



Self-Supervised Visual Preference Alignment

Ke Zhu^{1,2} Liang Zhao⁴ Zheng Ge^{3,4} Xiangyu Zhang^{3,4}

¹State Key Laboratory for Novel Software Technology, Nanjing University, China

²School of Artificial Intelligence, Nanjing University, China

³MEGVII Technology

⁴StepFun Intelligent Technology

zhuk@lamda.nju.edu.cn, {zhaoliang06, gezheng, zhangxiangyu}@megvii.com

Abstract

This paper makes the first attempt towards unsupervised preference alignment in Vision-Language Models (VLMs). We generate chosen and rejected responses with regard to the original and augmented image pairs, and conduct preference alignment with direct preference optimization. It is based on a core idea: properly designed augmentation to the image input will induce VLM to generate false but hard negative responses, which helps the model to learn from and produce more robust and powerful answers. The whole pipeline no longer hinges on supervision from GPT-4 or human involvement during alignment, and is highly efficient with few lines of code. With only 8k randomly sampled unsupervised data, it achieves 90% relative score to GPT-4 on complex reasoning in LLaVA-Bench, and improves LLaVA-7B/13B by 6.7%/5.6% score on complex multi-modal benchmark MM-Vet. Visualizations shows its improved ability to align with user-intentions. A series of ablations are firmly conducted to reveal the latent mechanism of the approach, which also indicates its potential towards further scaling. Code are available in <https://github.com/Kevinz-code/SeVa>.

1. Introduction

Large Vision-Language Models have recently emerged and greatly advanced current researches [2, 25, 34, 40]. Representative methods in this family, such as MiniGPT4 [46], LLaVA [25] and InstructBLIP [8], try to properly align vision signals to Large Language Models (LLMs) to better conduct multi-modal comprehension. These methods usually undergo a pretraining stage with abundant image-text pairs for modality alignment before supervised fine-tuned (SFT) with academic [8] or GPT-4 generated [4, 42] instruction following data. There are variants of them

(e.g., LLaVA-based) that try to improve the VLMs’ instruction following ability by leveraging high quality pretraining pairs [4, 23] or scaling up SFT database [22, 36, 42].

Despite their success [4] in boosting the comprehension skills of current VLMs, these models are not properly aligned with user-intentions. As a result, they lack the ability to reject samples and could induce more unintended output. For instance, Instruct4V [36] construct 220k SFT data from GPT-4, but still struggle on hallucination benchmarks POPE [20]. The same can be observed in Fig. 6, where LLaVA failed to follow user instructions and provide meaningless information in its answers. Recently, there are few trials that incorporate alignment techniques into vision-language fields [19, 35, 44]. However, we found that they only emphasize on quite limited task domains [44], and, more importantly, *their data construction pipeline requires extra knowledge source either from GPT-4 or human feedbacks*. This might hinder them from further data scaling, as preference data is *not cheap* (e.g., a 10k collected human-evaluated instances requires a cost of 3000\$ [21]).

In this paper, we propose a self-supervised paradigm that can *freely* obtain arbitrary number of preference samples with minimal code implementation (c.f., Alg. 1). Simply but, we have enjoyed the advantage of this pipeline and observed a significant improvement in the capabilities of current VLMs from various aspects: stronger chain-of-thought skills, better OCR [28] ability, proper alignment with user-intentions, less hallucinations, etc (*cf.* Fig. 6).

Our motivation came from an interesting discovery in Fig. 2, where we apply commonly adopted image augmentations in *visual contrastive learning* [14, 47, 48] during LLaVA inference. The effect of all these test-time augmentations are evaluated on three common multi-modal benchmarks (MMVet [39], POPE [20], MMBench [26]). It is not surprising to observe in Fig. 2 that *vision-language models are quite sensitive to data-augmentations* and, slightly

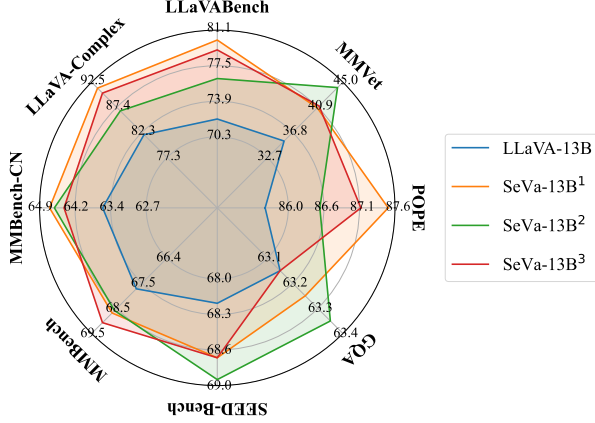


Figure 1. Illustration of the baseline LLaVA-13B (v1.5) and the proposed SeVa-13B. Here we demonstrate three variants of SeVa with different sampled seed to obtain the un-labeled dataset (the image-text pairs used for DPO sample generation, cf. Alg. 1).

Algorithm 1 Pseudocode of SeVa in a PyTorch style.

```
# Q, I: question set, image set
# model: the SFT vision-language model
# C, R: chosen and rejected answer set
# T: sampled data augmentation

# randomly sample data pair, generate answer
Q, I = random.sample(data)
C, R = model(I, Q), model(T(I), Q)

# filtering equal answers
idx = (C != R)
C_t, R_t = C[idx], R[idx]
Q_t, I_t = Q[idx], I[idx]

# DPO training, omit reference model for simplicity
DPOTrain(model, (Q_t, I_t), (C_t, R_t))
```

distortion will induce the model to output different semantic tokens. Then, a natural question arises: *are the original and distorted responses valuable to construct preference data for DPO training?* Since this will totally free the data construction from *any* extra source, and might relieve the difficulty of scaling up preference data as well [41].

Our motivation was then firmly verified by quantitative experiments in Table 1, in which we randomly sampled 8k image-question pairs from the subset of LLaVA665k [24] (cf. Sec. 4.1 for details), and choose 5 augmentations to generate the preference data using LLaVA-7B. All these self-generated data are again fed into the *same* LLaVA-7B model for DPO training. As Table 1 shows, all the augmentations are helpful for model comprehension, especially on GPT-4 evaluated benchmark MM-Vet, which makes our motivation valid. In the meanwhile, we found that either too strong (e.g., Diffusion-S) or too weak data augmentation (e.g., RandFlip) are sub-optimal for the whole pipeline, and medium is the best (e.g., Diffusion-W, MOCO). Our final conjecture is: *self-generated augmentation pairs are*

suitable preference sample to improve multi-modal abilities, and hard negatives are most preferred.

We name our method: **Self-supervised Visual preference alignment** (SeVa), and summarize its whole pipeline in Alg. 1. Though being embarrassingly simple with few lines of code to implement, we mathematically showcase its strong relation with visual contrastive learning in maximizing a preference distribution, where SeVa could probably be viewed as a special form of contrastive learning with one negative sample. This makes SeVa easily extendable if more negatives are involved (cf. appendix for details).

Experiments in common multi-modal benchmarks demonstrate the effectiveness of our methods, where SeVa improves the VLM’s comprehension ability by a large margin (e.g., SeVa-7B even surpasses LLaVA-1.5-13B by 1.8% in MMVet). In addition, we carefully design detailed ablations to reveal the latent mechanism of SeVa from different angles. We found that SeVa shows surprisingly results like detailed descriptions, less hallucinations, stronger OCR skills and chain-of-thought ability, etc.

In summary, our contributions are:

- For the first time, we conduct visual preference alignment in an un-supervised manner. The whole pipeline *do not* require any GPT-4 or costly human annotated data.
- We found such an alignment pipeline have numerous benefits such as enhanced multi-modal comprehension ability, better alignment with user-intentions, etc. Visualizations and ablations firmly verify our hypothesis as well.
- Our method, SeVa, enjoys efficiency in pipeline and simplicity in implementation, which paves way for future preference alignment in visual-language domain.

