Symmetrical Visual Contrastive Optimization: Aligning Vision-Language Models with Minimal Contrastive Images

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, Nick Haber
Stanford University

https://s-vco.github.io/

Abstract

Recent studies have shown that Large Vision-Language Models (VLMs) tend to neglect image content and over-rely on language-model priors, resulting in errors in visually grounded tasks and hallucinations. We hypothesize that this issue arises because existing VLMs are not explicitly trained to generate texts that are accurately grounded in fine-grained image details. To enhance visual feedback during VLM training, we propose S-VCO (**S**ymmetrical **V**isual **C**ontrastive **O**ptimization), a novel finetuning objective that steers the model toward capturing important visual details and aligning them with corresponding text tokens. To further facilitate this detailed alignment, we introduce MVC, a paired image-text dataset built by automatically filtering and augmenting visual counterfactual data to challenge the model with hard contrastive cases involving **M**inimal **V**isual **C**ontrasts. Experiments show that our method consistently improves VLM performance across diverse benchmarks covering various abilities and domains, achieving up to a 22% reduction in hallucinations, and significant gains in vision-centric and general tasks. Notably, these improvements become increasingly pronounced in benchmarks with higher visual dependency. In short, S-VCO offers a significant enhancement of VLM's visually-dependent task performance while retaining or even improving the model's general abilities.

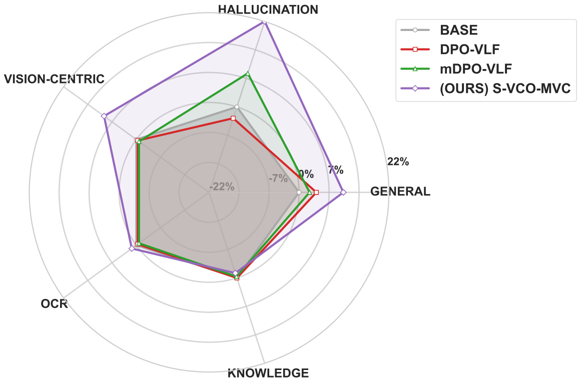


Figure 1: **Improvement over the base-VLM grouped by the test ability domain of benchmarks.** Our S-VCO delivers the most significant overall improvement across nearly all domains, with particularly strong gains in reducing visual hallucinations. In vision-centric and general capability domains, S-VCO also achieves considerable performance boosts over the base-VLM, outperforming existing preference tuning methods including DPO and mDPO (discussed in more detail in §5.2).

1 Introduction

Large Vision-Language Models (VLMs) tend to over-rely on their language models, leading to neglect of visual content. This problem manifests as visual hallucinations across tasks like perception (Tong et al., 2024b), reasoning (Chen et al., 2024), and in-context learning (Jia et al., 2024). Studies like Tong et al. (2024a) show that VLMs exhibit only limited performance gains when vision inputs are enabled compared to having no vision inputs. This behavior is reflected in the metrics across many popular vision-language benchmarks (Lu et al., 2022; Yue et al., 2024; Lu et al., 2023; Kembhavi et al., 2016; xAI, 2024; Singh et al., 2019). Our own perplexity-based evaluations reveal similar patterns of **visual neglect**. As illustrated in Fig. 2, we measured a base VLM's perplexity (PPL) when generating a caption matched to one image (*e.g.*, "image$_{\text{match}}$" with a "dog") while contradicting the other image (*e.g.*, "image$_{\text{mismatch}}$" with a "cat"). We also tested the model's PPL without any vision input ("no image"). Intuitively, the model should exhibit the lowest PPL when given the matching image, as the aligned visual information would make the caption easier to generate. However, results reveal the opposite pattern: the model's PPL lowest when no image input is provided at all, and highest when presented with the correct image (the "dog") – higher than the mismatched image (the "cat") as condition. Across

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30284–30297
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

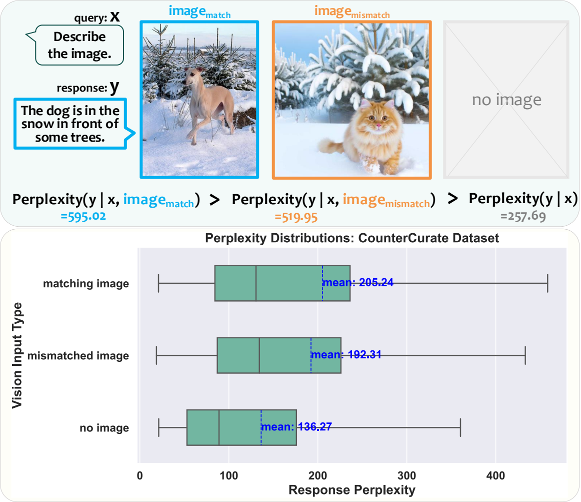


Figure 2: **Upper Part**: A comparison of a VLM's perplexity (PPL) for generating a text caption when the input is an image matching the text, an image contradicting the text, or no image input at all. Intuitively, the PPL should be lowest when the image matches the text. However, the current VLM exhibits the lowest PPL without any image input and the highest PPL given the matching image. **Lower Part**: This counterintuitive pattern holds across 1, 000 random examples with these visual counterfactual pairs extracted from CounterCurate dataset (§4.1).

1, 000 random samples from CounterCurate (Zhang et al., 2024) — a large collection of such paired counterfactual images — the average perplexity distribution follows this counterintuitive pattern. This highlights the model's tendency to disregard visual information even when they are critical for generating accurate texts.

Recent works (Wang et al., 2024; Xie et al., 2024; Jiang et al., 2024a; Luo et al., 2024; Fu et al., 2025) attempt to address similar issues of visual neglect by adapting Direct Preference Optimization (DPO) (Rafailov et al., 2024) to compare image inputs. However, these methods treat visual supervision as a "preferential" tuning paradigm, where an original image is preferred, and a cropped or noisy version of that image is dispreferred. As shown in Fig. 3, these noisy images lack meaningful connections to their paired texts, making it easy for the model to learn shortcuts by rejecting the corrupted image versions without fully understanding visual details and aligning them to the corresponding texts.

Instead, we posit that the "**preference**" paradigm should be improved with a purely **contrastive** framework, as the "dispreferred" image is merely another image misaligned with a given text. Building upon this insight, we propose **S**ymmetrical **V**isual **C**ontrastive **O**ptimization (**S-VCO**), a finetuning objective that enforces precise correspondence between visual details and textual tokens. S-VCO rewards the model for attending to matching images and strongly rejecting contradictory images with incorrect details. To further avoid shortcut learning, S-VCO incorporates **symmetry** by flipping the objective for contradictory responses, allowing the "negative" image to serve as the "preferred" visual condition when paired with its corresponding text.

Additionally, as cropping (Wang et al., 2024) or adding diffusion noise to an original image (Jiang et al., 2024a) fails to provide meaningful comparisons in visual details, we construct **MVC**, a dataset of paired images with **M**inimal **V**isual **C**ontrasts, each matched to contrastive textual responses given a shared query. Building on recent visual counterfactual data sources (Zhang et al., 2024; Liu et al., 2024c), we implement a vision-centric filter to select visually challenging pairs and an LLM augmentation scheme to diversify texts, forming an instruction-response-styled dataset suitable for VLM finetuning.

