kernel φ is a conditionally symmetric positive-definite kernel [18] defined on $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$.*

Proof. (i) for two slice $a, b \in \mathbb{R}^d$, meets $|a| = |b| = 2$, $d = d_1 + d_2$, $a = \text{concat}(a_1, a_2), b = \text{concat}(b_1, b_2)$, $a_1, b_1 \in \mathbb{R}^{d_1}, a_2, b_2 \in \mathbb{R}^{d_2}$ and $|a_1| = |a_2| = |b_1| = |b_2| = 1$:

17

let $a' = \text{concat}(a'_1, a'_2, a'_3)$, $a' = \text{concat}(b'_1, b'_2, b'_3)$: when $a'(b')$ is from text modal, we let $a' = \text{concat}(a_1, a_2, \mathbf{0})(b' = \text{concat}(b_1, b_2, \mathbf{0}))$, and $a' = \text{concat}(a_1, \mathbf{0}, a_2)(b' = \text{concat}(b_1, \mathbf{0}, b_2))$ for image modal. $\mathbf{0} \in \mathbb{R}^{d_2}$. we can unify the Equation 4 first and second case in:

$$
\begin{aligned}
\varphi(a, b) &= \varphi(a', b') \\
&= (|a'_2||b'_2| + |a'_3||b'_3|)||a'_1 - b'_1||^2 + (1 - |a'_2||b'_2|)||a'_2 - b'_2||^2 \\
&\quad + (1 - |a'_3||b'_3|)||a'_3 - b'_3||^2 \\
&\geq (|a'_2||b'_2| + |a'_3||b'_3|)||a'_1 - b'_1||^2 + 0 + 0 \\
&\geq 0
\end{aligned}
$$

As $(1 - |a'_2||b'_2|) \geq 0$ and $(1 - |a'_3||b'_3|) \geq 0$. for any family $\alpha_1, \alpha_2, ...\alpha_n \in \mathcal{X}$ and $c_1, c_2, ..., c_n \in \mathbb{R}$, we have that

$$
\sum_{i,j} c_i c_j \varphi(x_i, x_j) = \sum_{i,j} c_i c_j \varphi(x'_i, x'_j) \geq 0
$$

(ii) for two slice meets $|a| = |b| = 1$: for any family $\alpha_1, \alpha_2, ...\alpha_n \in \mathcal{X}$ and $c_1, c_2, ..., c_n \in \mathbb{R}$, we have that

$$
\sum_{i,j} c_i c_j \varphi(x_i, x_j) = \sum_{i,j} c_i c_j ||x_i - x_j||^2 \geq 0
$$

Additionally, it's evident that for both (i) and (ii), $\varphi(a, b) = \varphi(b, a)$. Therefore, triple kernel φ is a conditionally symmetric positive-definite kernel defined on $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ □

Theorem 3. *when both $\mathbf{x}$ and $\mathbf{y}$ contain only one slice, GAK is monotonically increasing with the directly calculated cosine similarity.*

Proof. let $\mathbf{x} = (x)$, $\mathbf{y} = (y)$, and cosine similarity of x and y is $cos = cos < x, y >$. we can get

$$
\sigma = \delta \sqrt{\frac{M + N}{2}} = \delta \sqrt{\frac{1 + 1}{2}} = \delta
$$

and we have $\varphi(a, b) = |a|^2 + |b|^2 - 2cos < a, b >= 2(1 - cos < a, b >)$, thus:

$$
\begin{aligned}
\varphi_\sigma &= \frac{1}{2\sigma^2} \varphi\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right) + \log\left(2 - e^{-\frac{\varphi\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right)}{2\sigma^2}}\right) \\
&= \frac{1}{2\delta^2} \varphi\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right) + \log\left(2 - e^{-\frac{\varphi\left(x_{\pi_1(i)}, y_{\pi_2(i)}\right)}{2\delta^2}}\right) \\
&= \frac{1 - cos < a, b >}{\delta^2} + \log\left(2 - e^{-\frac{1 - cos < a, b >}{\delta^2}}\right)
\end{aligned}
$$

Letting $t = \frac{1 - cos < a, b >}{\delta^2}$ and substituting the result of $\varphi_\sigma = t + \log(2 - e^{-t})$ into Equation 2, we obtain:

$$
\begin{aligned}
K(\mathbf{x}, \mathbf{y}) &= \sum_{\pi \in \mathcal{A}(\mathbf{x}, \mathbf{y})} \prod_{i=1}^{|\pi|} e^{-\phi_\sigma} \\
&= e^{-\phi_\sigma} \\
&= e^{-t + \log(\frac{1}{2 - e^{-t}})} \\
&= \frac{e^{-t}}{2 - e^{-t}}
\end{aligned}
$$

Letting $s = e^{-t}$, we can further obtain:

$$
K(\mathbf{x}, \mathbf{y}) = \frac{s}{2 - s}
$$

As $\cos\langle a, b\rangle \in [-1, 1]$, s strictly increases with $\cos\langle a, b\rangle$ and $s \in [e^{-\frac{2}{\delta^2}}, 1]$ when the hyperparameter δ is fixed. Derivative of $K(\mathbf{x}, \mathbf{y})$ can be obtained:

$$K(\mathbf{x}, \mathbf{y})' = \frac{2 - s + s}{(2 - s)^2} = \frac{2}{(2 - s)^2} > 0$$

Overall, when both $\mathbf{x}$ and $\mathbf{y}$ contain only one slice, GAK is monotonically increasing with the directly calculated cosine similarity. $\qquad\square$

D   Method Details

D.1   Architecture Details

We adopt the manifold multimodal model architecture [55, 52, 11, 6, 34], formulated as follows:

Visual Representation. We first process x_{img} subject to a visual representation backbone V_ω that outputs a sequence of features $p_{\text{img}} \in \mathbb{R}^{L \times h_{\text{vision}}}$ where $p_{\text{img}} = V_\omega(x_{\text{img}})$. As an example, p_{img} might be the patch features output by a Vision Transformer.