2. Related Works

Large Vision-Language Models (VLMs). The great development of Large Language Models (LLMs) has facilitated the community in both academics [8, 9, 37, 38, 46] and industries [1, 7, 34]. Recently, large vision-language models emerged, aiming to extend the reasoning brain of LLM to vision modality [43]. The majority of VLMs undergo a two-stage training manner, with the pretraining period focusing on aligning the vision and text signals before finetuned with instruction following data in the second stage. LLaVA [25], as one of its representatives, has attracted numerous researchers [4, 22, 23, 36] since it provide a valuable opportunity for them to reproduce and built upon. Subsequent works based on LLaVA try to leverage more image-text data pairs of higher quality [4] or to construct more abundant instruction following database [36]. Despite their achievements, these model are not preference aligned with user intentions neither implicitly nor explicitly, which might limit their further deployment. Therefore, an suitable alignment technique is of great importance.

Preference alignment in LLM/VLM. Training an LLM

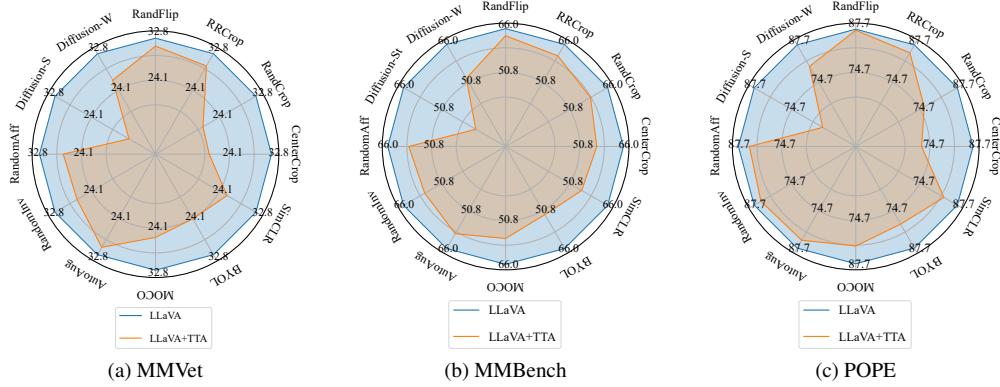


Figure 2. The test-time image augmentations (TTA) plugged into LLaVA-1.5 on three benchmarks. We involve standard augmentation: RandFlip, RandomResizedCrop (‘RRCrop’), RandomCrop, CenterCrop, RandomAffine, RandomInvert and AutoAug; diffusion noise augmentation: Diffusion-Weak (‘W’) and Diffusion-Strong (‘S’); mixtures: strategies adopted in MOCO [14], BYOL [12] and SimCLR [5].

DPO data	MMVet							MMBench		POPE			
	All	rec	ocr	know	gen	spat	math	en	cn	All	rand	pop	adv
N/A	30.5	35.7	21.9	17.4	19.7	24.7	7.7	64.3	58.3	85.9	89.5	86.7	81.7
RandFlip	33.7	37.2	26.4	21.8	23.9	29.1	7.7	64.4	58.3	86.7	89.2	87.1	83.9
RRCrop	33.8	39.2	23.2	21.9	24.5	27.7	7.7	65.5	59.5	86.8	89.4	87.4	83.6
AutoAug	35.0	38.5	26.8	22.9	24.4	29.3	15.0	65.4	59.6	86.5	88.8	86.8	83.9
Diffusion-W	37.2	40.2	29.9	21.8	23.9	34.3	11.2	65.6	59.2	86.7	89.4	87.1	83.6
MOCO	37.0	40.9	29.0	21.3	25.6	32.5	15.0	65.2	59.8	86.6	89.1	87.1	83.7
Diffusion-S	34.6	38.8	26.5	20.5	23.4	32.0	11.5	65.2	58.2	86.6	89.2	87.5	83.3

Table 1. Six data augmentations selected to generate preference data for DPO training (w/ LLaVA-1.5). Each strategy was adopted to distort the image (together with the questions) to produce rejected response, which is paired with the chosen response from the original image input (*cf.* Alg. 1). The models after DPO training with these preference pairs are then evaluated on MMVet [39], MMBench [26] (in english ‘en’ and chinese ‘cn’) and POPE [20] benchmarks, respectively. Please refer to Sec. 3.2 and visualizations in appendix.

to align with human or user preference is called RLHF [29], which usually happens after the instruction following stage. The core concept of RLHF is to reduce un-intended or toxic output produced by LLMs [11, 49]. Variants of RLHF include, but not limited to, DPO [31], PPO [32] and RLAIIF [3]. Alignment has been fruitfully researched in natural language processing (NLP) domains [3, 13], but relatively less visited in vision-language fields. There are some latest attempts [19, 35, 44] that integrate preference alignment with DPO in vision-language domains. However, they only emphasize on task specific ability (e.g. hallucinations in HA-DPO [44]), and, more importantly, it requires tedious GPT-4 or human interventions to construct the preference data, which may diminish the pipeline efficiency as well. In this paper, we propose an unsupervised data construction pipeline that not only solve the data hungry issue in preference alignment, but also greatly boosts the comprehension ability of current VLMs, as well.

Contrastive learning. Self-supervised learning (SSL) of visual representation are popularized in the past years [5, 12, 14, 47]. Traditional SSL are mostly contrastive based,

where strongly augmented positive views of the same image *will* share similar deep semantic in the hidden space that the model could capture. Nevertheless, we found that a similar augmentation pipeline *does not* holds true in VLMs. Inspired by unsupervised property of SSL, we design an self-supervised pipeline to construct the preference data, and empower the current VLM with improved capability.

3. Method

We will first show the preliminaries of language modeling and direct preference optimization, then move onto the procedure of the proposed SeVa pipeline. Finally, we discuss its strong relation with visual contrastive learning.

3.1. Preliminaries

Language modeling. We start from the vision-language modeling perspective and take LLaVA [25] as an example for illustration. Given an image input I , it is first passed through a vision encoder (e.g., ViT-L [30]) to get:

$$\mathbf{h} = g(I), \quad (1)$$

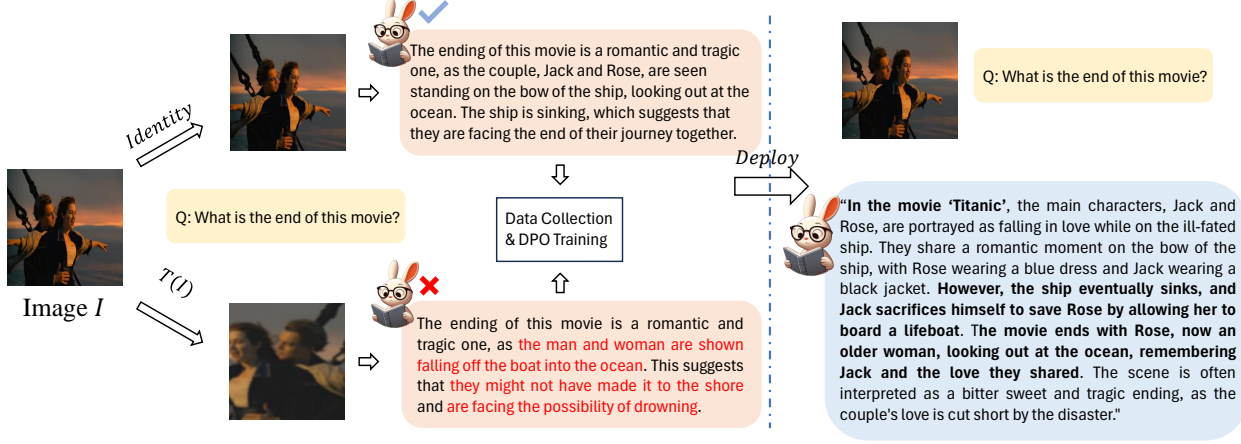


Figure 3. The pipeline of SeVa. For each image I in the selected dataset, we transform it with data augmentation T to obtain the distorted one, while keeping a copy of the original image to form a pair. The shared questions are acted on the paired images to get the chosen and rejected responses, respectively, which undergo a data collection (e.g., filtering) process before the DPO training. In the left part, incorrect words or sentences are red color coded, while in the right part (the improved version of the model), we highlight excellent content with bold phase. Note that in the picture, we show the *same* image both for training and for testing, but actually the data distribution between them are different (cf. Sec. 4.1). This figure is best viewed in color.