Experiments demonstrate the effectiveness and versatility of our approach, as S-VCO consistently enhances VLM performance across diverse benchmarks spanning various abilities and domains. As shown in Fig. 1, S-VCO achieves significantly greater improvements over the base-VLM, particularly in reducing visual hallucinations, while excelling in vision-centric tasks and offering considerable gains on general benchmarks. Compared to existing VLM preference tuning methods (DPO, Rafailov et al., 2024; mDPO, Wang et al., 2024), S-VCO combined with MVC delivers more substantial and comprehensive performance boosts.

In summary, our main contributions are:

- S-VCO, a novel VLM finetuning objective that enforces strict and balanced visual contrastive supervision through the symmetrical alignment of image-text pairs;
- MVC, a dataset of minimal contrastive image pairs accompanied with corresponding textual responses for a shared query. MVC is constructed automatically through vision-centric filtering and augmentation based on visual counterfactual data;
- The combination of S-VCO and MVC significantly boosts VLM performance across diverse benchmarks, particularly in visually dependent tasks, without compromising general capabilities.

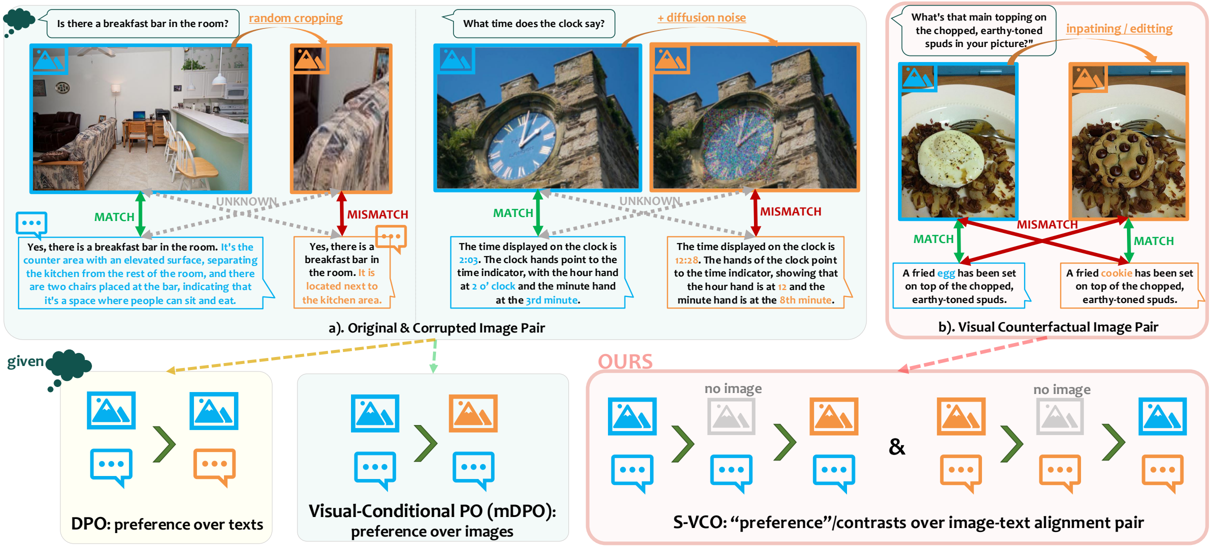

Is there a breakfast bar in the room?
random cropping
UNKNOWN
MATCH
MISMATCH
Yes, there is a breakfast bar in the room. It's the counter area with an elevated surface, separating the kitchen from the rest of the room, and there are two chairs placed at the bar, indicating that it's a space where people can sit and eat.
Yes, there is a breakfast bar in the room. It is located next to the kitchen area.
What time does the clock say?
+ diffusion noise
UNKNOWN
MATCH
MISMATCH
The time displayed on the clock is 2:03. The clock hands point to the time indicator, with the hour hand at 2 o' clock and the minute hand at the 3rd minute.
The time displayed on the clock is 12:28. The hands of the clock point to the time indicator, showing that the hour hand is at 12 and the minute hand is at the 8th minute.
a). Original & Corrupted Image Pair
What's that main topping on the chopped, earthy-toned spuds in your picture?"
inpatining / editting
MATCH
MISMATCH
MATCH
A fried egg has been set on top of the chopped, earthy-toned spuds.
A fried cookie has been set on top of the chopped, earthy-toned spuds.
b). Visual Counterfactual Image Pair
given
DPO: preference over texts
Visual-Conditional PO (mDPO): preference over images
OURS
no image
&
no image
S-VCO: "preference"/contrasts over image-text alignment pair

Figure 3: **Upper Part: MVC of visual counterfactual images [b)] in comparison to the image pair data used in prior work [a)] (§4).** MVC's image pair differs in meaningful visual details that are also grounded in the associated texts [b)], while corrupting original images with random cropping or adding noise leads to images that are not aligned with the texts directly derived from language preference data [a)]. **Lower Part: S-VCO in comparison to existing VLM preference tuning paradigms DPO and visual-conditional PO (§3).** Unlike prior methods that treat visual supervision as uni-modal preferences, S-VCO considers the contrast of the image-text pair as a whole. It rewards the model for attending to matching images and rejecting contradictory ones (§3.1), while using a symmetrical mechanism to switch the role of each image-text pair, thus avoiding shortcut learning (§3.2).

2 Preliminaries

Our visual contrastive objective is inspired by Direct Preference Optimization (DPO) (Rafailov et al., 2024) that contrasts pairs of textual responses. When applied to VLMs, DPO incorporates the image as an additional prefix condition (Zhou et al., 2024; Li et al., 2023a; Yu et al., 2024; Sarkar et al., 2024). Let π_θ be the policy VLM to be optimized, and π_{ref} the fixed reference model (typically the unfinetuend VLM in DPO's framework) used to measure how the finetuned policy π_θ improves. Given a query q (an instruction prompt), an image i, and a pair of responses – one preferred y_w (*winning*) and one dispreferred y_l (*losing*) – the **DPO** objective for VLMs can be formulated as:

$$L_{\text{DPO}} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|i,q)}{\pi_{\text{ref}}(y_w|i,q)} - \beta\log\frac{\pi_\theta(y_l|i,q)}{\pi_{\text{ref}}(y_l|i,q)}\right) \quad (1)$$

As illustrated in Fig. 3, the DPO formulation focuses on supervising differences between **language responses**, which however, often pertain to wording choices and stylistic variations rather than reflecting meaningful visual distinctions. Consequently, the VLM is being trained like a language model to weigh over text formulations, rather than to fully utilize the image as a grounding input.

Recent approaches, such as Wang et al. (2024); Jiang et al. (2024a), adapt the DPO framework to contrast image conditions. The original image i becomes the preferred image i_w (*winning*), while a negative image i_l (*losing*) is derived through random cropping (Wang et al., 2024) or adding diffusion noise (Jiang et al., 2024a) (see Fig. 3). The updated **Vis**ual **Con**ditional objective is:

$$L_{\text{VisCon}} = \quad (2)$$
$$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|i_w,q)}{\pi_{\text{ref}}(y_w|i_w,q)} - \beta\log\frac{\pi_\theta(y_w|i_l,q)}{\pi_{\text{ref}}(y_w|i_l,q)}\right)$$

The above visual conditional preference formulation **shifts the target of preference onto the images**, where the original image i_w is always preferred over the corrupted version i_l. However, this approach shares similar limitations with DPO, as it prioritizes distinguishing image differences without necessarily grounding those differences in the associated texts. Since i_l is typically a noisy variant of i_w with no definitive relations to y_w, the model could easily rely on superficial visual features to discern the image pair. This encourages shortcut learning, where the model rejects "unrealistic" images like i_l without examining the visual details related to the text tokens.