Vision-Language Projector. Next, we map p_{img} to a sequence of *embeddings* $e_{\text{img}} \in \mathbb{R}^{L \times h_{\text{text}}}$ via a learned projector F_ψ, where $e_{\text{img}} = F_\psi(p_{\text{img}})$.

Language Model. Finally, we concatenate the sequence e_{img} with the text prompt embeddings $e_{\text{prompt}} = \text{embed}(u_{\text{prompt}})$, passing the result to the language model. Generally, we have the interleaved image-text input x_{input} by concatting all the e_{prompt} and e_{img}. The language model generates output text $u_{\text{gen}} = \text{LM}_\theta(x_{\text{input}})$.

Retrieval Projector. For discriminative tasks, we select the token d_i from MLLM's hidden state and map it to r_i via a learned projector F_φ.

In Implementation, we utilize CLIP ViT-L/14 [73] as the visual encoder, and Vicuna 1.5 [14] as the language model.

D.2   Sequence Alignment

An alignment π of length $|\pi| = p$ between two sequences $\mathbf{x}$ and $\mathbf{y}$ is a pair of increasing p-tuples (π_1, π_2) such that

$$
\begin{aligned}
1 = \pi_1(1) \leq ... \leq \pi_1(p) = n, \\
1 = \pi_2(1) \leq ... \leq \pi_2(p) = m,
\end{aligned}
\tag{5}
$$

We write $\mathcal{A}(\mathbf{x}, \mathbf{y})$ for the set of all possible alignments between $\mathbf{x}$ and $\mathbf{y}$. Intuitively, an alignment π between $\mathbf{x}$ and $\mathbf{y}$ describes a way to associate each element of a sequence $\mathbf{x}$ to one or possibly more elements in $\mathbf{y}$, and vice versa. Such alignments can be conveniently represented by paths in the $n \times m$ grid displayed in the left of Figure 3.

with unitary increments and no simultaneous repetitions, that is $\forall 1 \leq i \leq p - 1$,

$$
\begin{aligned}
\pi_1(i + 1) \leq \pi_1(i) + 1, \quad \pi_2(j + 1) \leq \pi_2(j) + 1, \\
(\pi_1(i + 1) - \pi_1(i)) + (\pi_2(i + 1) - \pi_2(i)) \geq 1.
\end{aligned}
\tag{6}
$$

The score on a path is defined as:

$$S(\pi) = \sum_{i=1}^{|\pi|} \varphi(x_{\pi_1(i)}, y_{\pi_2(i)}) \tag{7}$$

E Experimental Details

E.1 Datasets

Training Data. Our vision-language task datasets are a subset of VILA [52], including MMC4 [104], COYO [9], LLaVA-1.5 SFT dataset [55].

We use a prompt template formatted as (system-message is a system prompt from Vicuna, and the following messages all have the same meaning.):

```
{system-message}. USER: <image>\n {question}. ASSISTANT: {answer}.
```

For interleaved vision-language datasets, the template is formatted as:

```
{system-message}. USER: {interleaving question}. ASSISTANT: {answer}.
```

Training Strategy. To jointly train the discriminative loss and generative loss, we calculate the loss as follows. Since the last token of an image is often a padding token, we take all 576 hidden state tokens before the LM head for images and apply average pooling to obtain a single token. For text, we directly take the last toke in the hidden state of MLLM.

During training, We calculate the discriminative loss using the last token from either the end of the text or the image in the MLLM's hidden state. Notably, in an interleaved input sequence with multiple texts or images, we randomly select multiple last tokens from the same sequence to more efficiently utilize the samples.

Evaluation Data. For generative tasks, we first evaluate on a wide range of question-answering tasks and some MLLM-oriented comprehension benchmarks, including VQA-v2 [27], GQA [35], VizWiz [28], ScienceQA-IMG [61], TextVQA [76], POPE [51], MME [95], MMBench [58], LLaVA-Bench (In-the-Wild) [56], MM-Vet [99] and. The split of test sets and the evaluation metrics are aligned with those described in VILA[52] and LLaVA [55].

To test the generative ability in interleaving tasks, we use DEMON [47], a comprehensive benchmark that demonstrative instruction following ability, including a wide variety of multi-modal datasets from different fields and scenarios.

For generation tasks, our evaluation encompasses MSCOCO [38] for image-text retrieval, Visual Storytelling (VIST) [32] for interleaved retrieval and Winoground [79] for fine-grained retrieval.

E.2 Training

We train the parameters for both the LLM and the MLP for embedding the MLLM's hidden state, initializing from VILA [52]. To enhance efficiency for the LLM, we employ LoRA tuning [30] on the W_q and W_v matrices using low-rank adaptation. In our implementation, we set the rank $r = 128$ and $\alpha = 256$. We utilize the AdamW optimizer [60] in conjunction with a cosine learning rate scheduler. The hyperparameters for the AdamW optimizer are configured with a warm-up ratio of 0.03 and a maximum learning rate of $1e - 4$. Training is conducted on 8 x A800 GPUs for approximately 12 hours.