where g denotes a combination of vision encoder and projection layer. \mathbf{h} represents a sequence of deep image embeddings. The embeddings before the last transformer layer are considered in LLaVA-1.5 [24], which are then concatenated with the question token q and are fed into the autoregressive LLM π that sequentially generates the next target token y :

$$\pi_{\theta}(y|x) = \prod_{i=1}^L \pi_{\theta}(y_i|y_{<i}, x), \quad (2)$$

where $x = (\mathbf{h}, q)$, and is considered as prompt sent to the LLM parameterized by θ . $y_{<i}$ means the generated token before the current prediction y_i , and L is the length of the produced token sequence. We denote the LLM trained with SFT data as π_{SFT} for clarity.

Direct preference optimization. DPO is first proposed in [31], and can be viewed a new parameterization of the reward model in RLHF [29] that can directly optimize the policy (the LLM’s parameter θ). Here we start with RLHF for better illustrations. In RLHF, a Bradley-Terry (BT) reward model is often adopted, which stipulates the human preference distribution as:

$$p^*(y_c \succ y_r | x) = \frac{\exp(r^*(x, y_c))}{\exp(r^*(x, y_c)) + \exp(r^*(x, y_r))}. \quad (3)$$

The y_c and y_r are the chosen and rejected response conditioned on the LLM’s prompt x . The reward model $r_{\phi}(\cdot)$ is parameterized by ϕ and can be obtained with maximum likelihood using preference database \mathcal{D} annotated by human [31]:

$$\mathcal{D} = \{x^{(j)}, y_c^{(j)}, y_r^{(j)}\}_{j=1}^N. \quad (4)$$

Then we maximize the a preference policy as follows:

$$\max_{\theta'} \mathbb{E}_{x,y} \{r_{\phi}(x, y) - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta'}(y|x) | \pi_{\text{ref}}(y|x)]\}. \quad (5)$$

The reference models π_{ref} are usually initialized by π_{SFT} , which prevents the learned parameter θ' from too much deviation. The finally learned policy $\pi_{\theta'}(\cdot)$ are better equipped with human or user intentions.

DPO uses a closed form [31] derived from Eq. 5 to represent the optimal reward r^* model by the learned optimal preference model π^* as (with K the constant factor):

$$r^*(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + K. \quad (6)$$

By substituting the r^* in Eq. 3 with Eq. 6, we get the final optimized loss function of DPO as follows (cf. [31]):

$$\mathcal{L}_d = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta'}(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \beta \log \frac{\pi_{\theta'}(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right) \right] \quad (7)$$

The trainable parameter Eq. 7 is θ' , which is the same with RLHF. But it do not need any online optimization process like PPO [32]. Instead, DPO samples (x, y_c, y_r) from the preference database \mathcal{D} and optimize the VLM’s parameters θ' with a simple classification loss.

3.2. Visual preference alignment

This section will introduce our pipeline: self-supervised visual preference alignment. For a language model π_{SFT} after supervised finetuning stage, we feed it with the original and

augmented image input as (for simplicity, we omit the generated token as conditional input shown in Eq. 2):

$$y_c^{(j)} = \pi_{\text{SFT}}(g(I^j), q^j), \quad (8)$$

$$y_r^{(j)} = \pi_{\text{SFT}}(g(\mathcal{T}(I^j)), q^j). \quad (9)$$

The $y_c^{(j)}$ and $y_r^{(j)}$ are regarded as a pair of raw chosen and rejected response. For each image I^j and its question q^j , we generate only one paired preference data $(y_c^{(j)}, y_r^{(j)})$. Suppose there are total M image-question pairs and its corresponding generated preference pairs, we simply filter those equaled responses to get remained preference data pairs (supposed N_d), which constitute the final preference database $\mathcal{D}_{\text{SeVa}}$ for DPO training (cf. Alg. 1 and Eq. 7).

$$\mathcal{D}_{\text{SeVa}} = \{I^j, q^j, y_c^{(j)}, y_r^{(j)}\}_{j=1}^{N_d}, \quad (10)$$

$$\text{s.t. } \forall j, y_r^{(j)} \neq y_c^{(j)}. \quad (11)$$

We choose 6 different data augmentation strategies to construct our database and conduct DPO training. As shown in Table 1, all self-constructed preference data are helpful for multi-modal comprehension, showing the validity of this pipeline. What surprises us is that training samples built from ‘RandFlip’ are quite helpful (e.g., an increase of 3.2% score on MMVet). We conjecture that this might derive from the natural property of the unsupervised textdata, since flipped images are can make OCR and recognition harder and induce potential negative responses (cf. appendix for more details). Besides, all models have seen a noticeable improvement in adversarial settings of POPE, indicating the models trained by our self-supervised pipeline might be less inferior to potential interruptions.

3.3. Connection with contrastive learning

In contrastive learning, augmented views of the same image will share similar semantic in its deep embeddings, which optimize the following InfoNCE loss:

$$\mathcal{L}_{\text{in}} = -\log \frac{\exp(q \cdot k_+ / \tau)}{\exp(q \cdot k_+ / \tau) + \sum_i^n \exp(q \cdot k_-^i / \tau)}, \quad (12)$$

in which q, k_+ are the positive embeddings from augmented views of the same image, while k_- could be negatives embeddings from other images [5] or memory bank [14]. If we consider only one negative, and denote $f(q, k) = (q \cdot k) / \tau$ as the scoring function, we can reformulate Eq. 12 as:

$$\mathcal{L}'_{\text{in}} = -\log \frac{\exp(f(q, k_+))}{\exp(f(q, k_+)) + \exp(f(q, k_-))}. \quad (13)$$

It is not hard to find that the optimized loss function derived from Eq. 3 are quite similar with Eq. 13. The main difference lies in the *definition of negative samples*. Unlike contrastive learning, the predicted tokens from the augmented

views of the same image will be counted as negatives, but not positives. If we imitate traditional InfoNCE loss in Eq. 12, we could inject multiple negative reward terms in Eq. 3 (by augmenting multiple views of the same image to produce multiple responses), and thus derive a more general form of DPO. Please refer to appendix for more details, and we will leave this as future work.

4. Experiment

In this section, we conduct experiments to verify the effectiveness of the proposed SeVa pipeline. Firstly, we provide the experimental settings (e.g., the data construction process). Then we show the results on multi-modal comprehension benchmarks. Finally, we provide fruitful ablations and visualizations to analyze SeVa with more details. For simplicity, we name *LLaVA-1.5-7B/13B trained with SeVa pipeline as SeVa-7B, SeVa-13B respectively*.