Our S-VCO addresses these limitations by introducing a stricter visual-conditioned objective (§3.1) and a symmetrical construct that aligns both i_w and

i_l with their respective y_w and y_l (§3.2). Fig. 3 illustrates S-VCO in comparison to existing preference tuning paradigms, showcasing its emphasis on grounded visual-textual alignment.

3 S-VCO: Symmetrical Visual Contrastive Optimization

3.1 Visual Contrastive Supervision

S-VCO enforces a strict visual focus by optimizing the model for two key behaviors:

1) Attending to matching images. The model is rewarded for prioritizing relevant visual details in the matching image i_w as a condition when predicting the corresponding response y_w. This directly addresses the tendency of VLMs to overlook visual content (§1) and is achieved through the term:

$$L_{\text{Attend}}(i_w, y_w) = \\ -\log \sigma \left(\beta_1 \log \frac{\pi_\theta(y_w|i_w,q)}{\pi_{\text{ref}}(y_w|i_w,q)} - \beta_1 \log \frac{\pi_\theta(y_w|q)}{\pi_{\text{ref}}(y_w|q)} \right). \tag{3}$$

2) Rejecting contradictory images. When presented with a contrastive image i_l containing visual details that directly contradict the response y_w, the model must strongly reduce the likelihood of predicting y_w under this incorrect image condition. Intuitively, the model should assign minimal probability to a response that directly contrasts with the visual input. This behavior is modeled as:

$$L_{\text{Reject}}(i_l, y_w) = \\ -\log \sigma \left(\beta_2 \log \frac{\pi_\theta(y_w|q)}{\pi_{\text{ref}}(y_w|q)} - \beta_2 \log \frac{\pi_\theta(y_w|i_l,q)}{\pi_{\text{ref}}(y_w|i_l,q)} \right) \tag{4}$$

By combining these two components, our strict visual contrastive objective is defined as:

$$L_{\text{VCO}}(i_w, y_w, i_l) = L_{\text{Attend}}(i_w, y_w) + L_{\text{Reject}}(i_l, y_w) \tag{5}$$

3.2 Symmetrical Alignment

Unlike prior "preference"-based approaches, S-VCO treats i_w and i_l as mere images with contrastive details, where either can serve as the correct (*i.e.*, "preferred") condition depending on the paired textual response (y_w or y_l). While we inherit the notation of i_w and i_l, S-VCO does not assign an inherent "winning" or "losing" property to the images. Instead, an image is considered "winning" only when paired with its corresponding text. For instance, i_l, typically labeled as "losing" in preference tuning methods, becomes a "winning" (preferred) condition when the target response is y_l that matches its visual details.

A one-sided formulation that consistently favors i_w over i_l risks encouraging shortcut learning, where the model rejects i_l based on superficial, text-unrelated features (*e.g.*, visual style differences). This issue could arise because most of the contrasting images are synthesized via inpainting or image editing (see §4). To address this, S-VCO introduces symmetry by flipping the objective in Eq. 5, treating i_l as the preferred condition when paired with its corresponding response y_l. This effectively switches the roles of "winning" and "losing" for the image pair:

$$L_{\text{VCO}}(i_l, y_l, i_w) = L_{\text{Attend}}(i_l, y_l) + L_{\text{Reject}}(i_w, y_l). \tag{6}$$

As defined above, this encourages the model to attend to i_l and reject i_w in the flipped case where the target response is y_l.

The complete S-VCO objective incorporates symmetry by summing over both roles of i_w and i_l given their corresponding target responses:

$$L_{\text{S-VCO}} = L_{\text{VCO}}(i_w, y_w, i_l) + L_{\text{VCO}}(i_l, y_l, i_w) \tag{7}$$

This symmetrical alignment ensures balanced optimization, allowing both i_w and i_l to contribute equally to the model's learning. It promotes true alignment between images and texts without relying on shortcuts, such as rejecting either the image (Eq. 2) or the text (Eq. 1). In essence, S-VCO optimizes for the alignment of **image-text pairs** rather than simply one modality.

4 Minimal Visual Contrasts Dataset

4.1 Visual Counterfactual Data

To complement S-VCO, we introduce **MVC**: a dataset of paired visual contrastive samples designed to enhance the model's ability to discern visual details. Built on existing sources (*e.g.*, Zhang et al., 2024; Liu et al., 2024c; Gaur et al., 2024), MVC contains image pairs with minimal but meaningful variations, accompanied by corresponding contrastive texts. The curated data includes four key contrast types, as shown in part **a).** of Fig. 4: **Object Replacement** (changing a specific object); **Attribute Replacement** (modifying an object's features like color, shape, or size); **Count Modification** (altering the number of objects); **Position Flipping** (reversing relative positions of objects). Except for the last type, these counterfactual images are generated through controlled image synthesis, including inpainting, editing, and generation (Li et al., 2023b; Zhuang et al., 2025; Betker et al.).

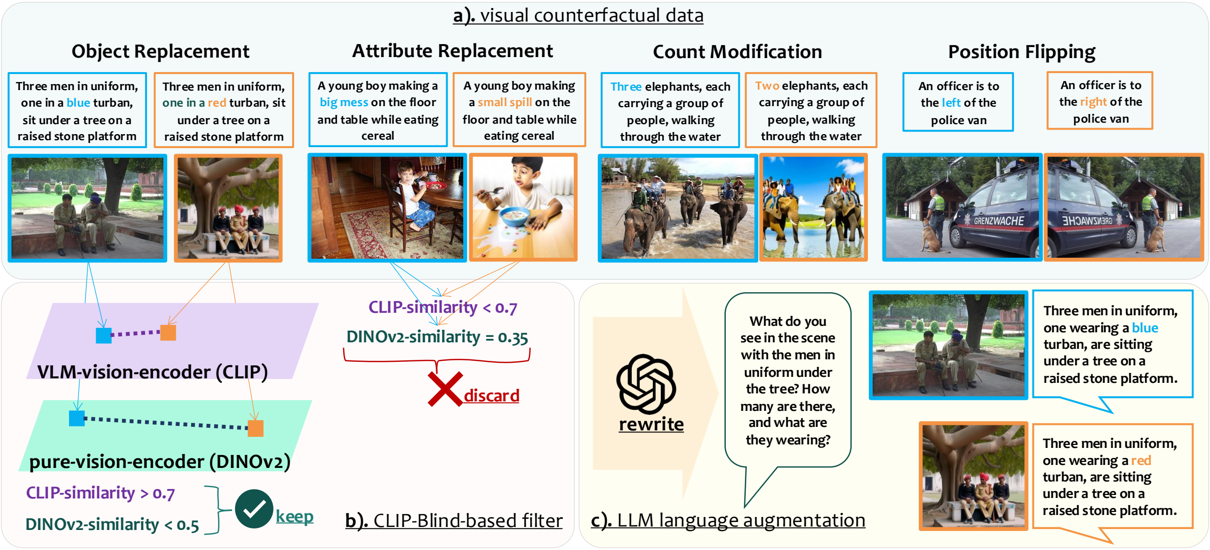

a). visual counterfactual data
Object Replacement
Attribute Replacement
Count Modification
Position Flipping
Three men in uniform, one in a blue turban, sit under a tree on a raised stone platform
Three men in uniform, one in a red turban, sit under a tree on a raised stone platform
A young boy making a big mess on the floor and table while eating cereal
A young boy making a small spill on the floor and table while eating cereal
Three elephants, each carrying a group of people, walking through the water
Two elephants, each carrying a group of people, walking through the water
An officer is to the left of the police van
An officer is to the right of the police van
GRENZWACHE
VLM-vision-encoder (CLIP)
pure-vision-encoder (DINOv2)
CLIP-similarity > 0.7
DINOv2-similarity < 0.5
keep
CLIP-similarity < 0.7
DINOv2-similarity = 0.35
discard
b). CLIP-Blind-based filter
rewrite
What do you see in the scene with the men in uniform under the tree? How many are there, and what are they wearing?
Three men in uniform, one wearing a blue turban, are sitting under a tree on a raised stone platform.
Three men in uniform, one wearing a red turban, are sitting under a tree on a raised stone platform.
c). LLM language augmentation