E.3 Introduction Experiment Details

(a) WebQA. The original WebQA contains two types of questions: *"Qcate": "text"* (open-ended questions) and *"Qcate": "YesNo"* (binary judgment questions). For ease of evaluation, we only used the second type. We selected 500 samples of the "YesNo" question type from WebQA [10], each containing one relevant image-text pair and five unrelated image-text pairs. Since the original data provides responses in declarative sentences, we modified the answers of these samples to be either "yes" or "no" by prompting with *"please answer the question in Yes or No."*

We transformed this dataset into a question-answering format. Each question takes the following form (Due to display problems, we have performed line breaks, the same below.):

```
{system-message}. USER: {qustion}\n{image-text pairs}.\n
please answer the question in Yes or No.\n
ASSISTANT: {answer}.
```
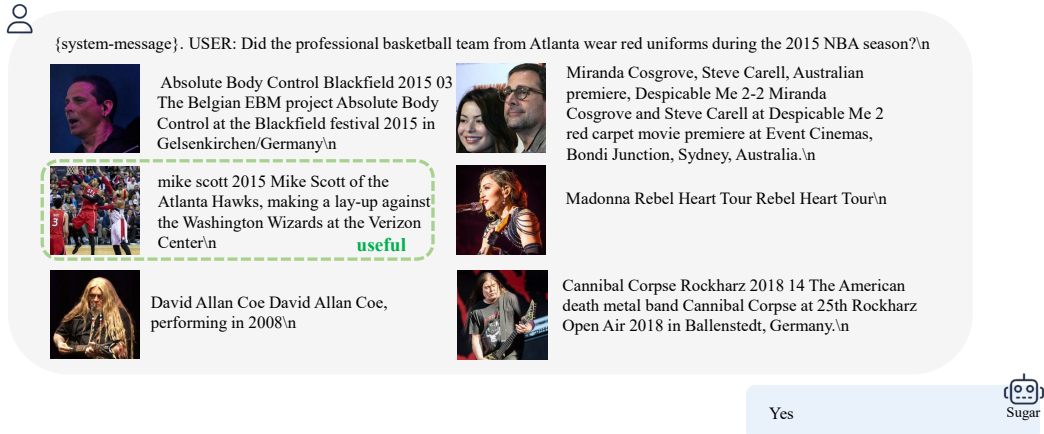
Here is a case for one sample in Figure 5:

Figure 5: A Case for WebQA. The index for the useful pair is three.

In WebQA, the accuracy roughly forms a "U" shape curve when the relevant image-text pair for a question appears at different positions. While Sugar also shows similar trends, it tends to be more stable overall. Specific numerical results can be found in Table 5.

Index	1	2	3	4	5	6
VILA	50.0	49.0	47.4	44.4	44.4	49.0
Sugar	63.8	60.4	59.6	59.2	60.4	61.0

Table 5: Specific accuracy (%) values displayed on WebQA. The index indicates the position of the useful image-text pair, denoting which position it occupies in the sequence.

(b) **MMVP-VLM Benchmark.** MMVP-VLM [81] contains 30 carefully annotated images in each dimension of capability, with pairs of images being highly similar to each other (as indicated by their high similarity scores in CLIP). To evaluate the discriminative ability of generative models on these finely nuanced images, we transformed this dataset into a question-answering format. Each question takes the following form:

```
{system-message}. USER: First Image:<image>\nSecond Image:<image>\n

which choice meets the first image:\n

A.{data["Statement"]}\nB.{data["Statement2"]}\n.please answer in A or B

ASSISTANT: {answer}.
```

Among them, both Statement 1 and Statement 2, as well as Image 1 and Image 2, are highly similar, with only subtle differences. Furthermore, there is a corresponding relationship between Image 1 and Statement 1, and between Image 2 and Statement 2. We employed a random seed to ensure that the correct answer is equally distributed between option A and option B. The specific values for the experiment are provided in Table 6.

	Image Size	⊘	⦿	⟳	↕⁝	⚲	🎨	⚙	A	◉	Average
OpenAI ViT-L-14 [73]	224^2	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [73]	336^2	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [100]	224^2	26.7	20.0	53.3	40.0	20.0	**66.7**	40.0	20.0	**53.3**	37.8
SigLIP ViT-SO-14 [100]	384^2	20.0	26.7	60.0	33.3	13.3	**66.7**	33.3	26.7	**53.3**	37.0
DFN ViT-H-14 [20]	224^2	20.0	26.7	73.3	26.7	26.7	**66.7**	46.7	13.3	**53.3**	39.3
DFN ViT-H-14 [20]	378^2	13.3	20.0	53.3	33.3	26.7	**66.7**	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [88]	224^2	13.3	6.7	**66.7**	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [88]	224^2	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [78]	224^2	6.7	26.7	40.0	6.7	13.3	**66.7**	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [78]	224^2	13.3	20.0	**66.7**	26.7	26.7	**66.7**	26.7	20.0	33.3	33.3
VILA-7B [52]†	336^2	36.7	46.7	53.3	43.3	50.0	60.0	50.0	46.7	50.0	48.5
Sugar†	336^2	**56.7**	**50.0**	63.3	**50.0**	**60.0**	**66.7**	**56.7**	**63.3**	**53.3**	**57.8**

Table 6: Performance Comparison of VILA and Various CLIP-Based Models on Different Visual Patterns in MMVP-VLM Benchmark. For most of the visual patterns, all CLIP-based methods show struggle, as evident from the scores. Sugar achieves state-of-the-art performance on the majority of tasks, demonstrating its powerful discriminative ability. We use symbols for visual patterns due to space limit: ⊘: Orientation and Direction, ⦿: Presence of Specific Features, ⟳: State and Condition, ↕⁝: Quantity and Count, ⚲: Positional and Relational Context, 🎨: Color and Appearance, ⚙: Structural and Physical Characteristics, A: Texts, ◉: Viewpoint and Perspective. † indicates that we use question-answering as the test method, instead of dot product.