4.1. Settings

Data construction. The source data we obtained are from LLaVA665k SFT dataset [24], and we choose image-question pairs from TextVQA [33] and OCRVQA [28] (denoted as ‘text+ocr’ for simplicity) in LLaVA665k to generate the DPO preference data. *This setting is kept by default throughout all the experiment unless otherwise noted.* For each data instance in ‘text+ocr’ of LLaVA665k, we randomly choose 2 questions in its multi-turn dialogue, and pair each question with the image. We treat each image-question pair as *one data instance*, which leads to nearly doubled instance as the original ‘text+ocr’. Then we randomly choose 8k such image-question pairs in ‘text’ and ‘ocr’ each, to form the original data source of 16k. Finally, these 16k image-question data instances will go through SeVa pipeline to produce DPO preference data with filtering (cf. Alg. 1), which are cut down by half to about 8k. For clarity in our later ablations, we denote training data num as those unfiltered preference pairs instead of the filtered ones (e.g., 16k but not 8k), since the raw unfiltered data is fixed and immune to filtering process (cf. appendix).

Data augmentation choice. As verified in Table 1, all the selected data augmentations are all helpful for multi-modal comprehension. In our later experiment, we choose *diffusion noise* as the default augmentation in SeVa pipeline, since it can easily control the intensity of distorted level, which is helpful for ablations (cf. Fig. 7). The noise steps for training SeVa-7B and SeVa-13B are set as 800 and 500, respectively. As MOCO augmentation strategy is also highly effective as shown in Table 1, we are glad to verify more data augmentation strategies in our future work.

Training details. Following previous works [4, 21, 22, 36], we choose LLaVA-1.5-7B/13B as our base models in all our experiments. The LLaVA’s weights are pretrained and SFT tuned before our DPO training, which is the by

Method	Language model	MMVet	LLaVA ^W	MMB	MMB ^{CN}	POPE	SEED ^I	SHR (\downarrow)	SQA	GQA
BLIP-2	FLAN-T5	22.4	38.1	—	—	85.3	46.4	—	61.0	41.0
InstructBLIP	Vicuna-7B	26.2	60.9	36.0	23.7	—	53.4	—	60.5	—
InstructBLIP	Vicuna-13B	25.6	58.2	—	—	78.9	—	51.2	63.1	49.5
Shikra	Vicuna-13B	—	—	58.8	—	—	—	—	—	—
IDEFICS-9B	LLaMA-7B	—	—	48.2	25.2	—	—	—	—	38.4
IDEFICS-80B	LLaMA-65B	—	—	54.5	38.1	—	—	—	—	45.2
Qwen-VL	Quen-7B	—	—	38.2	7.4	—	56.3	—	67.1	59.3
Qwen-VL-chat	Quen-7B	—	—	60.6	56.7	—	58.2	—	68.2	57.5
LLaVA	Vicuna-7B	26.7	63.0	34.1	14.1	—	25.5	—	38.5	—
LLaVA-1.5	Vicuna-7B	30.5	63.4	64.3	58.3	85.9	65.7	36.7	66.8	62.0
SeVa-7B	Vicuna-7B	37.2	72.2	65.6	59.2	86.7	65.8	34.9	67.5	<u>60.7</u>
LLaVA-1.5	Vicuna-13B	35.4	70.7	67.7	63.6	85.9	68.2	37.2	71.6	63.3
SeVa-13B	Vicuna-13B	41.0	80.1	68.7	64.8	87.4	68.6	36.6	<u>71.2</u>	63.4

Table 2. Comparison with state-of-the-art methods on 9 benchmark datasets. SeVa consistently improves LLaVA-1.5-7B/13B on 8 out of 9 benchmarks, namely, MMVet [39], LLaVA-bench-in-the-wild [25], MMBench [26], MMBench-Chinese [26], POPE [20], SEED-Image [18], SHR [44], SQA [16], GQA [16]. Among them, SHR are newly proposed [44] to evaluate the hallucination ratio of VLMs.

Method	Unsupervised	data source	data scale	MMVet	MMB	POPE	Training cost
LLaVA-1.5	—	—	—	30.5	64.3	85.9	—
cont. SFT (2%)	X	LLaVA665k	15k	31.9	63.7	86.0	0.1h
cont. SFT (10%)	X	LLaVA665k	66k	32.8	64.9	86.0	1.2h
cont. SFT (20%)	X	LLaVA665k	132k	33.9	64.2	86.1	2.4h
cont. SFT	X	text+ocr	102k	32.5	65.2	86.7	1.2h
SeVa (<i>ours</i>)	✓	text+ocr	8k*	34.8	65.3	86.2	0.3h
SeVa (<i>ours</i>)	✓	text+ocr	16k*	37.2	65.6	86.7	0.7h

Table 3. Comparison between SeVa with continual (‘cont.’) SFT in a 7B setting. We randomly select 15k (2%), 66k (10%) and 132k (20%) out of LLaVA665k as SFT data. ‘Unsupervised’ means no answering labels are required. We also sample *all* the TextVQA and OCRVQA data in LLaVA665k for comparison (since SeVa utilize its unsupervised data version). * means the unfiltered preference instances (the actually kept data for DPO training are mostly by half, cf. Sec. 4.1). Training cost are all evaluated in a same experiment settings.

default pipeline as in many previous literatures [13, 31, 32, 44]. We adopt deepspeed ZeRO stage-3 [24] during DPO training and use Vicuna-7B/13B, CLIP-ViT-L-336px as our LLM and vision encoder, respectively. The total epoch, batch size, learning rate and weight decay are set as 1, 128, 2e-6 and 0, respectively, following previous work [44]. The hyper-parameters of lora r are set as 1024, and the scale parameter β in DPO is fixed as 0.1. DPO optimization are conducted with 8-A800 GPU, with LLaVA-1.5-7B/13B tuned for about 0.7/1.3 hours.

4.2. Compare with state-of-the-art

Firstly, we compare the proposed SeVa pipeline with state-of-the-art VLMs. A total of 9 benchmarks are included, with multi-modal benchmark as: MMVet [39], LLaVA-Bench-in-the-wild [24] (LLaVA^W), MMBench [26] (MMB), MMBench-Chinese (MMB^{CN}), POPE [20], SEED-Bench-Image [18] (SEED^I) and SHR [44], and traditional Question-Answer benchmarks ScienceQA [27] (SQA) and

GQA [16]. Among them, MMVet, LLaVA^W are two GPT-4 evaluated benchmarks. MMB and MMB^{CN} are multiple choice question answer dataset that requires minimal GPT-4 involvement (e.g., answer choice aligning [26]). POPE and SHR are two hallucination benchmarks, and SQA & GQA are two traditional QA benchmarks.

As shown in Table 2, SeVa consistently improve the LLaVA-1.5 7B/13B models on 8 out of 9 benchmark datasets. Specifically, it improve LLaVA-v1.5-7B/13B models by 6.7% and 5.6% on the most complex multi-modal dataset MMVet, showing the great power the proposed pipeline. The same can be observed in LLaVA^W, where a 13B SeVa model achieves a 80% relatively score to GPT-4, boosting LLaVA-1.5-13B by a large margin of 9.4%. The great boost on GPT-4 involved datasets suggested better alignment with user-intentions, as demonstrated in Fig. 5 and Fig. 6. We also observe a consistent improve on benchmarks POPE and SHR, which indicates the SeVa pipeline is also helpful for relieving object hal-

Unsupervised data	Data Scale			
	2k	4k	8k	16k
TextVQA [33]	31.8	32.1	34.8	35.8
OCRVQA [28]	32.1	32.3	32.8	34.5
GQA [16]	31.4	31.8	34.1	35.9
COCO [14]	31.7	31.2	34.2	34.4
Visual Genome [24]	31.4	32.0	33.4	34.1