Figure 4: **MVC dataset outline**: **a).** Types of visual counterfactuals sourced from Zhang et al. (2024); Liu et al. (2024c); **b).** Our vision-centric filter that keeps only image pairs whose CLIP-similarity > 0.7 to select hard samples for current VLMs, while ensuring meaningful visual differences with DINOv2-similarity < 0.5; **c).** Rewriting captions into conversational queries and responses without changing the explicit minimal visual contrasts.

In this work, we use CounterCurate (Zhang et al., 2024) and FineCops-Ref (Liu et al., 2024c) as our visual counterfactual data sources. Refer to Appx.A for dataset statistics.

4.2 Filtering and Language Augmentation

Filter. Existing visual counterfactual datasets offer a large quantity of detailed contrasts, but their quality is inconsistent due to the synthetic nature of the data, which could lead to pairs where the generated image fails to truly contradict the original. To address this, we implement a vision-centric filter inspired by the CLIP-Blind concept (Tong et al., 2024b) to select image pairs based on the following two criteria, as shown in part **b).** of Fig. 4: **1. Different in detailed visual features**: Image pairs must exhibit meaningful contrasts, especially in detailed visual features. To achieve this, we employ DINOv2 (Oquab et al., 2023), a vision-only model with more robust visual feature representations (Singh et al., 2023). Pairs with high similarity in DINOv2's purely visual representation space are discarded, as they may not contain the desired degree of contrasts. **2. Semantically close & Hard for VLMs**: Image pairs must also be semantically similar overall and difficult for current VLMs to distinguish. Therefore, we embed images using the same CLIP vision encoder used by the VLM (Radford et al., 2021; Zhai et al., 2023) and retain pairs with relatively high similarity in the CLIP space that focuses on images' overall visual semantics. In this way, we exclude pairs with overly distinct content and less significant contrasts in visual details. In practice, we use `DINOv2-Large` (Oquab et al., 2023) with a similarity threshold of 0.5, and `SigLIP-400M` (Zhai et al., 2023) as the CLIP encoder with a similarity threshold of 0.7. Together, these thresholds ensure that MVC focuses on meaningful visual contrasts while maintaining semantic relevance and difficulty for the VLM.

Language Augmentation. Although visual counterfactual data sources provide image contrasts, their original textual descriptions are typically short captions, which are unideal for VLM finetuning. To address this, we augment the data using a two-step process with a strong LLM (`gpt-4o`), as shown in part **c).** of Fig. 4: **1. Generating queries**: For each pair of captions, we prompt the LLM to generate a conversational question as if the captions were natural responses to that question, while ensuring that the contrasts remain explicit. **2. Rewriting and diversifying**: We prompt the LLM to rephrase the original captions lacking clarity on key contrasts, thus enabling a more natural flow of language given the generated question and the contrastive details. Overall, the augmentation step results in conversational instruction-response pairs that are more suited for VLM finetuning, and more aligned with the visual details in the contrasting image pairs. Our prompt templates are provided in Appx.C.

After the filtering and language augmentation, our MVC dataset comprises over $11,000$ pairs of minimal contrastive images matched with accurate and diverse conversational queries and responses (Appx.A). The effectiveness of these data processing steps is empirically discussed in §5.3.

5 Experiments

5.1 Setup

Baseline methods. We compare S-VCO against DPO (Rafailov et al., 2024) and mDPO (Wang et al., 2024)[1] containing the visual-conditional objective in Eq. 2 as baseline finetuning approaches.

Training data. We use two datasets: our proposed MVC with minimal contrastive image-text pairs and VLFeedback (VLF) (Li et al., 2023a), a classical instruction-tuning dataset for VLMs that includes preferred and dispreferred response pairs. We follow mDPO (Wang et al., 2024)'s sample of ~10,000 data points from VLF. Refer to Appx.A for more dataset implementation details.

Base VLMs. We use two pretrained models hosted on Huggingface: `LLaVA-1.5-7B` (Liu et al., 2024a) – denoted as LV-1.5(-7B), and `LLaVA-Next-Interleave-7B` (Li et al., 2024b) – denoted as LV-INT(-7B).

Evaluation. We evaluate the models on a wide range of benchmarks spanning various ability domains. Following the categorization by Tong et al. (2024a), these benchmarks include: *General*: LLaVABench (Liu et al., 2024b), MMVet (Yu et al., 2023); *Hallucination*: MM-Hal (Sun et al., 2023); *Vision-Centric*: CVBench (Tong et al., 2024a), MMVP (Tong et al., 2024b), RealworldQA (xAI, 2024); *OCR*: TextVQA (Singh et al., 2019); *Knowledge*: SQA (Lu et al., 2022).

Implementation details. Refer to Appx.B for detailed training and inference configurations.

5.2 Main Results

Tab. 1 presents detailed evaluation results for different methods across various benchmarks grouped by ability domains. The final column shows the average improvement (%) over the base VLM across all metrics, summarizing each model's overall performance. Key findings are highlighted below.

S-VCO achieves consistently superior performance across benchmarks in various domains. Fig. 1 illustrates performance improvements over the base-VLM (LV-INT) for each benchmark category. Our S-VCO consistently outperforms baseline methods across nearly all domains, with only a slight drop in knowledge-heavy tasks like ScienceQA that relies minimally on visual input (discussed more below). The most significant gain is in visual hallucination tasks, where S-VCO enhances the base model by **~22%**. Considerable improvements are also observed in vision-centric (+**~10%**) and general domains (+**~11%**). Compared to baseline methods, DPO and mDPO trained on the standard instruction-tuning dataset VLF, S-VCO with MVC delivers much greater and more consistent improvements across domains. While recent approaches like mDPO (Wang et al., 2024) improve visual hallucination metrics, its effects on other abilities are less notable. In contrast, S-VCO not only excels in visually demanding tasks but also achieves considerable gains in other domains.

S-VCO shows increasing benefits as benchmarks become more visually dependent. We quantify **visual dependency** of a metric as the percentage drop in a base-VLM's performance when image inputs are removed. Using LV-INT as the base model, we rank benchmarks by visual dependency (*e.g.*, MMVP accuracy drops to 0 without image inputs), and plot the improvement trends of different methods in Fig. 6. Dotted lines show fitted regressions, and shaded areas represent variances. S-VCO demonstrates increasingly pronounced improvements as visual dependency rises, aligning with its design focus on strengthening visual detail recognition (§3). On highly visually-dependent benchmarks such as MM-Hal, LLaVABench, MMVet and MMVP, our S-VCO delivers the most substantial gains. In contrast, on SienceQA (SQA), which benefits merely **~4%** from image inputs, we observe a minor drop in performance after S-VCO, as visual information plays very little role in this task. Compared to other methods, SFT degrades performance on visually dependent metrics (§5.3), while preference tuning approaches show positive trends. Among all, S-VCO achieves the most significant trend of gains with increasingly vision-intensive tasks.