E.4 Retrieval-Augmented Generation.

For performing retrieval-augmented generation (RAG), we selected two tasks, namely VizWiz and SQA^I, as they provide held-in data not seen during model training. We did not use VQA^{v2}, GQA, and VQA^T because their held-in data is a subset of the widely-used LLaVA-1.5 SFT. Benchmarks like POPE, MMB, and others lack held-in data. Therefore, we focused on VizWiz and SQA^I for our experiments. We utilized a mixed set comprising the widely-used LLaVA-1.5 SFT subset and the held-in dataset of the two tasks as the knowledge base and employed different retrieval modules to retrieve relevant knowledge for the MLLM. Similar to common practice, we average the similarity scores for CLIP (We choose CLIP ViT-L/14@336px). For BLIP-2, we compute the similarity using its multimodal token's CLS token. Figure 7 shows some specific retrieval results on test data, demonstrating that Sugar can better integrate information from both images and text, retrieving more similar data as external knowledge.
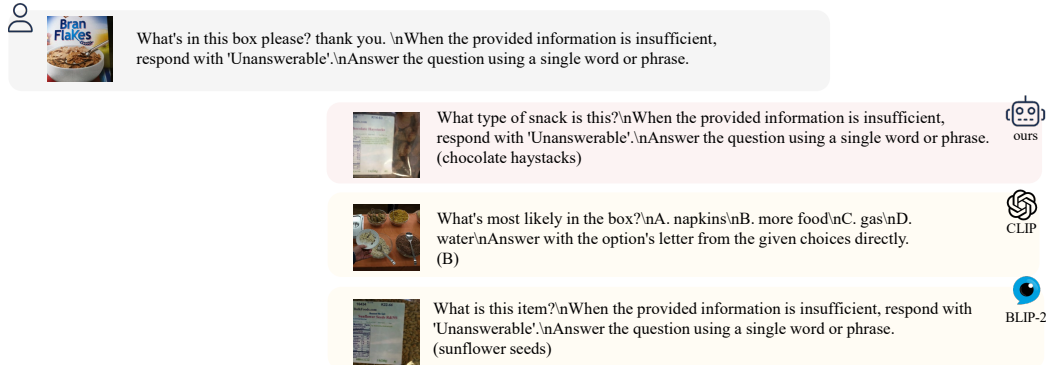

Figure 6: Selected examples from do retrieval-augmented generation.Sugar can retrieve more useful knowledge compared with CLIP and BLIP-2. Inside the parentheses are the answers, note that the When retrieving, we will only retrieve the questions, not the answers, which are shown here for convenience only.
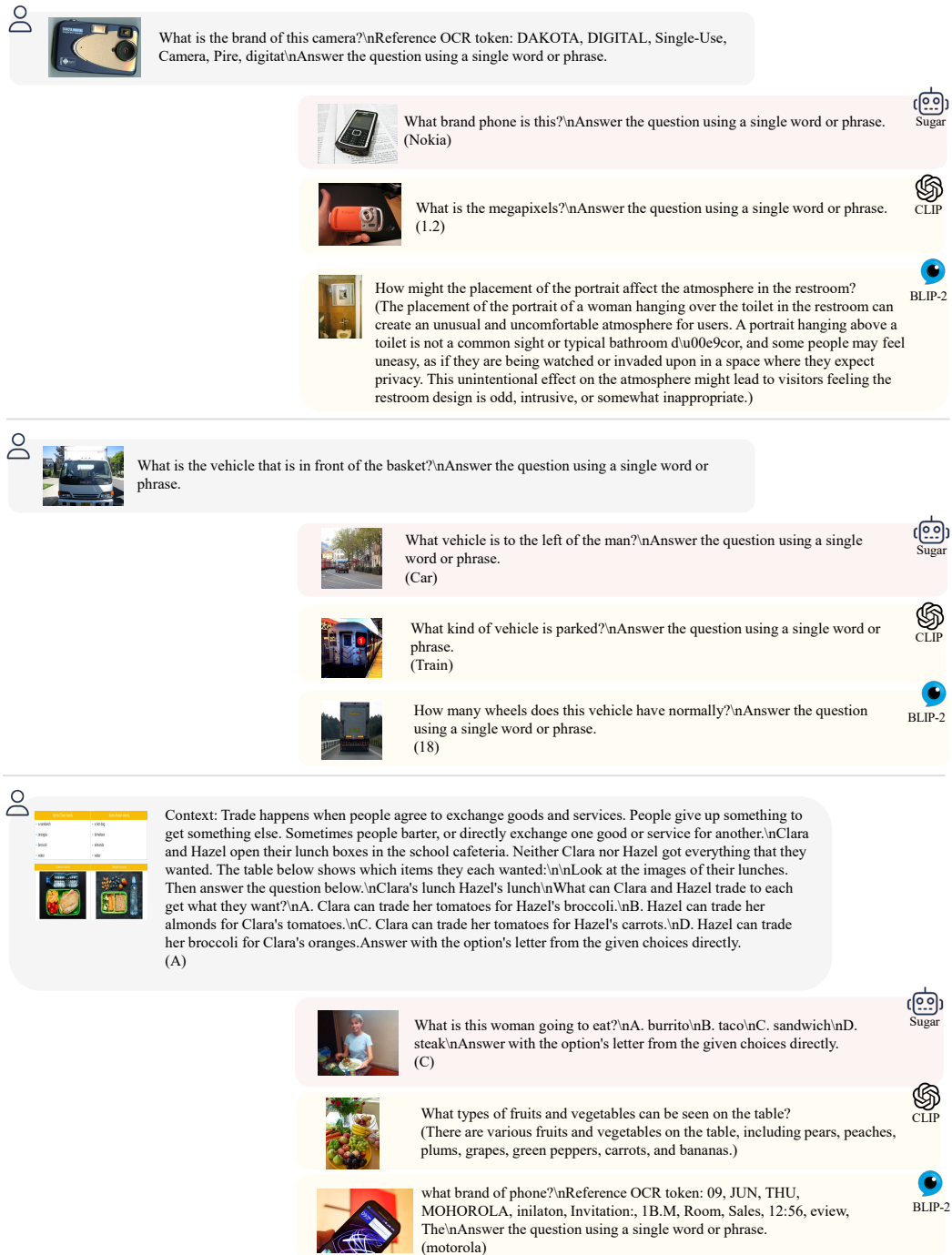
Figure 7: Selected examples from do retrieval-augmented generation (continued for Figure 6).

F More Results

F.1 Details of Retrieval

Image-text Retrieval. FROMAGe [40] was evaluated on the 5K validation set of MSCOCO 2017. Due to the split method confusion in FROMAGe, we report our image-text retrieval results on MSCOCO val2014's 5K val set following UniIR [84] and the Karpathy split [38]. What's more we then utilize FAISS [37], a powerful library for efficient similarity searches in dense vector spaces, to

index and retrieve candidates. Therefore, the results may exhibit slight differences when compared under identical settings. The results in Table 3(a) are provided for reference only.

Interleaved Retrieval. We conduct evaluations across several experimental configurations, following the same setup as FROMAGe [40]. The settings are as follows:

1. Retrieval of the last image given the descriptions of the preceding 5 images. This evaluates models' ability to condition on temporally dependent language.

2. Retrieval of the last image given the descriptions of the preceding 5 images and the 4 preceding images. This assesses models' capability to process interleaved image-and-text context.

Fine-grained Retrieval. Winoground [79] is designed to evaluate the ability of vision and language models to perform vision-linguistic compositional reasoning. The task involves matching two images with two captions, where both captions contain an identical set of words/morphemes arranged in different orders. This dataset, meticulously hand-curated by expert annotators, includes a rich set of fine-grained tags to facilitate detailed performance analysis.

F.2 Quality Results

To analyze Sugar's emergent behaviors and observed weaknesses, we present additional qualitative samples that were not included in the main paper due to space constraints. Please note that for brevity, we have omitted the system prompts and the line breaks after the images for all the quality examples.

We hope these additional results and observations showcase the potential of Sugar in various application areas. In future work, it is important to investigate these emergent behaviors more thoroughly and to understand the underlying mechanisms that enable Sugar to demonstrate such generalization abilities. This will pave the way towards building better MLLMs, including enhancing robustness, reducing biases, and improving the alignment and scope of the learned vision-language representations.

World Knowledge: We observe that Sugar can leverage the world knowledge [26] embedded within the LLM to enhance performance on multimodal tasks. For example, as shown in Figure 4, the model understands that during Halloween, people typically dress up in various ways to portray scary, funny, or creative characters, such as ghosts and skeletons.

Retrieval the Same Sequence at Different Place: One interesting emergent behavior of Sugar is its ability to retrieve sequences from different positions within the input interleaved sequence, demonstrating flexibility and high sample efficiency, as shown in Figure 4. Unlike CLIP, which requires encoding each sample separately, Sugar can encode sequences of varying lengths for the same multi-modal document in a single forward pass.

What's more, Sugar is capable of both retrieval and generation tasks. Below in Figure 8 are some examples from the VIST dataset.
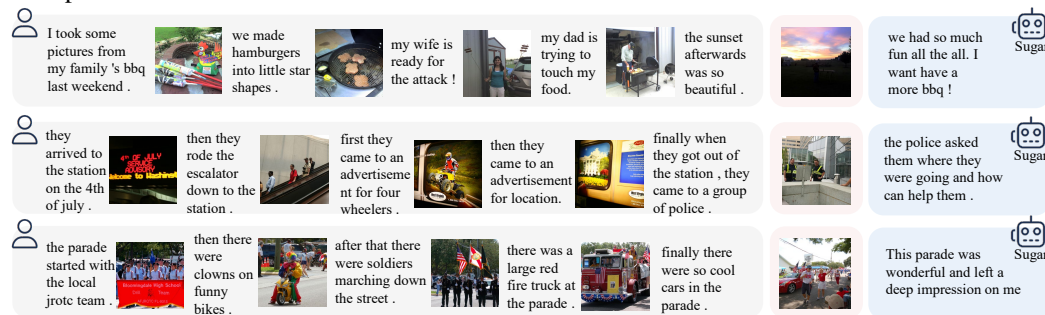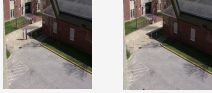

Figure 8: Selected examples for various image-text tasks. The pink background indicates retrieval results, while the blue background indicates generated results.

Fine-grained Image Discrimination: As shown in Figure 9, Sugar excels at accurately discerning subtle differences between images and identifying detailed objects and their attributes. VILA, on the other hand, tends to describe the content of the images without pinpointing the precise differences between them. In contrast, Sugar provides more concise and direct answers.

What distinguishes these two images?   Difference:

VILA: The main difference between the two images is the perspective. In the first image, the view is from a higher angle, giving a broader view of the parking lot and the building. In contrast, the second image is taken from a lower angle, focusing on the entrance of the building. This difference in perspective allows for a more detailed view of the entrance and the surrounding area in the second image.