Table 4. MMVet evaluation score with different data (image-question) pairs and scale to conduct the SeVa-7B pipeline. The highest score in each dataset are highlighted with bold symbol. In SeVa pipeline, we choose TextVQA+OCRVQA by default. All the data listed are sourced from LLaVA665k [24].

temp (t)	Q-Consistency (\uparrow)		A-Consistency (\uparrow)	
	LLaVA	SeVa	LLaVA	SeVa
0.2	7.30	7.75	6.45	6.95
0.4	7.11	7.98	6.28	7.32
0.5	7.39	7.93	6.52	7.29
0.7	7.81	8.18	7.10	7.64
0.9	8.43	8.53	7.87	7.98
50	7.75	8.47	6.94	8.03

Table 5. GPT-4 evaluated consistency score from 1-10. Q-consistency measures how the model’s answers are aligned with the asked question, while A-consistency measures how answers ($t > 0$) are aligned with its generated tokens w/o sampling ($t = 0$).

lucinations, and thus we *may not* need specialized hallucination methods [17, 44] to handle them. Please note that the improvement on hallucinations are *not marginal*, since previous work that constructed huge amount of SFT data can lead to *decreased performance* on POPE (e.g., LVIS-Instruct4V [36]). There is also a higher multiple choices accuracy in MMB, MMB^{CN} and SEED^I, showing the broad adaptability of SeVa. In Table 2, a little bit drop was found on traditional benchmarks SQA/GQA. We conjecture that this phenomenon might arise from a trade-off between stronger instruction following ability in traditional QA and better comprehension ability in modern VLMs benchmarks, which is also pointed out in [13] that stronger instruction following do not guarantee a more helpful and better VLMs. In fact, how to better trade-off between benchmarks remains a discussing issue [4] and is out-of-scope of this paper.

4.3. Ablations

Now we move onto ablating the factors in the proposed pipeline, to help readers better understand the success behind SeVa and its generalized ability. We choose SeVa-7B as the default settings unless otherwise noted.

Compare with SFT We start from comparing with continual SFT, as this is the most straight forward method to verify its effectiveness with supervised training (note SeVa

Representative questions in each dataset:

Question textvqa:

- Provide a one-sentence caption for the provided image.
- Reference OCR token: [task-specific prompt]

Question ocrvqa:

- Who is the author of this book?
- What is the title of this book?

Question gqa:

- Is the cheese to the left of the food on the plate?
- Which kind of furniture are the cups hanging on?

Question coco:

- Where is the cat positioned in the image?
- Please provide a short description for this region: [task-specific prompt]

Question vg:

- Please provide a short description for this region: [task-specific prompt]
- Provide the bounding box coordinate of the region this sentence describes. [task-specific prompt]

Figure 4. Illustration of representative questions in five different database from LLaVA665k [24]. In our main experiment, we adopt a combination of ‘textvqa’ and ‘ocrvqa’. The results of applying the other 3 question types in SeVa can be found in Table 4.

do not need target answer labels). We conduct three different types of SFT: a random selection of 2%/10%/20% in LLaVA665k and a collection of TextVQA and OCRVQA in 665k (denoted as ‘text+ocr’). We evaluate its downstream performance on MMVet, MMB and POPE, which can be found in Table 3. For a fair comparison, we strictly follow LLaVA-1.5’s SFT settings in experiments. It can be observed that all SFT data are helpful for multi-modal comprehension. Specifically, text related data are more helpful for dealing with model hallucinations (e.g., on POPE). In comparison, our SeVa achieves the overall best results on all benchmarks with both less training time cost and data usage. More importantly, SeVa enjoys its unsupervised nature, making it more generalizable in reality.

Generalize to other datasets. Since we adopt the questions in TextVQA and OCRVQA as a default fashion, we now ablate on more dataset selection on top of these two dataset. For more clarity, we choose other part of LLaVA665k including GQA part, COCO part and Visual Genome part to form the image-question pairs. We randomly select 2k, 4k, 8k and 16k raw image-question pairs and conduct SeVa pipeline to collect data and apply DPO training with 7B models. The results are then evaluated on MMVet. As shown in Table 4, all the chosen dataset lead to consistent improvement over LLaVA-1.5-7B (30.5% score), showing the generalized ability of SeVa to different image and question domains. Besides, involving more unsupervised data leads to better performance boost, which

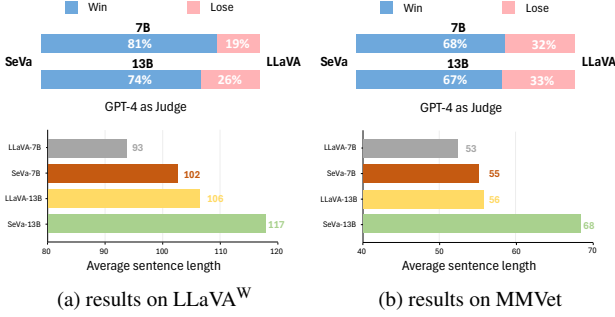


Figure 5. The pair-wise competition and output sentence length (by token) in LLaVA^W and MMVet, respectively. We compare between SeVa and LLaVA-1.5 models in 7B and 13B settings.

indicates the great power of SeVa towards further scaling.

Why SeVa works? Attentive readers might have a natural question: *why rejecting inferior answers lead to better model output than the original?* Here, we conduct a detailed experiment to show that, SeVa can be viewed as a special form of model calibration and rejecting negatives *does help*. Specifically, we collect image-question pairs from LLaVA^W to prompt LLaVA and SeVa with higher generation temperatures t for sampling. Note that $t=0$ means no sampling in token generation, which is the default evaluation settings and is most stable. Then we utilize GPT-4 (the detailed prompt can be found in appendix) to evaluate the its output consistency with scoring from 1 to 10, namely Q-consistency (how the model’s answer are aligned with the input questions) and A-consistency (how the model’s answer are aligned with its generated tokens w/o sampling). As shown in Table 5, the consistency score of both models fluctuates as t grows. However, SeVa has seen an consistent better score in all different t s, especially when temperature becomes higher. Since VLMs are easily affected by randomness in generating tokens [19], SeVa has potentially made a calibrating effect by reducing the sampled probability of those *bad* tokens, thus being more tolerable to interruption (e.g., temperatures). As a result, the model trained with SeVa can produce more robust and correct answers.

Besides, we have also observed a longer response in LLaVA^W and MMVet (*cf.* Fig. 5). This indicates that models trained with SeVa pipeline could potentially produce more detailed and meaningful responses that greatly benefits multi-modal comprehension.

Hard negatives matter. As we previously observed the phenomenon of hard negatives in Table 1, we now design a controllable experiment to study how strength of data augmentation affect the final training performance. Specifically, diffusion gaussian noise with different steps are chosen to generate preference data for DPO training, which are then evaluated on MMVet benchmark. As shown in Fig. 7a, larger noise steps will increase the reward margin

LoRA r	LoRA α	MMVet	MMB	POPE
—	—	30.7	64.3	85.9
64	128	32.8	64.7	86.5
128	256	33.9	65.0	86.4
256	512	34.0	65.1	86.6
512	1024	35.5	65.5	86.8
1024	2048	37.2	65.6	86.7
2048	4096	33.5	65.0	84.5

Table 6. The effect of hyper-parameter in LoRA [15] during our DPO training. We range the rank of r from 64 to 2048, while keeping the weight ratio fixed as 2 (α is always twice as r).

(*cf.* DPO [31] for more details) during training, indicating the increased preference divergence. In the meanwhile, a best trade-off exists in the noise steps (possibly between 500 and 800, as shown in Fig. 7b). We thus attribute the preference samples produced within this regime as *hard negatives* that are most valuable for DPO training.