Qualitative performance highlights S-VCO's superior visual understanding. In Fig. 5, qualitative examples from various benchmarks further

[1] In practice, mDPO combines textual DPO (Eq. 1) with the visual-conditional version (Eq. 2) and an absolute reward regularization.

	Benchmarks	Hallucination		Vision-Centric			General		OCR	Knowledge	TOTAL
		MMHal		CVBench	MMVP	RQA	MMVet	LVBench	TextVQA	SQA	avg_impr.
	Models$_{\text{TrainSet}}$	score	hal_rate↓	acc.	acc.	acc.	score	score	acc.	acc.	over BASE
LV-1.5-7B	BASE	2.16	57%	59.3	21.3	35.3	30.46	61.2	46.40	66.78	0%
	DPO$_{\text{VLF}}$	2.06	65%	57.0	16.7	39.5	31.65	68.1	49.16	66.71	−1.25%
	DPO$_{\text{MVC}}$	2.45	53%	63.2	22.0	42.1	33.53	66.0	**49.43**	66.07	+8.11%
	mDPO$_{\text{VLF}}$	2.39	57%	53.2	18.7	**44.2**	31.79	68.4	41.71	66.52	+2.11%
	mDPO$_{\text{MVC}}$	2.29	56%	59.4	20.7	35.2	31.51	63.0	46.42	66.80	+1.26%
	S-VCO$_{\text{MVC}}$	**2.75**	**46%**	**63.5**	**25.3**	43.0	**34.68**	**69.5**	49.16	**67.22**	**+14.26%**
LV-INT-7B	BASE	2.74	46%	67.6	41.3	57.5	41.01	74.5	59.45	74.77	0%
	DPO$_{\text{VLF}}$	2.70	48%	68.1	42.7	54.4	40.87	81.0	59.20	**74.79**	+0.10%
	DPO$_{\text{MVC}}$	2.92	38%	68.8	46.7	57.9	41.51	83.5	59.76	73.80	+5.78%
	mDPO$_{\text{VLF}}$	2.97	42%	68.0	42.0	54.8	40.46	79.3	58.97	74.53	+2.07%
	mDPO$_{\text{MVC}}$	3.04	38%	70.0	44.7	57.9	41.24	80.2	59.43	74.65	+5.43%
	S-VCO$_{\text{MVC}}$	**3.28**	**35%**	**71.0**	**50.7**	**58.2**	**44.04**	**85.0**	**60.26**	73.87	**+10.47%**
	S-VCO$_{\text{MVC-Raw}}$	3.10	35%	71.3	46.0	58.8	40.55	86.5	59.22	73.87	+7.73%
	VCO$_{\text{MVC}}$	3.16	39%	70.1	46.0	56.2	42.61	84.8	59.87	74.44	+6.82%
	SFT$_{\text{MVC}\times 2}$	2.68	46%	66.6	38.7	56.5	38.94	70.7	59.32	74.77	−2.45%

Table 1: **Performance of different methods applied to two base-VLMs, tested across benchmarks grouped by ability domains**. VLF refers to VLFeedback used in mDPO; MVC is our minimal visual contrastive dataset; RQA and SQA represent RealworldQA and ScienceQA (§5.1). The last column shows the average percentage of improvement over the base-VLM across all metrics. **Best scores are in boldface**, second-best underlined. Our S-VCO demonstrates consistent improvement across domains, achieving the most significant enhancement over the base-VLMs overall. The last three rows present ablation results (§5.3): 1. S-VCO on unfiltered and unaugmented visual counterfactual data (MVC-Raw); 2. S-VCO without the symmetrical construct (VCO, §3.1); 3. SFT using both sides of image-text pairs from MVC (SFT$_{\text{MVC}\times 2}$). These results highlight the importance of our data preprocessing (§4.2) and the symmetrical objective (§3.2) for optimal performance.

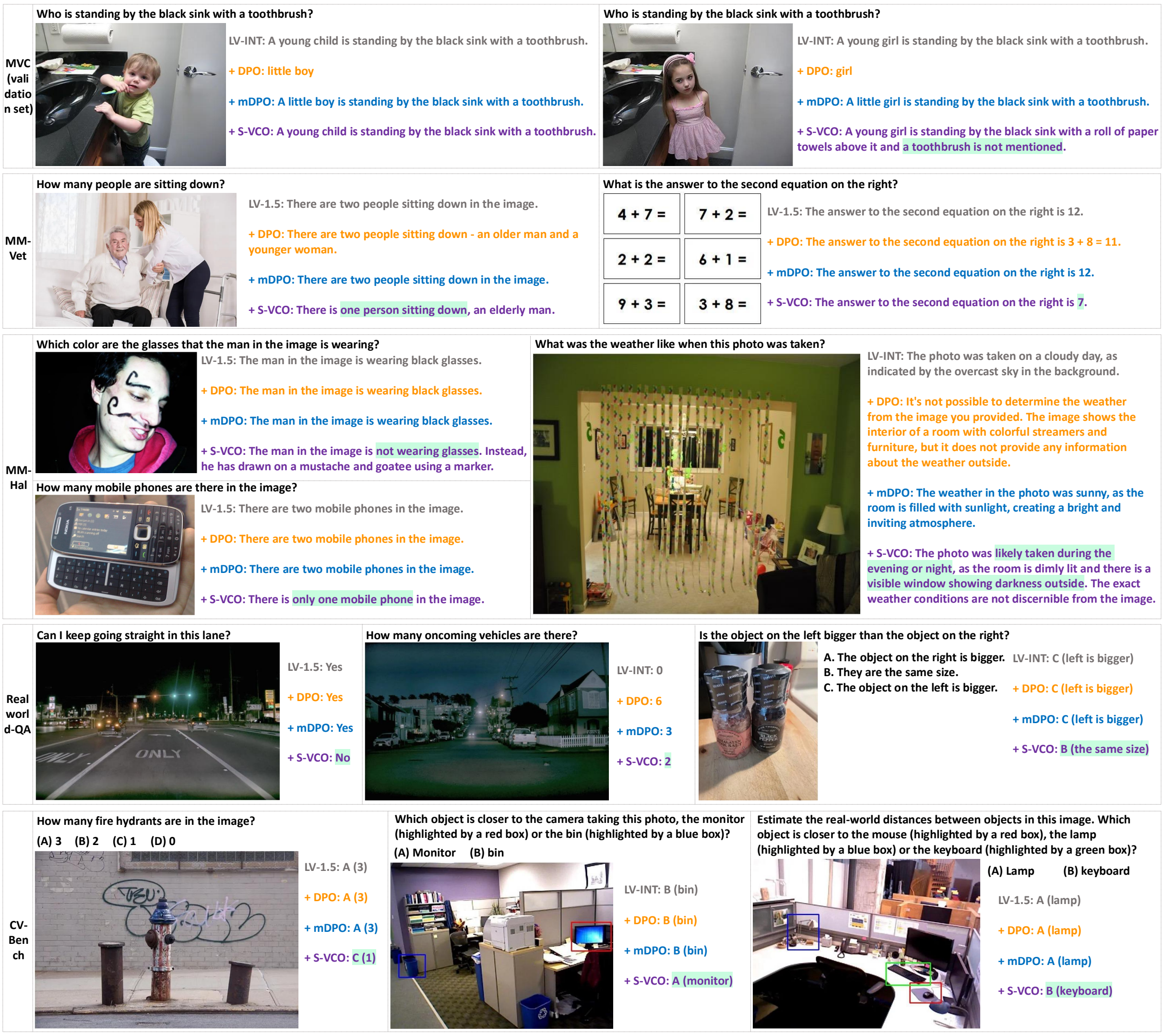
demonstrate S-VCO's superior ability to process fine-grained visual details and reason about complex scenes. S-VCO excels in recognizing subtle yet critical visual distinctions (*e.g.*, identifying the absence of a toothbrush) and remains robust against visual hallucinations (*e.g.*, differentiating marker-drawings, slide-phones, and fire hydrants). Moreover, S-VCO shows strong visual reasoning capabilities by interpreting complex scenarios such as drive-lane conditions. It also captures intricate details in scenes with greater depth and contextual awareness (*e.g.*, discerning weather conditions through a window or identifying oncoming vehicles in low-light settings).