Sugar: the two people that were standing on the sidewalk are gone.

 which one of following meets the image:\n
A.a car facing back.\n
B.a car facing front.\n
please answer in A or B with the content of the choice.

VILA: B.a car facing front.

Sugar: A.a car facing back.

 which one of following meets the image:\n
A.A painting on the wall behind the bed.\n
B.Empty on the wall behind the bed.\n
please answer in A or B with the content of the choice.

VILA: B.Empty on the wall behind the bed.

Sugar: A. A painting on the wall behind the bed.

 which one of following meets the image:\n
A.points at \"0\"\n
B.points at \"80\"\n
please answer in A or B with the content of the choice.

VILA: A.points at \"0\".

Sugar: B.points at \"80\".

Source Image: Target Image: Please give a editing Request to describe the transformation from the source image to the target image.

VILA: The image depicts a young girl with blonde hair, wearing a purple shirt, standing in front of a white house. The girl is smiling and looking directly at the camera. The image has been edited to remove the background, focusing solely on the girl. The background has been replaced with a solid color, which is not specified in the image description. The girl's position relative to the house has not been altered, and she is still standing in front of it. The image does not contain any text or additional objects. The girl's smile and direct gaze are the main focus of the image.

Sugar: Change the girls face Zoom it in slightly.

Figure 9: Selected examples. Sugar excels at accurately discerning subtle differences between images and identifying detailed objects and their attributes.

Style Following: Sugar exhibits a certain degree of in-context style following capability. As shown in Figure 10, with the aid of external knowledge, Sugar partially adopts the style of retrieved results, resulting in more accurate and detailed answers compared to scenarios without retrieval augmentation.

What's the best weather, season, time of the day of visiting this place? Is the time when this photo was taken a good time to visit this place?

Question

Sugar

Yes.

Answer

+RAG

What are the things I should be cautious about when I visit here?

When visiting this serene lakeside location, be cautious about the following:\n\n1. Wooden pier safety: Ensure the T-shaped wooden pier is sturdy and safe to walk on, and be mindful of any loose planks or slippery surfaces.\n\n2. Swimming conditions: If you plan to swim in the lake, check for potential hazards such as submerged objects or strong underwater currents, and be aware of the water temperature.\n\n3. Wildlife: Some lakes may have local wildlife, such as snakes or alligators, which could pose a threat. Remain cautious and observant, especially near the water's edge.\n\n4. Weather conditions: The sky is cloudy, so be prepared for potential rainfall or changes in temperature, and monitor weather forecasts before your visit.\n\n5. Hiking in the surrounding area: If you plan to explore the green trees and mountains in the background, be cautious of potential hazards like steep and slippery terrain or unstable trail conditions. Carry appropriate hiking gear and wear sturdy footwear.\n\n6. Protect the environment: Be respectful of the natural beauty of the area, and avoid littering, damaging any trees or plants, or disturbing wildlife.

Retrieved External Knowledge

The best time to visit this serene lakeside location is during the day, as the weather is clear and the lighting is good for photography. The photo was taken during the day, and the weather appears to be clear, making it a good time to visit.

Answer

What is unusual about this image?

Question

Sugar

A man ironing clothes on car.

Answer

+RAG

What is the problem this city might be facing? What are some possible solutions?

The city might be facing traffic congestion issues, particularly in the four-lane sections of the highway since most cars have their brake lights on, indicating slow or stopped traffic. Possible solutions could include:\n\n1. Implementing traffic management strategies, such as ramp meters to control the rate at which cars enter the highway or optimizing traffic signal timings to improve traffic flow.\n2. Encouraging the use of public transportation, carpooling or ride-sharing to reduce the number of cars on the road.\n3. Expanding or adding more lanes to ease traffic congestion in the four-lane sections of the highway.\n4. Exploring alternative routes to divert traffic from the congested areas.\n5. Improving walking and biking infrastructure in the city, thus promoting more sustainable and non-motorized modes of transport.

Retrieved External Knowledge

The unusual aspect of this image is the presence of a man standing on the hood of a car while ironing clothes. This is not a common sight and could potentially be dangerous.

Answer

Figure 10: Selected examples from LLaVA-Bench(In-the-wild). Using external knowledge, Sugar partially follows the style of retrieved results, providing more accurate and detailed answers compared to not using retrieval augmentation.

Interleaved Comprehension: As demonstrated by results from DEMON [47], Sugar exhibits superior interleaved comprehension capabilities compared to VILA, particularly in tasks requiring fine-grained analysis and an understanding of global context. For instance, in the third example of Figure 11, VILA confuses character names, whereas Sugar maintains narrative coherence while adhering to the style of the preceding text. Similarly, in the third example of Figure 12, VILA provides an irrelevant response, while Sugar delivers a more contextually appropriate answer. Additionally, Figure 13 demonstrates Sugar's ability to effectively capture global information, identifying the relevant images and text within the sequence to provide accurate responses.

👤 Given two images and a sentence outlining their relation, your responsibility is to judge whether the sentence correctly describes the connection between them.\n
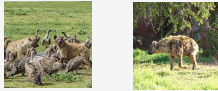
Relation: The left image shows at least one rear-facing hyena standing in front of two larger standing animals with horns.

🤖 True
VILA

🤖 False
Sugar

👤 Having viewed the images, can you use the information presented to answer the following question?\n

 Image Caption #1: StadiumTrack Track inside Olympic Stadium for the Monster Spectacular . \n

 Image Caption #2: Stade Olympique Monteal A match of football of Montreal Impact at the Oympic Stadium of Montreal. \n

Question: Is the center of the Olympic Stadium of Montreal always the same?