Effect of Lora. For completeness, the hyper-parameters of LoRA’s low rank are ablated on three multi-modal comprehension benchmarks, which are shown in Table 6. In PETL [10], this parameter can be quite important, as it decides how much new knowledge the model can be absorbed during finetuning. As shown in the table, a relatively higher r is helpful to downstream tasks, showing the increased acquired knowledge of the VLM. However, too larger r will decrease the model performance, which can be attributed to the catastrophic forgetting in training a LLM (a similar point also pointed out in ShareGPT4V [4]). In conclusion, we set r as 1024 to seek a best trade-off, and always keep the delta importance as 2 (the α is always twice as r).

Aligning with user-intentions. Finally, we visualize the improved VLM with SeVa. We selected image-question pairs from MMVet and LLaVA^W, which are sent to the original VLM (LLaVA-1.5-7B) and the improved model (SeVa-7B), respectively. As shown in Fig. 6, Our SeVa demonstrates superiority over LLaVA in various aspects: stronger OCR ability, where SeVa could recognize the exact number of the gas price; less hallucinations can be observed in the second picture, where SeVa accurately comprehends the process of ‘chocolate cake recipe’ and produces the correct answers. It is also surprising that SeVa could potentially recovers more world knowledge after DPO training, as it gives detailed and accurately response to the introduction of a movie (the third picture). Finally, we found that through our DPO training, the models provide more detailed and helpful answers, as suggested by the last picture of a meal ordering. Following Vicuna [6], we also conduct a competition game between SeVa and LLaVA-1.5 under 7B and 13B settings on LLaVA^W and MMVet, respectively. We introduce GPT-4 as judge to evaluate the score of each model’s responses (similar to the evaluation process in these two



Figure 6. Four examples to illustrate the improved ability of our SeVa, including chain of thought (COT), stronger OCR, less hallucinations, world knowledge and more detailed and clearer responses. Since we built-on LLaVA-1.5, we directly make comparison between them to show the effectiveness of our proposed method. For clarity, we highlight the incorrect response in LLaVA-1.5 with red color, and emphasize ours with blue italic character. This figure is best viewed in color.

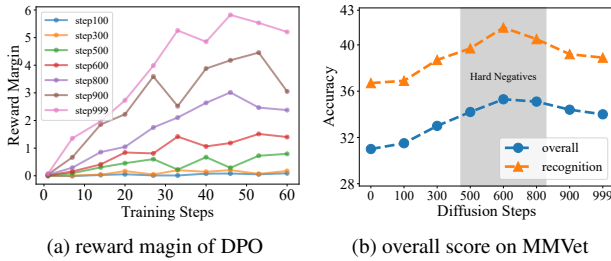


Figure 7. Exploration of hard-negatives. We plot the DPO reward margin between chosen and rejected samples with regard to different diffusion noise steps in data generation process (cf. 7a). We conduct DPO with preference data from diffusion noise augmentations, and show their evaluation scores on MMVet (cf. 7b).

benchmarks) and calculate the statistics of ‘win’ and ‘lose’. This could serve as a results that indicates how the models are aligned with user-intentions [45]. Note we didn’t include ‘tie’ [6], since we observed a majority of meaningless ‘tie’, where both SeVa and LLaVA achieve a score of zero. As shown in Fig. 5, SeVa has occupied most of the winning rounds in comparison with LLaVA, across both model sizes and benchmark datasets. Together with the visualizations in Fig. 6, we believe that our SeVa could probably serve as a strong pipeline to improve current VLM’s’ chat ability,

making them more suitable towards practical usage.

5. Conclusions and future work

In this paper, we introduce SeVa: Self-supervised visual preference alignment, for multi-modal comprehension. It first went through a data collection and filtering process with response from the original and distorted image, respectively. Then, the standard DPO training are applied to improve the model’s capability. Experiments in various benchmarks clear verify and efficiency of the approach. We also conduct ablations and quantitative visualizations to reveal the latent mechanism of SeVa, which demonstrate the great merit of aligning large models in vision-language domains.

In the future, we might consider generalize SeVa to various domains. We will study the effect of data scale: how the model will benefit from DPO training if more unlabeled data are involved, On top of that, we would investigate the potential to generalize our SeVa pipeline RLHF [29] domain. (e.g., whether the generated preference data could be helpful in building a reward model).

6. Acknowledgement

We thank Yin-Yin He (M.Sc. degree in Nanjing University, now in ByteDance) for his helpful connections. We also

appreciate the discussions with Jie Shao (Ph.D student in Nanjing University) about LLMs.

A. Relations with contrastive learning

As discussed before in Sec. 3.3, the optimized function in SeVa has strong relations with visual contrastive learning. In this section, we will derive a more general form of DPO loss that can be easily adapted to our SeVa pipeline. We first rewrite the InfoNCE loss here for clarity:

$$\mathcal{L}_{\text{in}} = -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+/\tau) + \sum_i^n \exp(q \cdot k_-^i/\tau)}, \quad (14)$$

Inspired by Eq. 14 that multiple negative terms are involved, we can reformulate the preference distribution in Eq. 3 as:

$$\begin{aligned} p_{\text{multi}}^* &= p^*(y_c \succ \mathbf{Y}_r | x) \\ &= \frac{\exp(r^*(x, y_c))}{\exp(r^*(x, y_c)) + \sum_{y_r \in \mathbf{Y}_r} \exp(r^*(x, y_r))}, \end{aligned} \quad (15)$$

where one positive preference data is paired with multiple rejected samples that are represented by a union \mathbf{Y}_r . In this case, the preference database changes to:

$$\mathcal{D}^{\text{multi}} = \{x^{(j)}, y_c^{(j)}, \mathbf{Y}_r^{(j)}\}_{j=1}^{N_d}. \quad (17)$$

Note that in SeVa pipeline, this new preference database in Eq. 17 can be easily constructed (e.g., we can do data augmentation sampling across range of available ones to obtain more negatives for \mathbf{Y}_c). Now, by considering the closed form in of the reward function in Eq. 6 and utilizing samples in Eq. 17 (to do maximum likelihood estimation), we derive a more general form of DPO as:

$$\begin{aligned} \mathcal{L}_{\text{dpo}}^{\text{multi}} &= -\mathbb{E}_{\mathcal{D}^{\text{multi}}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta'}(y_c|x)}{\pi_{\text{ref}}(y_c|x)} \right. \right. \\ &\quad \left. \left. - \sum_{y_r \in \mathbf{Y}_r} \beta \log \frac{\pi_{\theta'}(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right) \right], \end{aligned} \quad (18)$$

and optimize it using pseudo-constructed database by our SeVa pipeline. At present, this generalized form is out of the scope of this paper. We will leave this as future work to explore the potentials of our SeVa.

B. Data construction details

In Sec. 4.1, we have discussed the data construction process. Here, we want to emphasize the *counting* of our data instance. Note that for clarity in all our experiment, we choose the num of the unfiltered preference sample as the DPO instances we used. But please note that the actual preference data sent to DPO are much less than that. One is the

SeVa’s filtering mechanism as discussed before (the filtering process in Alg. 1 almost cut the instance num by half). The other is the data counting difference between LLaVA and SeVa: an instance of multi-turn dialogue will contribute to about 2 raw image-question pairs in our SeVa. As such, the data used in SeVa *can be* far less counted if we ‘regard’ this 2 image-question pairs as only one data instance that is sourced from a multi-dialogue.

C. More visualizations.

C.1. Data augmentations samples

In Fig. 9, we visualize the data augmentation samples to show how the negatives responses are obtained. We choose original image, random-flip (‘RandFlip’), random-resized-crop (‘RRCrop’), moco augmentations and diffusion noise for better visualizations. We highlight those in-correct of in-complete answers with red color.