MVC enhances previous preference tuning methods. Beyond its strong synergy with S-VCO, MVC dataset also strengthens the effects of existing preference tuning methods, particularly DPO. For LV-INT, both DPO and mDPO achieve greater overall improvements when trained on MVC compared to the textual-instruction-tuning-styled dataset VLF. For LV-1.5, DPO$_{\text{MVC}}$ outperforms DPO$_{\text{VLF}}$ by **~9%** on average across benchmarks, achieving second-best results on multiple metrics behind only our S-VCO. Fig. 6 further illustrates DPO and mDPO variants trained on MVC outperforming their VLF-trained counterparts, especially on visually dependent tasks. By providing visually challenging image pairs matched with texts, MVC proves to be a more effective training resource across preference tuning methods.

5.3 Ablations

Data filter and augmentation. We evaluate the impact of our filtering and augmentation step (§4.2) in MVC by training LV-INT-7B on unprocessed ("raw") visual counterfactual data directly sourced from CounterCurate (Zhang et al., 2024) and FineCops-Ref (Liu et al., 2024c). We sample $11,149$ random datapoints exactly matching MVC's size (Appx.A), and train the model with S-VCO using the same configurations. The resulting model, **S-VCO$_{\text{MVC-Raw}}$**, underperforms S-VCO$_{\text{MVC}}$ trained on filtered and augmented data, with drops across most benchmarks (Tab. 1). This highlights the importance of our filter and augmentation step in constructing high-quality contrastive data that better supports S-VCO. Notably, despite the data preprocessing omissions, S-VCO$_{\text{MVC-Raw}}$ still surpasses all other baseline preference tuning methods, demonstrating the robustness of S-VCO's symmetrical visual contrastive objective.

Symmetrical loss construction. A key feature of S-VCO is its symmetrical loss, which treats both sides of the texts y_w and y_l as preferred when aligned with their respective images i_w and



Who is standing by the black sink with a toothbrush?

LV-INT: A young child is standing by the black sink with a toothbrush.

+ DPO: little boy

+ mDPO: A little boy is standing by the black sink with a toothbrush.

+ S-VCO: A young child is standing by the black sink with a toothbrush.

Who is standing by the black sink with a toothbrush?

LV-INT: A young girl is standing by the black sink with a toothbrush.

+ DPO: girl

+ mDPO: A little girl is standing by the black sink with a toothbrush.

+ S-VCO: A young girl is standing by the black sink with a roll of paper towels above it and a toothbrush is not mentioned.

How many people are sitting down?

LV-1.5: There are two people sitting down in the image.

+ DPO: There are two people sitting down - an older man and a younger woman.

+ mDPO: There are two people sitting down in the image.

+ S-VCO: There is one person sitting down, an elderly man.

What is the answer to the second equation on the right?

LV-1.5: The answer to the second equation on the right is 12.

+ DPO: The answer to the second equation on the right is 3 + 8 = 11.

+ mDPO: The answer to the second equation on the right is 12.

+ S-VCO: The answer to the second equation on the right is 7.

Which color are the glasses that the man in the image is wearing?

LV-1.5: The man in the image is wearing black glasses.

+ DPO: The man in the image is wearing black glasses.

+ mDPO: The man in the image is wearing black glasses.

+ S-VCO: The man in the image is not wearing glasses. Instead, he has drawn on a mustache and goatee using a marker.

How many mobile phones are there in the image?

LV-1.5: There are two mobile phones in the image.

+ DPO: There are two mobile phones in the image.

+ mDPO: There are two mobile phones in the image.

+ S-VCO: There is only one mobile phone in the image.

What was the weather like when this photo was taken?

LV-INT: The photo was taken on a cloudy day, as indicated by the overcast sky in the background.

+ DPO: It's not possible to determine the weather from the image you provided. The image shows the interior of a room with colorful streamers and furniture, but it does not provide any information about the weather outside.

+ mDPO: The weather in the photo was sunny, as the room is filled with sunlight, creating a bright and inviting atmosphere.

+ S-VCO: The photo was likely taken during the evening or night, as the room is dimly lit and there is a visible window showing darkness outside. The exact weather conditions are not discernible from the image.

Can I keep going straight in this lane?

LV-1.5: Yes

+ DPO: Yes

+ mDPO: Yes

+ S-VCO: No

How many oncoming vehicles are there?

LV-INT: 0

+ DPO: 6

+ mDPO: 3

+ S-VCO: 2

Is the object on the left bigger than the object on the right?

A. The object on the right is bigger.
B. They are the same size.
C. The object on the left is bigger.

LV-INT: C (left is bigger)

+ DPO: C (left is bigger)

+ mDPO: C (left is bigger)

+ S-VCO: B (the same size)

How many fire hydrants are in the image?

(A) 3 (B) 2 (C) 1 (D) 0

LV-1.5: A (3)

+ DPO: A (3)

+ mDPO: A (3)

+ S-VCO: C (1)

Which object is closer to the camera taking this photo, the monitor (highlighted by a red box) or the bin (highlighted by a blue box)?

(A) Monitor (B) bin

LV-INT: B (bin)

+ DPO: B (bin)

+ mDPO: B (bin)

+ S-VCO: A (monitor)

Estimate the real-world distances between objects in this image. Which object is closer to the mouse (highlighted by a red box), the lamp (highlighted by a blue box) or the keyboard (highlighted by a green box)?

(A) Lamp (B) keyboard

LV-1.5: A (lamp)

+ DPO: A (lamp)

+ mDPO: A (lamp)

+ S-VCO: B (keyboard)

Figure 5: **Qualitative examples extracted from various benchmarks** comparing **base-VLMs** (**LV-INT** or **LV-1.5**) to the results after finetuning with **DPO**, **mDPO** or **S-VCO** on MVC dataset. Accurate captions of visual information are highlighted . Our method S-VCO demonstrates superior understanding of **fine-grained visual details** (*e.g.*, identifying the absence of a toothbrush) and shows **strong resilience to visual hallucinations** (*e.g.*, recognizing marker-drawings, fire hydrants, slide-phones). Furthermore, S-VCO excels in more advanced **visual reasoning** (*e.g.*, interpreting drive-lane conditions & regulations, estimating object sizes & distances), and captures **complex scenes** with greater detailedness and depth (*e.g.*, identifying weather through the window, recognizing oncoming vehicles in low-light settings).

i_l, optimizing both simultaneously (§3.2). To assess the necessity of this symmetry, we train LV-INT-7B on MVC without the symmetrical term $L_{\text{VCO}}(i_l, y_l, i_w)$, reducing the objective to a one-sided preference tuning approach. The resulting model, $\textbf{VCO}_{\textbf{MVC}}$, underperforms S-$\text{VCO}_{\text{MVC}}$, as shown in Tab. 1, though it still surpasses all other baseline methods due to VCO's strong visual contrastive supervision (§3.1). The performance drop is most evident in hallucination and vision-centric tasks, highlighting the importance of optimizing both sides of the contrastive pair. Symmetry mitigates shortcut learning by encouraging the model to focus on meaningful image-text alignments rather than superficial image features.