🤖 The answer is yes.
VILA

🤖 No, the center of the Olympic Stadium of Montreal is not always the same.
Sugar

👤 Given the progression of the story with the first few images, can you write a fitting end considering the last image?\n

 Caption#1:Wilma and Betty are sitting on a couch in the living room. Wilma speaks to Betty and cries into a handkerchief.\n

Caption#2:Wilma and Betty are sitting on a couch in the living room. Wilma is crying and wiping her tears with a handkerchief while Betty speaks to her. \n Caption#3:Wilma and Betty are sitting on the couch in the living room in front of the television set. Wilma is holding a handkerchief and talking to Betty.\n

 Caption#4:Fred and Barney are in a car. Barney is looking at Fred and holding the steering wheel. Fred speaks, then Barney speaks.\n Caption#5:

🤖 Wilma and Betty are sitting on the couch in the living room in front of the television set. Wilma is holding a handkerchief and talking to Betty.
VILA

🤖 Fred and Barney are in a car. Barney is driving while Fred is talking to Barney while they look behind them.
Sugar

Figure 11: Selected examples for Interleaved Comprehension.

Utilizing the information, including images, text, and tables that I provide, could you provide a correct answer to the following question. \n

Global Table:

	Year	Title	Roles	Notes
0	2015	Señorita Pólvora	Juanita	Recurring role: 11 episodes
1	2015	El Dandy	Deyanira	Recurring role
2	2016	El Vato	Estela Núñez	Recurring role (season 1): 9 episodes
3	2016–2017	La Doña	Lálya Corona	Series regular (season 1): 102 episodes
4	2017	Las Malcriadas	Yuridia Corona	Recurring role: 11 episodes
5	2018	Tres Milagros	Milagros Cruz "Nikita"	Main role: 55 episodes
6	2018	Falco	Sonia García	Recurring role: 7 episodes
7	2018	Diablero	Keta	Recurring role: 8 episodes
8	2020	Narcos vs Zombies	TBA	

Context:

Question: How many people are in the car in diablero?

VILA: There are five people in the car in the image.

Sugar: There are four people in the car in the image.
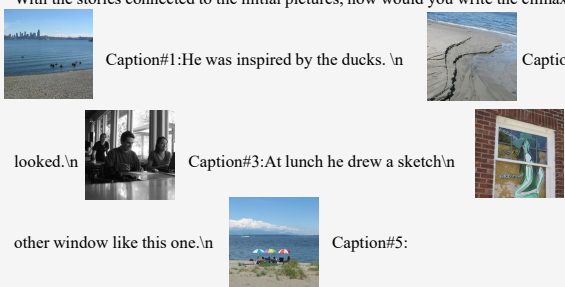
Provided with a series of diagrams from a textbook, your responsibility is to correctly answer the following question.\n

Question: How many layers of the atmosphere is identified in the diagram?

VILA: There are four layers of the atmosphere identified in the diagram.

Sugar: There are five layers of the atmosphere identified in the diagram.

With the stories connected to the initial pictures, how would you write the climax based on the last picture?

Caption#1:He was inspired by the ducks. \n

Caption#2:He liked the way the sand that had washed looked.\n

Caption#3:At lunch he drew a sketch\n

Caption#4:He would hang the painting in the other window like this one.\n

Caption#5:

VILA: He was inspired by the ducks.

Sugar: Then he would take a picture of the beach to put in the other window.

Figure 12: Selected examples for Interleaved Comprehension (continued for Figure 11).

Did the Cadillac Series 61 Fastback come as a four door sedan only?\n

Image Caption #1: 1949 Cadillac Series 61 Fastback - Flickr - exfordy (1)

Image Caption #2: 1949 Cadillac Series 61 Fastback - Flickr – exfordy.

Image Caption #3: Cadillac Series 62 51 Cadillac Limousine.\n

Image Caption #4: 1949 Cadillac Series 62 Convertible – fvr.\n

Image Caption #5: 1949 Cadillac Series 62 Convertible – fvl.\n

Image Caption #6: 1949 Cadillac Series 62 Convertible - fvr2. \n
please answer the question in Yes or No.

Yes.   VILA

No.   Sugar

Are there any trees near the HSBC Hong Kong Headquarters building which are taller than the building?\n

Image Caption #1: HSBC Hong Kong headquarters building, Hong Kong, Mar 06.\n

Image Caption #2: HSBC Headquarters Building, Hong Kong, detail of top showing cannons. \n   \ Image Caption #3: Hong Kong HSBC headquarters building IMG 5378 HSBC Hong Kong headquarters building, Hong Kong.\n

Image Caption #4: Hong Kong HSBC headquarters building IMG 5377 HSBC Hong Kong headquarters building, Hong Kong.\ n

Image Caption #5: HSBC Hong Kong Headquarters.\n   Image Caption #6: HK HSBC Main Building 2008 Facade of HSBC Hong Kong headquarters building, and others.“ \n please answer the question in Yes or No.

Yes.   VILA

No.   Sugar

Figure 13: Selected examples for Interleaved Comprehension (continued for Figure 12).

Sensitivity with Detailed Semantics: Sugar can address various examples inspired by the Winograd schema [43]. These examples consist of multiple sentences that differ only by a single word, leading to different resolutions of ambiguity. Sugar can accurately match images and text, demonstrating its sensitivity to even minor changes in input prompts. Figure 14 showcases some cases that align with the Winograd schema from Winoground.