C.2. GPT-4 prompt template

Please refer to Fig. 8 for the GPT-4 prompt template we adopted to evaluate the consistency score in Table 5.

C.3. More examples of SeVa

Please refer to Fig. 10-11. In Fig. 10, we showcase more examples of SeVa-13B to illustrate its multi-modal comprehension ability, including complex reasoning about a wedding, chain-of-thought power to make a drink and the capability to read scientific figures. In Fig. 11, we compare SeVa-13B with more VLMs, namely LLaVA-1.5-13B and Qwen-VL-chat. We found our model demonstrate a detailed and faithful answers, which is on par with (or even better than) Qwen-VL-chat that possibly requires more training data source and larger model capacity.



ChatGPT for Consistency Evaluation

You are an AI assistant to help me identify the consistency of two set of answers generated by 2 Vision-Language models with different weights. The answers are generated according to the input image (invisible in your system), and an asked question (the 'prompt' dict in each question answer pair). The input to you are shown in the following format:

[task-specific input to the GPT]

Illustrations for input-specific information

The "prompt" value in the dict are the same questions asked to the two models (as mentioned before), while the 'text' value in each dict are the answers to Model1 and Model2, respectively. Now please help me evaluate two metrics, called "Question consistency of model1", "Question consistency of model2" and "consistency between model1 and model2" with an integer score from 1 to 10 (1 is the least consistency, and the 10 is the most consistent). Note that DO NOT writing any code to evaluate, but only evaluate your own, that is in your own system!

Illustrations for output-specific information

Here is the meaning of the three metrics

"A1, Question consistency of model1": the answered information ("text") of the model1 with regard to the asked questions ("prompt").

"A2, Question consistency of model2": the answered information ("text") of the model2 with regard to the asked questions ("prompt").

"A3, consistency between model1 and model2": The consistency of model2's answer compared with model1's answer. If the model2 outputs very different answers from model1, you should give this metric a relatively low score, and vise verse.

You should evaluate the consistency between the following 3 aspects:

1. wording: is the words used in the sentence consistency of not (score from 1-10, with 1 the least)
2. sentiment: is the sentiment conveyed consistent? (score from 1-10, with 1 the least)
3. content: is the overall content of facts consistent? (score from 1-10, with 1 the least)

Now I will show you one paired input, please output the metrics as follows

Requirements for output format

You need to figure out the [score] in the [OUTPUT] format. Also, print the SIMPLIFIED FORMAT as follows, and fill the correct number in the corresponding metrics.

Output examples

OUTPUT: [evaluation-specific output format]

SIMPLIFIED FORMAT:

A1,[wording],[content],[sentiment],[overall]

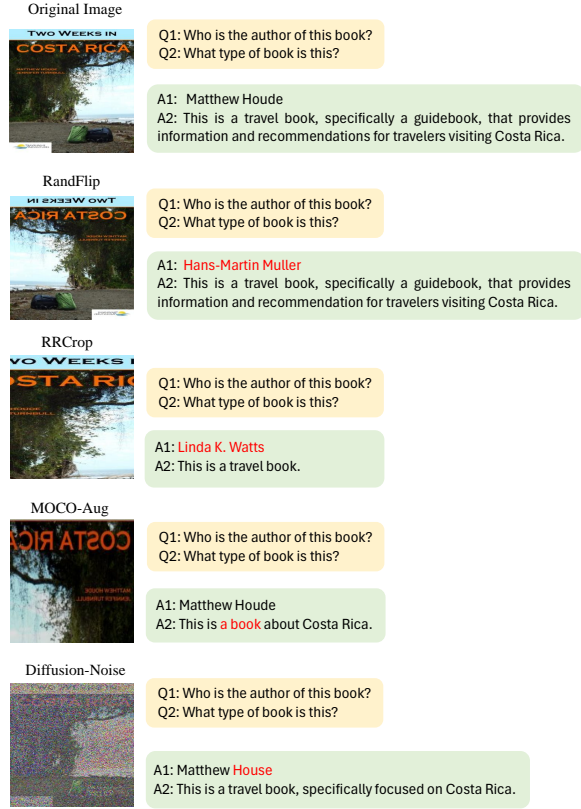
A2,[wording],[content],[sentiment],[overall]

A3,[wording],[content],[sentiment],[overall]

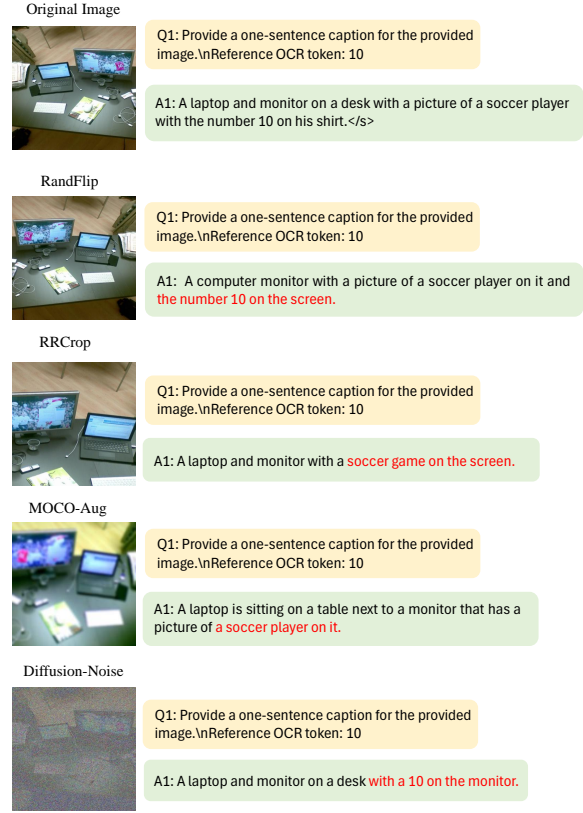
The INPUT Prompt is:

Prompt [Show your input to GPT for evaluation of consistency score]

Figure 8. Prompt template to evaluate the consistency score. The numerical results are show in Table 5.



(a) data augmentations of OCRVQA [28] images



(b) data augmentations of TextVQA [33] images

Figure 9. Visualizations of the distorted images and their model answers. In-complete or in-correct tokens are highlighted with red colors.



Q: Why are the man and woman in white so close to each other? Answer it and give the rationale.



SeVa-13B:

The man and woman in white are close to each other because they are a bride and groom celebrating their wedding. In the image, the man is holding a glass of champagne and toasting, while the woman is also holding a glass of champagne and drinking. This scene represents a joyful moment during their wedding reception, where the couple is sharing a toast and enjoying each other's company. The close proximity of the couple symbolizes their unity, love, and commitment to each other as they embark on their new journey as husband and wife.



Q: How to make this?

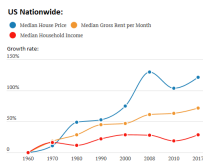


SeVa-13B:

To make this refreshing citrus drink with grapefruit balls and a sprig of mint, follow these steps:

1. Gather ingredients:
 - Fresh grapefruit: peel and segment the grapefruit to remove the bitter pith and membranes. You can use a grapefruit zester or a sharp knife to do this.
 - Fresh orange: peel the orange to remove the outer layer and segment it to remove the seeds and pith.
 - Lime: peel the lime and remove seeds.
 - Mint leaves: fresh or dried, for garnish
 - Simple syrup (optional): you can make your own or use store-bought
 - Ice
 - Glass: a clear glass or a glass with a stem, preferably with a base wide enough to hold the grapefruit balls
 - Toothpick or skewer
2. Prepare the grapefruit balls:
 - Combine the grapefruit segments, simple syrup (if using), and a squeeze of lime juice in a bowl.
 - Mix well and refrigerate for at least 30 minutes to allow the flavors to meld.
 - Form the grapefruit segments into balls using your hands or a small cookie scoop.
3. Assemble the drink:
 - Fill the glass with ice.
 - Add a few grapefruit balls to the glass.
 - Garnish the drink with a sprig of mint and a toothpick or skewer to hold it in place.
 - Pour the prepared grapefruit juice over the ice and grapefruit balls.
 - Squeeze a wedge of lime over the drink and drop it in to release the juice.
 - Stir gently and serve immediately.