Comparing to SFT. To investigate whether standard supervised finetuning (SFT) on both sides of the image-text pairs could achieve similar results, we construct a new instruction-tuning dataset by including both (i_w, y_w) and (i_l, y_l) from MVC, effectively doubling its size (MVC × 2). Results for $\textbf{SFT}_{\textbf{MVC}\times\textbf{2}}$ are shown in the last row of Tab. 1. While SFT preserves performance on ScienceQA – a benchmark minimally reliant on visual inputs –

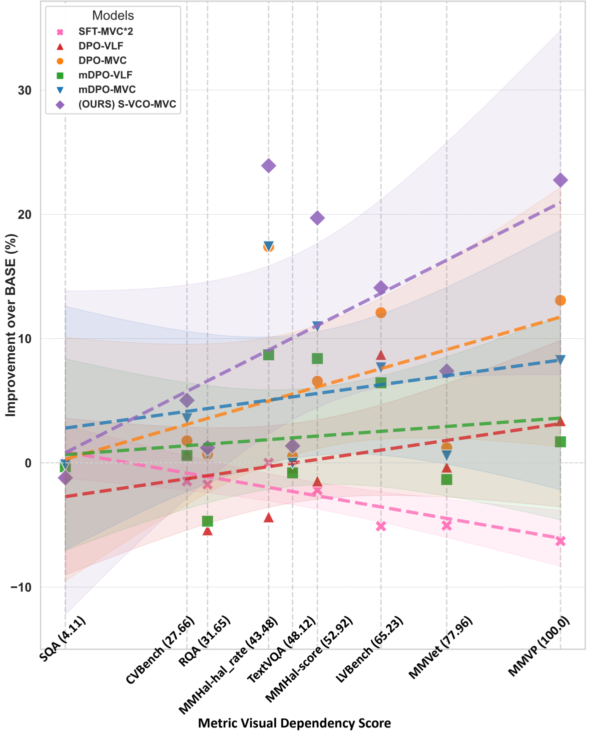


Figure 6: **Trend of improvement over the base-VLM as benchmarks become increasingly visually dependent.** A metric's visual dependency is measured as the performance drop of the base-VLM when no image input is provided. S-VCO exhibits the most significant trend of improvements with increasing task visual dependency, highlighting how its objective design (§3) enhances model's focus on critical visual details. MVC dataset also strengthens existing preference tuning methods (DPO and mDPO), while SFT (§5.3) degrades performance on more visually demanding benchmarks.

it performs significantly worse across all other domains, yielding the lowest results among all methods. Unlike S-VCO, SFT lacks the contrastive supervision necessary to highlight subtle visual-text alignment, thus leading to poor performance on tasks requiring strong visual grounding.

6 Related Work

VLM's Visual Hallucinations. Recent studies (Deng et al., 2024; Tong et al., 2024b; Chen et al., 2024; Jia et al., 2024) have shown that VLMs tend to hallucinate content not present in the visual input. VLMs also struggle with fine-grained visual understanding (*e.g.*, recognizing object attributes and relations) – especially when tested on confounding image pairs (Peng et al., 2024; Li et al., 2024a). This aligns with our findings in Fig. 2, suggesting a lack of robust and faithful multimodal grounding. To address this issue, several training-free methods have been proposed. Yang et al. (2023) introduced "Set-of-Mark" prompting that overlays spatial and textual markers on images to help models reference specific regions. Deng et al. (2024) employed CLIP-guided decoding to steer the language outputs with grounded visual cues. Architecture-wise, GRILL (Jin et al., 2023) incorporates object-level alignment during pretraining to promote visual grounding. Unlike previous approaches, our work focuses on finetuning with a novel objective (§3) and a data construction pipeline (§4) based on visual counterfactuals (Zhang et al., 2024; Liu et al., 2024c), targeting more precise alignment of multimodal details.

VLM Finetuning. Finetuning improves task-specific performance of VLMs and aligns them better with human preferences. SFT remains widely adopted to guide models toward towards following instructions (Sun et al., 2024; Jiang et al., 2024b). DPO (Rafailov et al., 2024) optimizes the margin between finetuned and unfinetuned model versions using paired preference data. Its extension to VLMs incorporates image as additional prefix condition (Zhou et al., 2024). Recent methods such as mDPO (Wang et al., 2024), MFPO (Jiang et al., 2024a), V-DPO (Xie et al., 2024), CHiP (Fu et al., 2025) and Image-DPO (Luo et al., 2024) further adapt the preference tuning paradigm to focus on image-side preferences over a pair of "good" and "bad" image, aiming to reduce visual hallucinations. Our approach S-VCO replaces the one-sided "preference" formulation with a stricter visual contrastive objective of symmetrical construct, treating "preference" explicitly as alignment over matching image-text pairs. This enables more comprehensive and robust VLM improvements across tasks.

7 Conclusion

This work introduces S-VCO, a novel VLM finetuning objective that enforces strict visual contrastive supervision within a symmetrical construct. Complementing this objective, we propose MVC, a dataset of paired images with minimal visual contrasts, each associated with corresponding contrastive texts. Experiments demonstrate that combining S-VCO with MVC consistently improves VLM performance across diverse benchmarks, with particularly significant gains on visually dependent tasks. Importantly, these improvements are achieved without compromising, and even enhancing VLMs' general capabilities.

Limitations

Our method incorporates multiple individual loss terms and weights (§3), which may require manual tuning to determine the optimal settings. Adjusting these hyperparameters could potentially further enhance performance.

The S-VCO objective works most effectively when the contrastive image pairs have meaningful differences in visual details. The current MVC derived from existing visual counterfactual data sources includes a limited set of operations for building the contrasts (§4.1). Our method could benefit from a more diverse set of contrasting visual details that should enable the model to potentially learn a broader range of visual features.

Ethics Statement

In this work, all data and pretrained models are publicly available. They are collected and processed in adherence to the respective data, checkpoints, and API usage policy. We acknowledge that our finetuned models may generate unsafe content, and we advise all users of careful verification before deploying this work in real-world applications.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *arXiv preprint arXiv:2403.18346*.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. 2025. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *arXiv preprint arXiv:2501.16629*.

Manu Gaur, Makarand Tapaswi, et al. 2024. Detect, describe, discriminate: Moving beyond vqa for mllm evaluation. *arXiv preprint arXiv:2409.15125*.

Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. *arXiv preprint arXiv:2411.11909*.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Yeying Jin, and Zuozhu Liu. 2024a. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*.

Xiaohu Jiang, Yixiao Ge, Yuying Ge, Dachuan Shi, Chun Yuan, and Ying Shan. 2024b. Supervised fine-tuning in turn improves visual foundation models. *arXiv preprint arXiv:2401.10222*.