The dog wears a hat but the person doesn't   Sugar

the person wearing neutral colors poses and the person wearing brighter colors takes a picture   Sugar

the person wears a hat but the dog doesn't   Sugar

the person wearing brighter colors poses and the person wearing neutral colors takes a picture   Sugar

the plant is eating the bug   Sugar

head in the clouds   Sugar

an inflatable flamingo on a person   Sugar

the bug is eating the plant   Sugar

clouds in the head   Sugar

a person on an inflatable flamingo   Sugar

Figure 14: Selected examples from Winoground. Sugar is Sensitivity with Detailed Semantics

29

G   Retrieval for Knowledge-based VQA

In this section, we use FVQA [83] and WebQA [10], two knowledge-based VQA datasets, to verify Sugar's effectiveness of combining retrieval and comprehension abilities in a single model, thereby avoiding compatibility issues and suboptimal performance.

Historically, solving FVQA has relied on modeling the knowledge database using a Knowledge Graph [13]. For WebQA, each question is associated with 10-20 knowledge bases, but only one is relevant to the image and caption. FVQA knowledge is textual, whereas WebQA knowledge consists of both text and pictures.

Implement Details. In this experiment, we used CLIP ViT-L/14@336px, and both experiments report the ROUGE-L Score.

For FVQA, answers originate from two sources: directly from the image or from the knowledge base. To minimize interference, we only tested questions requiring the knowledge base. We used the following prompt for FVQA: "Please answer the questions based on the pictures. If the reference information is useful, please use it. Otherwise, please ignore the reference information. Reference information: retrieved knowledge <image> question." The *baseline* without retrieval means we did not search for knowledge, but directly input the image and question for the model to answer. + *CLIP image* means using the image to retrieve knowledge, + *CLIP text* means using the text to retrieve knowledge, and 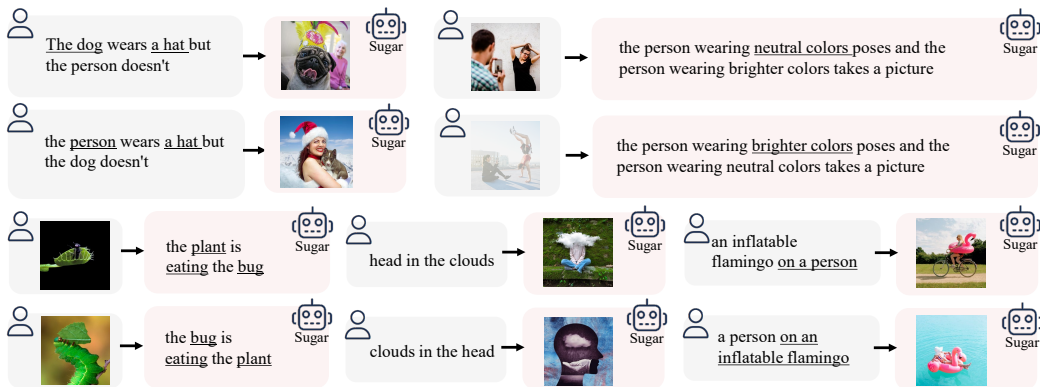+ *CLIP average* means using the average annotations of both image and text to retrieve knowledge. For our model, *sugar+rag* indicates the average result obtained using both image and text to retrieve knowledge.

For WebQA, each question has 10-20 negative captions and images. Due to context length limitations in LLaVA and VILA, we could not input all the data, necessitating a retrieval model to extract relevant knowledge. Due to the large dataset size, we randomly selected 1000 samples. For WebQA, +*CLIP image* means providing the positive image and using it to retrieve the most relevant text from the knowledge base, which is then used as input for the model to answer the question. Conversely, +*CLIP text* uses the text to retrieve relevant images. For our model, *sugar+rag* indicates the result obtained using the average similarity score of the aforementioned methods.

	FVQA	WebQA
LLaVA-1.5-7B	5.9	/
LLaVA-1.5-7B + CLIP image	6.8	81.8
LLaVA-1.5-7B + CLIP text	7.1	79.2
LLaVA-1.5-7B + CLIP (average)	7.9	/
VILA-7B	6.4	/
VILA-7B + CLIP image	9.0	80.0
VILA-7B + CLIP text	10.2	71.2
VILA-7B + CLIP (average)	11.0	/
Sugar	6.5	/
Sugar + rag	**20.7**	**88.7**

Table 7: Comparison between the independent generator + retriever and Sugar on knowledge-based VQA. '/' indicates not applicable.

Results. Based on Table 7, we can observe that while MLLM can answer a small portion of FVQA questions using its internal knowledge, it still requires the support of a retriever for enhanced accuracy. However, the impact of retrieval strategies on the results is inconsistent. For instance, using text retrieval often outperforms image retrieval in FVQA, whereas in WebQA, image retrieval is more effective. Additionally, there are compatibility issues between retrieval strategies and models. For example, in WebQA, VILA is more sensitive to CLIP's retrieval strategy, with fluctuations 3.4 times greater than those of LLaVA-1.5. Our integrated retriever and generator model, however, does not require an additional retriever and avoids the aforementioned optimization and selection issues.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims presented in the abstract and introduction provide an accurate representation of the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Appendix B.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We give the proof in Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification:We give experimental setup and implementation details in Section 4.1 and Appendix E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The codes will come soon and all the data is public to access.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: we have provided necessary implementation details of our method in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We followed the baseline settings on the evaluation benchmark.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We give the statements of experiments compute resources in Appendix E.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conducted in the paper conforms, in every respect, to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the Broader Impacts in Appendix A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have already cited all the original paper that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.