This citrus drink is perfect for warm weather or as a refreshing treat during any time of the year. Enjoy!



Q: Which year has the highest growth rate of median house price?



SeVa-13B:

The year with the highest growth rate of median house price is 2000, as indicated by the blue line on the chart. This suggests that during that time, there was a significant increase in the median house price, which could be attributed to various factors such as economic growth, low-interest rates, and increased demand for housing.

Figure 10. Visualizations of complex visual question answering with our SeVa-13B models.



Q: What are the meals that I can cook with these?

LLaVA-1.5-13B: With the ingredients available in the refrigerator, you can prepare a variety of meals. The refrigerator contains several carrots, which can be used as a healthy snack or incorporated into dishes like salads, stir-fries, or soups. The milk can be used to make smoothies, pancakes, or creamy sauces for pasta dishes. The yogurt can be used as a base for dips, dressings, or as a topping for oatmeal or granola. The strawberries can be eaten on their own or used as a topping for desserts or fruit salads. The eggs can be used to make omelettes, frittatas, or scrambled eggs. The juice can be consumed as a refreshing beverage or used as an ingredient in recipes like marinades or glazes. The bread can be used to make sandwiches, toast, or as a base for various savory or sweet dishes. The butter can be used for cooking or baking, and the cream can be used in desserts or sauces. The strawberries and cream can be combined to make a delicious fruit salad or dessert. Overall, the refrigerator offers a diverse range of ingredients that can be used to create a variety of healthy and delicious meals.



Qwen-VL-Chat:

Based on the ingredients visible in the image, you could prepare a variety of meals. Here are some ideas:

Yogurt Parfait: Layer Greek yogurt, fresh berries (strawberries and blueberries), and granola or honey for a healthy breakfast or snack.
 Smoothie Bowl: Blend frozen mixed berries, banana, and almond milk to make a smoothie base. Top it with fresh fruit, yogurt, and granola for a nutritious breakfast or snack.
 Carrot Salad: Mix shredded carrots with raisins, chopped apples, and a simple dressing made from olive oil, lemon juice, and honey. Serve as a side dish or light lunch.
 Fruit Salad: Combine strawberries, blueberries, and other fruits like bananas, grapes, or pineapple for a refreshing dessert or snack.
 Yogurt Dip: Mix Greek yogurt with garlic, lemon juice, and herbs like dill or parsley. Serve with carrot sticks or other vegetables for a healthy appetizer or snack.
 Remember to store your ingredients properly to maintain their freshness and quality.



SeVa-13B:

With the variety of food items in the refrigerator, you can prepare a range of meals and snacks. Some ideas include:

1. Fruit salad: Combine the strawberries and blueberries to create a refreshing fruit salad. You can also add other fruits like apples, oranges, or grapes for added variety.
2. Yogurt parfait: Use the yogurt as a base for a delicious yogurt parfait. Layer the yogurt with granola, fresh fruit, and nuts for a nutritious and satisfying breakfast or snack.
3. Carrot sticks with dip: Slice the carrots into sticks and serve them with a healthy dip, such as hummus, guacamole, or a yogurt-based dip.
4. Smoothie: Blend the yogurt, fruit, and any additional ingredients like nuts, seeds, or protein powder to create a nutritious and delicious smoothie.
5. Milk-based dessert: Use the milk to make a creamy dessert, such as a milkshake, pudding, or custard. You can also mix it with fruit or yogurt to create a refreshing dessert.
6. Oatmeal: Use the milk as a base for oatmeal, and add fruits, nuts, and other toppings for a hearty and filling breakfast.

These are just a few examples of the meals and snacks you can create using the food items in the refrigerator. With a little creativity and experimentation, you can come up with many more delicious and nutritious dishes.



Figure 11. Comparison of our SeVa-13B with other VLMs (LLaVA-1.5, Qwen-VL-Chat).

References

- [1] Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, and Hervé Le Borgne. Explicit knowledge integration for knowledge-aware visual question answering about named entities. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 29–38, 2023. **2**
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. **1**
- [3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. **3**
- [4] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. **1, 2, 5, 7, 8**
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. pages 1597–1607, 2020. **3, 5**
- [6] 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, 2023. **8, 9**
- [7] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023. **2**
- [8] Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, pages 49250–49267, 2023. **1, 2**
- [9] Runpei Dong, Chunrui Han, Yang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2024. **2**
- [10] Minghao Fu, Ke Zhu, and Jianxin Wu. Dtl: Disentangled transfer learning for visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12082–12090, 2024. **8**
- [11] Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*, 2023. **3**
- [12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: a new approach to self-supervised learning. In *NeurIPS*, pages 21271–21284, 2020. **3**
- [13] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023. **3, 6, 7**
- [14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. **1, 3, 5, 7**
- [15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. **8**
- [16] 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. **6, 7**
- [17] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. **7**
- [18] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. **6**
- [19] Shengzhi Li, Rongyu Lin, and Shichao Pei. Multimodal preference alignment remedies regression of visual instruction tuning on language model. *arXiv preprint arXiv:2402.10884*, 2024. **1, 3, 8**
- [20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. **1, 3, 6**
- [21] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. **1, 5**
- [22] Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. Stablelava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*, 2023. **1, 2, 5**
- [23] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoyebi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023. **1, 2**

- [24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#)
- [25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. [1](#), [2](#), [3](#), [6](#)
- [26] 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. [1](#), [3](#), [6](#)
- [27] Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. [6](#)
- [28] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. [1](#), [5](#), [7](#), [12](#)
- [29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. [3](#), [4](#), [9](#)
- [30] 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. [3](#)
- [31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. 2024. [3](#), [4](#), [6](#), [8](#)
- [32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. [3](#), [4](#), [6](#)
- [33] 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. [5](#), [7](#), [12](#)
- [34] Lina Sun, Yewen Li, and Yumin Dong. Learning from expert: Vision-language knowledge distillation for unsupervised cross-modal hashing retrieval. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 499–507, 2023. [1](#), [2](#)
- [35] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. [1](#), [3](#)
- [36] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. [1](#), [2](#), [5](#), [7](#)
- [37] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*, 2023. [2](#)
- [38] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. *arXiv preprint arXiv:2312.00589*, 2023. [2](#)
- [39] 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. [1](#), [3](#), [6](#)
- [40] Arun Zachariah and Praveen Rao. Video retrieval for everyday scenes with common objects. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 565–570, 2023. [1](#)
- [41] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. [2](#)
- [42] Bo Zhao, Boya Wu, and Tiejun Huang. Svlt: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. [1](#)
- [43] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. *arXiv preprint arXiv:2307.09474*, 2023. [2](#)
- [44] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. [1](#), [3](#), [6](#), [7](#)
- [45] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024. [9](#)
- [46] 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. [1](#), [2](#)
- [47] Ke Zhu, Minghao Fu, and Jianxin Wu. Multi-label self-supervised learning with scene images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6694–6703, 2023. [1](#), [3](#)
- [48] Ke Zhu, Yin-Yin He, and Jianxin Wu. Coarse is better? a new pipeline towards self-supervised learning with uncurated images. *arXiv preprint arXiv:2306.04244*, 2023. [1](#)
- [49] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024. [3](#)