Woojeong Jin, Subhabrata Mukherjee, Yu Cheng, Yelong Shen, Weizhu Chen, Ahmed Hassan Awadallah, Damien Jose, and Xiang Ren. 2023. Grill: Grounded vision-language pre-training via aligning text and image regions. *arXiv preprint arXiv:2305.14676*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024a. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023a. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. 2024c. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

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

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2024. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*.

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

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. 2024. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13279–13288.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*.

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

Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollar, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, and Ishan Misra. 2023. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5484–5494.

Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, Manling Li, Nick Haber, and Jiajun Wu. 2024. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. *arXiv preprint arXiv:2412.02193*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

xAI. 2024. Grok-1.5 vision preview. Accessed: 2024-12-12.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

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

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. 2024. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. *arXiv preprint arXiv:2402.13254*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. 2025. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer.

A Dataset Details

Tab. 2 details the dataset statistics of our visual counterfactual data sources CounterCurate (Zhang et al., 2024) and FineCops-Ref (Liu et al., 2024c), as well as our MVC after filtering and augmentation (disccused in §4.2). We discard the original "Order" category of FineCops-Ref for the frequent irrational cases under that category. For the non-synthetic "Left-Right" position flipping, we did not apply the filter but randomly sampled the data proportional to its original category distribution in the source datasets.

When training with MVC, we leave out 240 random samples for validation set (200 from CounterCurate and 40 from FineCops-Ref). This leads to a total training data size of $10,909$ with MVC. When training with the sampled VLFeedback data (Li et al., 2023a), used as in mDPO (Wang et al., 2024) (`https://huggingface.co/datasets/fwnlp/mDPO-preference-data`; noted as VLF), we leave out 200 random samples for validation set, leading to a total training data size of $9,222$ with VLF.

	CounterCurate	**FineCops-Ref**	**MVC**
Object Replacement	$26,164$	$4,171$	$7,189$
Attribute Replacement	$27,964$	$1,844$	
Count Modification	$10,010$	0	919
Position Change	$56,711$	$1,555$	$3,041$
Total	$120,849$	$7,570$	$11,149$

Table 2: **Statistics of visual counterfactual datasets** CounterCurate, FineCops-Ref, and our MVC after filtering and augmenting both data sources (§4). For MVC, the number of samples in the categories "Object Replacement" and "Attribute Replacement" are counted together.

B Implementation Details

Training. We set β_1 and β_2 in S-VCO's objective (Eq. 7) to 0.1, following the typical β values in DPO (Eq. 1) and mDPO (Eq. 2). All models are finetuned for 2 epochs with a batch size of 32 on 8 NVIDIA-A100x80G GPUs. When training on VLF, we retain the original learning rate of $1e{-}05$ used by DPO and mDPO. The text sequence length during finetuning is set to 1024 to accommodate VLF's text data length. When training on our MVC, we set a learning rate of $1e{-}06$ for all methods except mDPO on LV-1.5-7B, where a lower rate of $1e{-}07$ is used to better stabilize training. The text sequence length during finetuning is set to 128 given the MVC's shorter text data length.

Intermediate checkpoints are saved at an interval of 24, 31, and 62 steps for training on VLF, MVC, and MVC×2 (used for SFT ablation, see §5.3), respectively.

Evaluation. We report results using the best-performing checkpoint for each model configuration (Tab. 1), selected based on the highest average improvement over the base-VLM. For all benchmarks, the temperature of model predictions is set to 0. As the evaluation judge, we use `gpt-4-0613` for LLaVABench (Liu et al., 2024b) and MMVet (Yu et al., 2023), and `gpt-4-turbo` for MM-Hal (Sun et al., 2023).

C Prompt for MVC Language Augmentation

Below are our prompt templates for querying GPT4 (`gpt-4o`) to augment the queries and responses from visual counterfactual data sources (§4.2).

In the first step, we ask GPT4 to generate an natural and appropriate question that targets the subtle differences in the response-pair.

In the second step, we ask GPT4 to revise the instruction generated from the first step, and rephrase the original response pairs to make the whole instruction-response conversation sound more natural, while retaining the contrastive details in the responses.

In both steps, we set the temperature to 0.7 for more diversified wording.

GPT4 Language Augmentation Step 1

`[System]`
You are a helpful assistant that generates a natural-sounding instruction prompt for a vision-language scenario. Given two responses about an image: one 'chosen' and one 'rejected', your task is to produce a single instruction or question that encourages the user to naturally reveal the critical differences between the two responses. Focus on attributes that differ (like number, color, position, orientation). The prompt should sound like a normal request someone might ask when wanting more detail about the image. It should not sound overly forced or contrived, and it should not explicitly mention that there are two responses or that differences are being tested. Also, try to vary your phrasing, and do not always start the instruction with 'Could' or 'Can'.

`[User]`
Chosen response: `{ORIGINAL_RESPONSE}`
Rejected response: `{CONTRAST_RESPONSE}`

Generate a single, natural-sounding instruction or question that would prompt a user or model to include the detail of the collar in a natural way. Avoid making the prompt sound forced or unnatural, and do not explicitly mention comparing two descriptions.

GPT4 Language Augmentation Step 2

[System]
You are a helpful assistant that revises instruction prompts and their corresponding responses for vision-language scenarios. Given an initial instruction and two responses about an image: one ‘chosen’ and one ‘rejected’, your task is to:
1. Revise the instruction to make it sound more natural and conversational and ensure it seamlessly leads to the given responses. Also, try to vary your phrasing, and do not always start the instruction with ‘Could’ or ‘Can’.
2. Rephrase both the chosen and rejected responses to diversify the language without adding or removing any details.
3. Ensure that the differences between the chosen and rejected responses remain highlighted and are consistent with the original responses.

Do not introduce any new information or omit existing details. The revisions should maintain accuracy and ensure coherence between the instruction and responses.

[User]
Initial Instruction: {STEP1_GENERATED}
Chosen Response: {ORIGINAL_RESPONSE}
Rejected Response: {CONTRAST_RESPONSE}

Revise the instruction and both responses as described above.

Here are some examples:

1.
Initial Instruction: "What can you tell me about the hair color of the woman who is sweeping the floor?"
Chosen Response: "A blonde-haired woman wearing a white skirt, white shirt, white apron, and black shoes is sweeping the floor."
Rejected Response: "A black-haired woman wearing a white skirt, white shirt, white apron, and black shoes is sweeping the floor."
Revised Instruction: "Describe the woman’s hair color and her attire while she’s sweeping the floor?"
Revised Chosen Response: "The woman sweeping the floor has blonde hair and is wearing a white skirt, white shirt, white apron, and black shoes."
Revised Rejected Response: "The woman sweeping the floor has black hair and is wearing a white skirt, white shirt, white apron, and black shoes."

2.
Initial Instruction: "Where is the air stunt relative to the snowy mound in the image?"
Chosen Response: "An air stunt is above a snowy mound."
Rejected Response: "An air stunt is below a snowy mound."
Revised Instruction: "What do you notice about the position of the air stunt in relation to the snowy mound in the image?"
Revised Chosen Response: "The air stunt is positioned above the snowy mound."
Revised Rejected Response: "The air stunt is located below the snowy mound."

Now, revise the following instruction and responses:
Initial Instruction: {STEP1_GENERATED}
Chosen Response: {ORIGINAL_RESPONSE}
Rejected Response: {CONTRAST_RESPONSE}

Reply ONLY in the following format and no other text or notes:

Revised Instruction:
Revised Chosen Response:
Revised Rejected Response: