Trident: Self-Supervised Preference Alignment via Triplet Regularization

Yingnan Guo[1], Kejia Chen[1], Xiaofeng Zhang[2], Zifei Wu[1], Yu Zhang[1]*,
[1]Zhejiang University, [2]Shanghai Jiao Tong University
22432017@zju.edu.cn, chenkejia@zju.edu.cn, framebreak@sjtu.edu.cn
zifeiwu@zju.edu.cn, zhangyu80@zju.edu.cn

Abstract

Aligning Large Vision-Language Models (LVLMs) to mitigate hallucinations typically relies on high-quality preference data. However, in self-supervised settings, standard binary preference optimization (e.g., DPO) suffers from noisy supervision and semantic ambiguity, as automatically generated chosen responses are not guaranteed to be superior to rejected ones. In this work, we propose **Trident**, a fully self-supervised framework that ensures robust alignment via a structured triplet paradigm. Trident autonomously constructs reliable preference triplets—comprising semantically enriched (chosen), degraded (rejected), and neutral (anchor) responses—through automated visual perturbations and self-summarization. We further introduce *Trident Preference Regularization* (TPR), a novel objective that utilizes an adaptive margin to enforce semantic separation between the triplet components while preventing deviation from the pretrained distribution. Despite requiring no human annotations or external reward models, Trident consistently outperforms state-of-the-art RLHF and RLAIF baselines. For instance, on LLaVA-1.5-7B, it reduces the hallucination rate on AMBER to 11.3% and achieves 95.70% precision on POPE using only **4k self-generated** triplets and a **single epoch**. This validates structured triplet supervision as a scalable paradigm for robust self-supervised alignment.

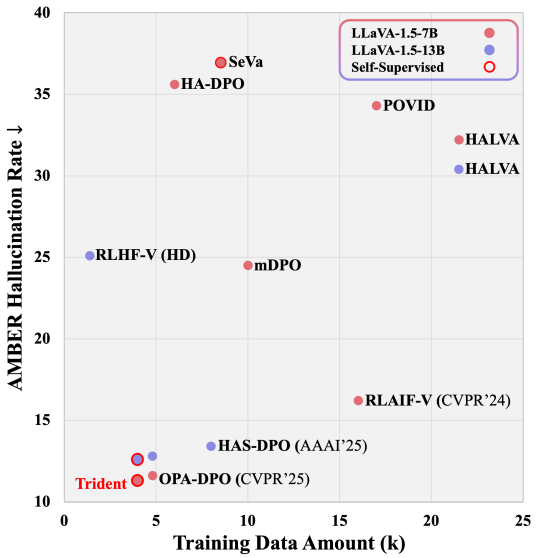

Figure 1: This figure illustrates AMBER hallucination rates of various algorithms across different training data sizes. Trident achieves the lowest hallucination rates with relatively minimal self-supervised data.

1 Introduction

Large Vision-Language Models (LVLMs) have achieved remarkable success (Sun et al., 2023a; Zachariah and Rao, 2023; Dong et al., 2023), enabling sophisticated multimodal capabilities ranging from visual question answering to detailed image description (Radford et al., 2021; Li et al., 2022; Dai et al., 2023; Alayrac et al., 2022; Chen et al., 2023b; Zhu et al., 2023). Despite their impressive performance, a critical challenge remains: aligning model outputs with visual reality and user intent. A frequent failure mode is hallucination, where the model generates text that is factually inconsistent with or unsupported by the visual input. To mitigate this, the community has increasingly turned to preference alignment techniques such as Reinforcement Learning from Human Feedback (RLHF) (Sun et al., 2023b; Yu et al., 2024) and Direct Preference Optimization (DPO) (Rafailov et al., 2023; Sun et al., 2023b). These methods fine-tune models on preference pairs to favor desired behaviors. DPO, in particular, has become a popular choice due to its stability and efficiency, as it dispenses with explicit reward modeling and complex reinforcement learning loops.

However, the efficacy of DPO hinges on access to a large corpus of high-quality preference data,

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 31668–31683
July 2-7, 2026

typically in the form of (y_c, y_r) pairs where a chosen response y_c is demonstrably superior to a rejected one y_r. While effective when sourced from human annotators or powerful proprietary models like GPT-4V (Chen et al., 2023a; Zhao et al., 2023a), this requirement creates a significant bottleneck in fully **self-supervised settings** (Zhou et al., 2024b; Zhu et al., 2024). When preference pairs are generated automatically—for instance, by comparing responses from augmented versus original inputs—there is no guarantee of a clear or reliable semantic gap between the chosen and rejected candidates. This semantic ambiguity introduces noisy supervision, which destabilize training and fail to effectively reduce hallucinations. Furthermore, the binary nature of DPO limits its ability to model the nuanced relationships between multiple potential responses, offering only a coarse directional signal.

To overcome these limitations, we introduce **Trident**, a self-supervised framework that transcends binary DPO via a structured triplet paradigm. Trident ensures robust alignment by automatically constructing triplets—comprising enriched (chosen), degraded (rejected), and stable anchor responses—to maximize semantic separation while preventing distribution drift. We further propose **Trident Preference Regularization (TPR)**, an adaptive margin objective that enforces these constraints. As shown in Figure 1, Trident achieves state-of-the-art performance on AMBER with remarkably high data efficiency. Our contributions are as follows:

- We identify key limitations of binary DPO in self-supervised settings and propose Trident, a triplet-based alignment framework that leverages an anchor to mitigate supervision noise.
- We introduce an adaptive margin objective Trident Preference Regularization (TPR), which dynamically enforces semantic separation, preventing model drift and ensuring groundedness without external rewards.
- Trident surpasses state-of-the-art RLHF and RLAIF baselines on multiple hallucination benchmarks, using only 4K self-generated triplets and a single training epoch.

2 Related Work

2.1 Preference Alignment in LVLMs

To reduce hallucinations in Large Vision-Language Models (LVLMs), recent works have explored preference-based alignment techniques such as Reinforcement Learning from Human Feedback (RLHF) (Sun et al., 2023b; Kaufmann et al., 2024; Yu et al., 2024) and its variant using AI-generated feedback, RLAIF (Zhao et al., 2023b; Xiao et al., 2025; Yu et al., 2025). These approaches fine-tune models using preference pairs in which the preferred response exhibits fewer hallucinations for the same image and prompt, thereby guiding the model toward more faithful outputs.

While Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) both aim to maximize rewards derived from the Bradley-Terry model (Bradley and Terry, 1952) under Kullback-Leibler (KL) divergence regularization (Kullback and Leibler, 1951), they differ significantly in training strategy. PPO is an on-policy algorithm that requires a reward model and online rollouts, introducing higher computational overhead. In contrast, DPO streamlines the process by relying solely on offline preference data collected from any policy, avoiding explicit reward modeling and online interaction (Li et al., 2024; Wang et al., 2024; Chen et al., 2025).

Due to its simplicity and scalability, DPO has emerged as a practical alternative for aligning LVLMs without compromising training efficiency, especially in settings where collecting high-quality offline preference data is feasible.

2.2 Triplet Loss in Machine Learning

Triplet loss has long been a cornerstone of metric learning, organizing embedding spaces by enforcing relative distances within an (anchor, positive, negative) tuple (Schroff et al., 2015; Hermans et al., 2017). Unlike simple binary classification, this paradigm explicitly models the relative structure of data. In the realm of preference alignment, however, dominant approaches like DPO (Rafailov et al., 2023) remain constrained to binary comparisons. This formulation is brittle in self-supervised settings (Chen et al., 2026), where the quality gap between automatically generated "chosen" and "rejected" responses is often ambiguous or noisy, leading to unstable optimization.

Drawing inspiration from the triplet paradigm, we propose to transcend this binary limitation by introducing a stable reference anchor into the preference objective. By enforcing structured semantic gaps among a semantically enriched chosen response, a degraded rejected response, and a neutral

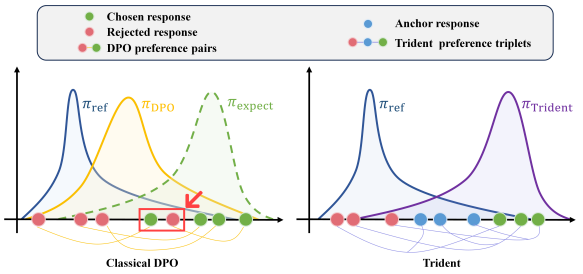

Figure 2: **Trident Motivation:** This figure illustrates the issue in self-supervised preference alignment where the chosen response may not be better than the rejected one, impacting preference alignment. This problem is mitigated through the structured triplet supervision used in Trident.

anchor, we can derive richer supervision signals than pairwise ranking allows. This structural extension is crucial for robust alignment in fully self-supervised settings, where reliable ground-truth preference labels are unavailable.

3 Method

3.1 Problem Formulation and Motivation

Preliminaries. To align the model with preferred responses, Direct Preference Optimization (DPO) (Rafailov et al., 2023) operates on binary preference pairs (x, y_c, y_r), where y_c is the preferred response and y_r the rejected one. DPO models pairwise preferences using the Bradley–Terry framework and defines the loss as:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \left[\log \frac{\pi_\theta(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \log \frac{\pi_\theta(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right] \right), \tag{1}$$

where π_{ref} is a frozen reference model, and β controls the sharpness of preference separation.

Motivation. While effective with high-quality data, standard DPO relies on the strict assumption that $y_c \succ y_r$. In self-supervised settings, this assumption is brittle: automatically generated "chosen" responses are not guaranteed to be semantically superior to "rejected" ones, leading to *supervision noise* (Yu et al., 2023). As illustrated in Figure 2, such ambiguity causes unstable alignment. Trident addresses this by extending the binary framework to a triplet formulation, introducing a neutral anchor to explicitly regularize the semantic space and ensure robust optimization.

3.2 Self-Supervised Triplet Construction

Constructing high-quality preference data without human annotations poses fundamental challenges (Awais et al., 2025), particularly in self-supervised regimes where no explicit preference signals are available. Traditional DPO relies on reliable preference pairs (y_c, y_r), often curated through human evaluation. However, in autonomous settings, naively pairing two model-generated responses may result in insufficient semantic contrast or even contradictory supervision. Furthermore, enforcing preference constraints without regularization can lead to degenerate optimization behaviors, such as semantic drift or loss collapse.

We propose a self-supervised triplet construction framework that simultaneously induces semantic variation and constrains learning through a stable anchor reference. For each image-question pair (I, q), we construct a triplet $(y_c, y_r, y_{\text{anchor}})$, where y_c is a semantically enriched response (chosen), y_r is a degraded response (rejected), and y_{anchor} is a reference response serving as a semantic midpoint. This triplet not only provides directional supervision but also preserves alignment with the pretrained model distribution.

Constructing Chosen Responses (y_c). To obtain the chosen response y_c, we apply a set of semantic-preserving transformations $\mathcal{T}_{\text{chosen}}$ to the image I. These include no perturbation, sharpening, small-angle rotation, contrast enhancement, and weak cropping—operations that preserve the overall content while introducing minor stylistic variation. Each transformed image $t(I)$ is encoded via the vision backbone $g(\cdot)$ and passed into a supervised finetuned policy π_{SFT} to yield a set of candidate responses:

$$\mathcal{Y}_{\text{chosen}} = \{\pi_{\text{SFT}}(g(t(I)), q) \mid t \in \mathcal{T}_{\text{chosen}}\}, \tag{2}$$

These responses are then aggregated using a *predefined summarization prompt* to obtain a unified, informative response:

$$y_c = \texttt{Self-Summarize}(\mathcal{Y}_{\text{chosen}}), \tag{3}$$

Importantly, this summarization is entirely self-supervised: it operates without human feedback, relying solely on prompt-guided distillation over multiple augmented views. This allows the model to consolidate its own semantic evidence into a single chosen response, thereby providing a higher-quality and more reliable supervision signal for preference optimization.

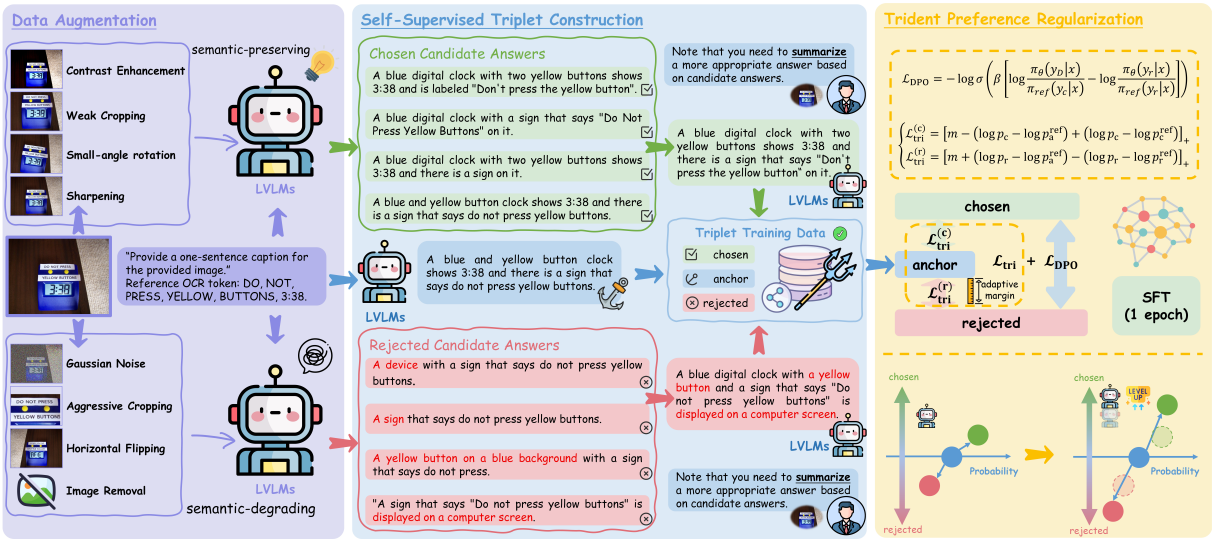

Figure 3: **Overview of the Trident pipeline.** Our framework begins with **Data Augmentation** (left) to generate diverse image inputs, followed by **Self-Supervised Triplet Construction** (center) that produces chosen, rejected, and anchor responses using a supervised finetuned policy and a frozen reference model. **Trident Preference Regularization** (right) then optimizes the model with a triplet-based loss incorporating an adaptive margin.

Constructing Rejected Responses (y_{r}). For the rejected response, we utilize a set of semantic-degrading transformations $\mathcal{T}_{\text{rejected}}$, which consists of semantic-degrading operations including Gaussian noise, aggressive cropping, horizontal flipping, and image removal. These transformations distort or remove critical visual cues, thereby reducing the model's ability to produce accurate and grounded responses. Given a sampled transformation $t_{\text{rej}} \in \mathcal{T}_{\text{rejected}}$, we compute:

$$\mathcal{Y}_{\text{rejected}} = \left\{ \pi_{\text{SFT}}(g(t(I)), q) \middle| t \in \mathcal{T}_{\text{rejected}} \right\}, \quad (4)$$

As with the chosen responses, we apply self-summarization to reduce variability and reinforce consistent degradation:

$$y_{\mathrm{r}} = \texttt{Self-Summarize}(\mathcal{Y}_{\text{rejected}}), \quad (5)$$

Integrating the Anchor (y_{anchor}). To stabilize preference alignment and prevent overfitting to noisy preference gaps in self-supervised settings, we introduce an anchor response y_{anchor}, generated from the unperturbed image I via a frozen reference model π_{ref}:

$$y_{\text{anchor}} = \pi_{\text{ref}}(g(I), q), \quad (6)$$

This response serves as a semantic baseline against which both y_{c} and y_{r} are compared, enabling the model to reason over relative preferences while preserving consistency with the pretrained model distribution.

The final training set thus comprises all triplets of the form:

$$\mathcal{D}_{\text{triplet}} = \{(x, y_{\mathrm{c}}, y_{\mathrm{r}}, y_{\text{anchor}})\}, \quad (7)$$

where $x = (g(I), q)$. This formulation facilitates more expressive supervision than binary preference alone by explicitly modeling the semantic gradient between diverse outputs, regularized by an intermediate reference. The use of self-summarized preferred responses ensures the entire triplet generation process remains fully self-supervised while benefiting from enhanced response quality. The overall process is summarized in **Algorithm 1**.

3.3 Trident Preference Regularization

To optimize preference alignment under the triplet structure, we propose **Trident Preference Regularization (TPR)**—a structured objective that extends Direct Preference Optimization (DPO) by integrating an adaptive margin-based constraint over the chosen, rejected, and anchor responses. This regularization encourages semantically meaningful separation while preserving alignment with pretrained model behavior.

Given the structured triplets $(x, y_{\mathrm{c}}, y_{\mathrm{r}}, y_{\text{anchor}})$ constructed in the previous section, Trident Preference Regularization augments the original DPO formulation by incorporating a margin-based regularizer that explicitly leverages the semantic structure induced by the triplet. The design is motivated

Algorithm 1 Self-Supervised Triplet Construction

Require: Image I, question q
Require: SFT model π_{SFT}, reference model π_{ref}
Require: Augmentation sets $\mathcal{T}_{\text{chosen}}$, $\mathcal{T}_{\text{rejected}}$
Ensure: Triplet $(x, y_{\text{c}}, y_{\text{r}}, y_{\text{anchor}})$

Initialize $Y_{\text{chosen}} \leftarrow \emptyset$
for $t \in \mathcal{T}_{\text{chosen}}$ **do**
  $I_{\text{aug}} \leftarrow t(I)$
  $y \leftarrow \pi_{\text{SFT}}(g(I_{\text{aug}}), q)$
  $Y_{\text{chosen}} \leftarrow Y_{\text{chosen}} \cup \{y\}$
end for
$y_{\text{c}} \leftarrow \texttt{Self-Summarize}(Y_{\text{chosen}})$
Initialize $Y_{\text{rejected}} \leftarrow \emptyset$
for $t \in \mathcal{T}_{\text{rejected}}$ **do**
  $I_{\text{deg}} \leftarrow t(I)$
  $y \leftarrow \pi_{\text{SFT}}(g(I_{\text{deg}}), q)$
  $Y_{\text{rejected}} \leftarrow Y_{\text{rejected}} \cup \{y\}$
end for
$y_{\text{r}} \leftarrow \texttt{Self-Summarize}(Y_{\text{rejected}})$
$y_{\text{anchor}} \leftarrow \pi_{\text{ref}}(g(I), q)$
if $y_{\text{c}} = y_{\text{r}}$ or $y_{\text{c}} = y_{\text{anchor}}$ or $y_{\text{r}} = y_{\text{anchor}}$ **then**
  Discard triplet
else
  $x \leftarrow (g(I), q)$
  Save $(x, y_{\text{c}}, y_{\text{r}}, y_{\text{anchor}})$
end if

by two key observations: (1) binary preference signals alone may be insufficient to capture nuanced differences between plausible responses; and (2) preference learning should maintain semantic consistency with the pretrained distribution, captured via the anchor response y_{anchor}.

We begin with the standard DPO loss, defined over a binary preference pair $(y_{\text{c}}, y_{\text{r}})$ under input x. Let p denote the likelihood probability assigned by the model. We define the log-probabilities under the policy model π_θ and the frozen reference model π_{ref} as follows:

$$\log p_{\text{c}} = \log \pi_\theta(y_{\text{c}} \mid x), \quad \log p_{\text{r}} = \log \pi_\theta(y_{\text{r}} \mid x),$$
$$\log p_{\text{c}}^{\text{ref}} = \log \pi_{\text{ref}}(y_{\text{c}} \mid x), \quad \log p_{\text{r}}^{\text{ref}} = \log \pi_{\text{ref}}(y_{\text{r}} \mid x),$$

The classical DPO term is:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\left(\beta \cdot \left[\log p_{\text{c}} - \log p_{\text{r}} - (\log p_{\text{c}}^{\text{ref}} - \log p_{\text{r}}^{\text{ref}})\right]\right), \quad (8)$$

where β is a scaling hyperparameter controlling preference sharpness.

To fully utilize the triplet structure, we introduce a margin-based regularization over the anchor. Let $\log p_{\text{a}} = \log \pi_\theta(y_{\text{anchor}} \mid x)$ and $\log p_{\text{a}}^{\text{ref}} = \log \pi_{\text{ref}}(y_{\text{anchor}} \mid x)$ denote the log-probabilities for the anchor under the policy and reference models. To dynamically adapt regularization strength, we define the average model shift:

$$\Delta = \frac{1}{2}\left(|\log p_{\text{c}} - \log p_{\text{c}}^{\text{ref}}| + |\log p_{\text{r}} - \log p_{\text{r}}^{\text{ref}}|\right), \quad (9)$$

and compute the adaptive margin as:

$$m = \max(\alpha \cdot \Delta, \epsilon), \quad (10)$$

where $\alpha > 1$ is a tunable scaling factor, and ϵ ensures numerical stability.

We then define two hinge-style penalties that enforce separation between the policy and anchor responses, regularized by model shift. The first term encourages the policy model to assign significantly higher likelihood to y_{c} than to y_{anchor}:

$$\mathcal{L}_{\text{tri}}^{(\text{c})} = \left[m - (\log p_{\text{c}} - \log p_{\text{a}}^{\text{ref}}) + (\log p_{\text{c}} - \log p_{\text{c}}^{\text{ref}})\right]_+, \quad (11)$$

while the second discourages the policy from assigning higher probability to y_{r} than to y_{anchor}:

$$\mathcal{L}_{\text{tri}}^{(\text{r})} = \left[m + (\log p_{\text{r}} - \log p_{\text{a}}^{\text{ref}}) - (\log p_{\text{r}} - \log p_{\text{r}}^{\text{ref}})\right]_+, \quad (12)$$

where $[\cdot]_+ = \max(0, \cdot)$ denotes the ReLU operator.

The total regularization term is:

$$\mathcal{L}_{\text{tri}} = \mathcal{L}_{\text{tri}}^{(\text{c})} + \mathcal{L}_{\text{tri}}^{(\text{r})}, \quad (13)$$

The final training objective of Trident Preference Regularization is:

$$\mathcal{L}_{\text{Trident}} = \lambda_{\text{dpo}} \cdot \mathcal{L}_{\text{DPO}} + \lambda_{\text{tri}} \cdot \mathcal{L}_{\text{tri}}. \quad (14)$$

where λ_{dpo} and λ_{tri} are hyperparameters balancing core DPO learning and anchor-aware regularization.

This formulation preserves the preference supervision from the DPO core, enforces semantic separation with respect to a stable anchor, and adapts regularization based on observed policy drift. Together, these components make Trident Preference Regularization a robust and scalable framework for self-supervised preference alignment.

4 Experiments

4.1 Experimental Setup

Implementation Details. We take LLaVA-1.5-7B / 13B as the backbone models and train both the 7B and 13B models on 4K self-supervised triplets

Table 1: Comparison of RLAIF/RLHF-based algorithms for enhancing LVLMs across benchmarks. Baseline results with official checkpoints are retested (marked §), while others are sourced from the respective paper († denotes (Yang et al., 2025)). All evaluations use greedy sampling for consistency, and the best results within each group are highlighted in **bold**, while the second-best performance is highlighted in underline.

Algorithm	Data Size	Feedback	AMBER				MMHal-Bench		Object HalBench		POPE Adversarial	
			CHAIR↓	Cover↑	HalRate↓	Cog↓	Score↑	HalRate↓	CHAIRs↓	CHAIRi↓	Acc.↑	Pre.↑
LLaVA-1.5-7B (Liu et al., 2024b,a)§			7.7	51.6	34.7	4.2	2.01	0.61	55.67	15.96	84.93%	89.10%
+LLaVA-RLHF§ (Sun et al., 2023b)	122k	Self-Reward	9.7	**53.2**	46.6	5.3	1.88	0.71	58.00	15.61	80.00%	87.19%
+HALVA† (Sarkar et al., 2024)	21.5k	GPT-4V	6.6	53.0	32.2	3.4	2.25	0.54	41.40	11.70	-	-
+mDPO† (Wang et al., 2024)	10k	GPT-4V	4.4	52.4	24.5	2.4	2.39	0.54	35.70	9.80	-	95.36%
+HA-DPO§ (Zhao et al., 2023b)	6k	GPT4	7.8	52.1	35.6	4.2	1.89	0.65	54.00	14.45	84.90%	90.42%
+POVID§ (Zhou et al., 2024a)	17k	GPT-4V	7.4	51.3	34.3	3.9	2.08	0.60	50.67	15.28	84.77%	89.01%
+RLAIF-V§ (Yu et al., 2024)	16k	LLaVA-Next	3.0	50.4	16.2	1.0	**3.00**	**0.38**	16.00	3.70	81.57%	94.97%
+SeVa§ (Zhu et al., 2024)	8k	**Self-Supervised**	7.3	54.0	37.3	2.9	2.12	0.57	-	-	86.70%	-
+OPA-DPO† (Xiao et al., 2025)	4.8k	GPT-4V	**2.2**	47.9	11.6	**0.9**	2.83	0.45	13.00	4.25	82.60%	95.61%
+Trident (ours)	**4k**	**Self-Supervised**	**2.2**	53.0	**11.3**	**0.9**	2.92	0.43	**12.81**	**3.68**	**85.10%**	**95.70%**
LLaVA-1.5-13B§ (Liu et al., 2024b,a)			6.8	51.9	31.8	3.3	2.48	0.52	51.00	13.71	85.50%	90.31%
+LLaVA-RLHF§ (Sun et al., 2023b)	122k	Self-Reward	7.7	52.3	38.6	4.0	2.27	0.64	44.67	11.83	82.47%	90.25%
+RLHF-V (HD)† (Yu et al., 2024)	**1.4k**	Human	6.3	46.1	25.1	2.1	2.81	0.49	-	-	-	-
+HSA-DPO† (Xiao et al., 2025)	8k	GPT-4/4V	**2.1**	47.3	13.4	1.2	2.61	0.48	-	-	84.00%	80.20%
+HALVA† (Sarkar et al., 2024)	21.5k	GPT-4V	6.4	**52.6**	30.4	3.2	2.58	0.45	45.40	12.80	-	-
+OPA-DPO† (Yang et al., 2025)	4.8k	GPT4V	2.4	48.3	12.8	**0.9**	3.07	0.39	16.33	5.48	82.63%	96.31%
+Trident (ours)	4k	**Self-Supervised**	**2.1**	52.5	**12.6**	**0.9**	**3.09**	**0.38**	**15.98**	**5.12**	**85.51%**	**96.40%**

for 1 epoch using DeepSpeed ZeRO-3 with BF16 precision and gradient checkpointing. LoRA is applied to the language model with rank 256 and scaling factor 512, while the vision encoder and multimodal projector remain frozen. We use a cosine learning rate schedule initialized at 2×10^{-6}, a batch size of 32, and set $\beta = 0.1$ in the DPO loss. As for the datasets, we curate 4,000 triplets using image-question pairs from OCRVQA (Mishra et al., 2019) and TextVQA (Singh et al., 2019). Additional details are provided in the appendix F.

Baseline Algorithms. We mainly compare our Trident with algorithms based on RLHF/RLAIF. As mentioned in Chapter 1, most algorithms, such as HALVA (Sarkar et al., 2024), POVID (Zhou et al., 2024a), RLHF-V (Yu et al., 2024), HA-DPO (Zhao et al., 2023b), HSA-DPO (Xiao et al., 2025), RLAIF-V (Yu et al., 2024), mDPO (Wang et al., 2024), and OPA-DPO (Yang et al., 2025) prefer to use DPO, while LLaVA-RLHF (Sun et al., 2023b) use PPO. In addition to these supervised or AI-feedback-based approaches, we also include **SeVa** (Zhu et al., 2024), a recent self-supervised preference alignment method which generates chosen and rejected responses by pairing the original image with its augmented counterparts.

Evaluation Benchmarks. We evaluate Trident on four widely used benchmarks for hallucination mitigation in LVLMs: **AMBER** (Wang et al., 2023), **MMHal-Bench** (Sun et al., 2023b), **Object HalBench** (Rohrbach et al., 2018), and **POPE** (Li et al., 2023).

4.2 Benchmark Evaluation Results

Table 1 presents a comprehensive comparison between our proposed Trident method and state-of-the-art preference alignment algorithms across four hallucination evaluation benchmarks. Unlike existing approaches that rely on human annotations (Yu et al., 2024) or high-quality feedback from powerful vision-language models such as GPT-4V (Xiao et al., 2025; Yang et al., 2025), Trident operates in a fully self-supervised manner. It requires neither manually curated preference labels nor external reward models, but instead constructs preference triplets automatically via a semantic augmentation pipeline introduced in Section 3.2. The triplet construction mechanism, which combines enriched chosen responses, degraded rejected responses, and stable reference anchors, provides fine-grained preference signals that surpass simple binary comparisons in effectiveness. Additionally, the adaptive margin regularization in TPR further stabilizes optimization, enabling effective training even with low compute and no human curation.

Despite this lack of external supervision, Trident demonstrates strong competitiveness. On LLaVA-1.5-7B, our method achieves a hallucination rate of 11.3% on AMBER and 0.43 on MMHalBench, outperforming or matching most baselines including those trained with tens of thousands of preference pairs. On the 13B model, Trident reduces the hallucination rate to 12.6% and obtains the best POPE precision of 95.92%, comparable to or exceeding models trained with GPT-4V (Hurst et al., 2024)

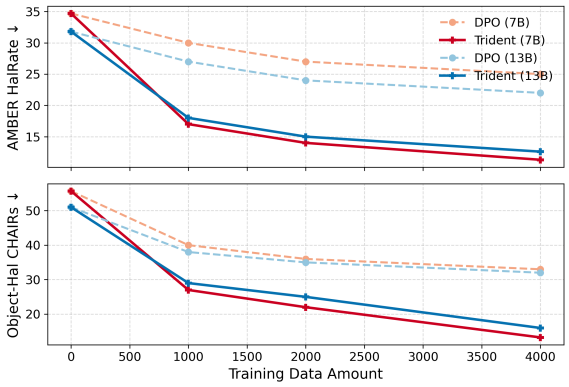

Figure 4: Performance of Trident with varying training data amounts.

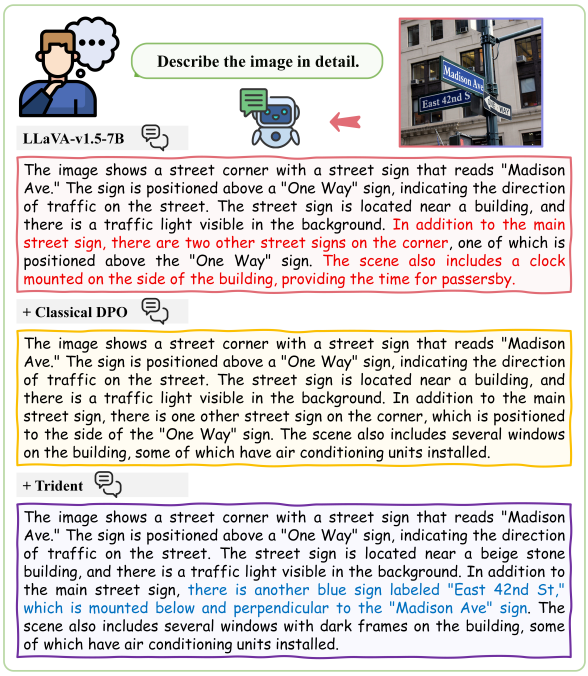

Figure 5: Qualitative example comparing outputs from LLaVA-v1.5-7B, Classical DPO, and Trident under the same image-text input. Hallucinated content is highlighted in red; missing details in blue. Trident offers the most grounded and semantically complete response.

feedback such as OPA-DPO (Yang et al., 2025) and HALVA (Sarkar et al., 2024).

It is worth highlighting that OPA-DPO (Yang et al., 2025), the strongest supervised baseline, is trained for 4 epochs on 4.8K GPT-4V (Hurst et al., 2024) annotated pairs, whereas our Trident only requires 1 epoch on 4K fully self-generated triplets. This underscores both the efficiency and scalability of our approach.

Compared to RLHF-based methods such as LLaVA-RLHF (Sun et al., 2023b), our method consistently outperforms them across hallucination metrics while avoiding the inefficiencies of reward modeling and online rollouts. Furthermore, unlike RLAIF approaches (Xiao et al., 2025; Yu et al., 2025) which rely on large-scale pretrained evaluators like GPT-4V (Hurst et al., 2024) to construct preference data, Trident is fully autonomous, making it particularly well-suited for scenarios where access to high-quality external feedback is limited or costly; in practice, it eliminates proprietary API and annotation expenses and reduces overall training overhead by enabling effective alignment with a lightweight fine-tuning schedule.

In summary, these results validate the core contribution of this work: that high-fidelity preference alignment in LVLMs can be achieved with purely self-supervised triplet construction, without compromising factual accuracy or requiring expensive external signals.

In order to demonstrate the effectiveness of the triplet-based approach, we present its performance across various amounts of training data. As shown in Figure 4, even with as few as 1000 data points, Trident outperforms most baseline algorithms in terms of hallucination-related metrics. Notably, increasing the data size does not lead to significant improvements in the traditional DPO method, while Trident shows substantial performance gains, underscoring its efficiency and scalability in preference alignment.

To further illustrate the benefits of triplet-based alignment over pairwise DPO, we provide a qualitative example in Figure 5. It compares model responses prompted with the same image and instruction. The hallucinated content is marked in red, while missing or incomplete details are highlighted in blue. Additional quantitative evidence is provided in Appendix D.

4.3 Generalization to Recent LVLMs

To evaluate the generality of our approach, we further test Trident on two recently released vision-language models: **Qwen2.5-VL** (Bai et al., 2025) (7B) and **DeepSeek-VL2** (Wu et al., 2024) (27B). We compare each base model's performance before and after alignment using two state-of-the-art hallucination mitigation methods, *SeVa* (Zhu et al., 2024) and *Critic-V* (Zhang et al., 2025), as well as our Trident. As shown in Table 2, Trident consistently achieves the lowest hallucination rates ($CHAIR_s$, $CHAIR_i$) and the highest factual alignment scores (POPE, MMVet) on both mod-

Table 2: Hallucination mitigation results on Qwen2.5-VL and DeepSeek-VL2. Lower $CHAIR_s$ / $CHAIR_i$ indicate fewer hallucinations; higher POPE / MMVet denote stronger factual alignment.

Models	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	POPE ↑	MMVet ↑
Qwen2.5-VL (7B)	37.1	9.4	91.3%	61.8
+ SeVa (Zhu et al., 2024)	21.5	6.2	94.9%	62.4
+ Critic-V (Zhang et al., 2025)	18.1	6.0	95.9%	64.4
+ Trident (Ours)	**10.5**	**3.6**	**96.8%**	**65.2**
DeepSeek-VL2 (27B)	41.3	11.7	88.8%	52.8
+ SeVa (Zhu et al., 2024)	24.2	8.3	93.1%	55.3
+ Critic-V (Zhang et al., 2025)	16.7	8.3	94.1%	56.0
+ Trident (Ours)	**12.5**	**4.7**	**95.6%**	**57.3**

els. These results demonstrate that Trident generalizes robustly to modern LVLM architectures, effectively curbing hallucinations while preserving factual grounding.

4.4 Ablation Studies

To understand the contributions of individual components in our framework, we perform ablation studies on both the 7B and 13B models. Each model is trained on the same 4K self-supervised triplets for one epoch. Evaluations are conducted on the AMBER and Object HalBench datasets using key grounding and hallucination metrics.

Efficacy of Trident Preference Regularization To elucidate the contribution of each component in TPR, we conduct an ablation study by progressively incorporating its submodules: the classical DPO term, the chosen–anchor margin regularization ($\mathcal{L}_{\text{tri}}^{(c)}$), and the rejected–anchor margin regularization ($\mathcal{L}_{\text{tri}}^{(r)}$). Notably, when the anchor-based regularization is entirely removed, the model degenerates into a standard DPO formulation operating on pairs with artificially enlarged semantic gaps, which, as shown in Table 3, yields suboptimal grounding and higher hallucination rates due to the lack of a stabilizing reference. Introducing either regularization term individually leads to marked improvements—$\mathcal{L}_{\text{tri}}^{(c)}$ strengthens semantic separation between chosen responses and the anchor, while $\mathcal{L}_{\text{tri}}^{(r)}$ effectively suppresses degraded responses close to the anchor. The full TPR yields the most substantial reduction in hallucination metrics and enhanced object coverage, confirming that these components act synergistically to enforce structured preference separation and maintain alignment with the pretrained distribution.

Fixed vs. Adaptive Margin To assess the effectiveness of our adaptive margin regularization, we conduct a controlled study on Object HalBench by replacing it with fixed margins $m \in \{0.8, 1.0, 1.2\}$. Table 4 shows that fixed margins are sensitive to scale: smaller margins under-regularize, while larger ones overly penalize minor deviations, both leading to suboptimal grounding. In contrast, the adaptive margin strategy consistently yields better results across CHAIRs and CHAIRi, by dynamically scaling with semantic drift and response uncertainty. See appendix H for additional ablations.

Table 3: Ablation of Trident Preference Regularization components on the AMBER benchmark.

Model	$\mathcal{L}_{\text{tri}}^{(c)}$	$\mathcal{L}_{\text{tri}}^{(r)}$	CHAIR↓	Cover↑	HalRate↓	Cog↓
Trident-7B			3.3	48.2	18.0	2.6
		✓	2.7	50.8	13.4	1.8
	✓		2.8	50.1	13.9	1.9
	✓	✓	**2.2**	**53.0**	**11.3**	**0.9**
Trident-13B			3.4	49.0	17.2	2.5
		✓	2.6	51.2	13.1	1.7
	✓		2.7	50.6	13.5	1.8
	✓	✓	**2.2**	**52.5**	**12.6**	**0.9**

Table 4: Comparison of fixed vs. adaptive margins on Object HalBench.

Model	Margin	CHAIRs↓	CHAIRi↓
Trident-7B	0.8	14.82	5.21
	1.0	13.61	4.63
	1.2	14.47	4.96
	Adaptive	**12.81**	**3.68**
Trident-13B	0.8	17.56	5.98
	1.0	16.40	5.32
	1.2	17.13	5.68
	Adaptive	**15.98**	**5.12**

5 Conclusion

We present Trident, a self-supervised preference alignment framework for vision-language models that extends Direct Preference Optimization to a triplet setting. To overcome the limitations of binary preference data in self-supervised regimes, we propose a structured triplet construction pipeline that systematically generates chosen, rejected, and anchor responses from a single image-question pair using semantic-preserving and semantic-degrading transformations. The resulting triplets enable a more expressive formulation of preference supervision. We further introduce a margin-based regularization mechanism, Trident Preference Regularization, which adaptively controls semantic separation while maintaining alignment with the pretrained model distribution. Despite relying solely on single epoch of fine-tuning on self-generated triplets without human annotations or external reward models, Trident achieves competitive or superior performance to state-of-the-art RLHF and RLAIF baselines across multiple hallucination benchmarks, demonstrating the promise of structured preference modeling under self-supervised settings.

Limitations

Despite the effectiveness of Trident in fully self-supervised preference alignment, several limitations merit further investigation. The current triplet construction procedure relies on a fixed set of static image augmentations, which, while empirically effective, may lack the flexibility required to accommodate the wide variability in visual content and scene complexity encountered in real-world data. Future work could explore adaptive or content-aware augmentation strategies that dynamically tailor transformations to the input image, thereby inducing more precise and informative semantic separation between chosen and rejected responses.

In addition, the quality of self-supervised supervision is inherently bounded by the model's zero-shot ability to follow a fixed self-summarization prompt. Although this design enables fully automated triplet generation, it may introduce prompt sensitivity and limit the expressiveness of the resulting chosen responses. An appealing direction is to adopt an iterative bootstrapping framework in which progressively aligned models are used to regenerate and refine preference triplets, allowing supervision quality to improve over successive iterations without human intervention.

Acknowledgments

This research was supported by STI 2030-Major Projects 2022ZD0208800, in part by NSFC 62088101 Autonomous Intelligent Unmanned Systems, and in part by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F030001.

References

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, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Kejia Chen, Jiawen Zhang, Jiacong Hu, Kewei Gao, Jian Lou, Zunlei Feng, and Mingli Song. 2025. Token-level inference-time alignment for vision-language models. *arXiv preprint arXiv:2510.21794*.

Kejia Chen, Jiawen Zhang, Jiazhen Yang, Mingli Song, and Zunlei Feng. 2026. Self-improved holistic alignment for preference enhancement. *Pattern Recognition*, page 113238.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and 1 others. 2023b. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, and 1 others. 2023. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback. *Preprint*, arXiv:2312.14925.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246.

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

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.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 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. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Data-augmented phrase-level alignment for mitigating object hallucination. *arXiv preprint arXiv:2405.18654*, 2(3):4.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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.

Lina Sun, Yewen Li, and Yumin Dong. 2023a. 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.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023b. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

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*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25543–25551.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 2025. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10610–10620.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 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.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, and 1 others. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.

Xiaotian Yu, Yang Jiang, Tianqi Shi, Zunlei Feng, Yuexuan Wang, Mingli Song, and Li Sun. 2023. How to prevent the continuous damage of noises to model training? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12054–12063.

Arun Zachariah and Praveen Rao. 2023. Video retrieval for everyday scenes with common objects. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 565–570.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, and 1 others. 2025. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9050–9061.

Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. 2023a. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023b. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

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

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024b. Calibrated self-rewarding vision language models. *Advances in Neural Information Processing Systems*.

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

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 291–300.

A Sensitivity of DPO to Supervision Noise

In Section 3.1, we discussed that Direct Preference Optimization (DPO) requires reliable binary preference pairs (y_c, y_r), where the chosen response y_c is assumed to be semantically superior to the rejected y_r. However, in self-supervised settings, this assumption is often violated, leading to noisy or even contradictory supervision. In this appendix, we provide theoretical analysis and controlled experiments to demonstrate how such noise degrades alignment performance, especially in hallucination-prone tasks.

A.1 Gradient Reversal Risk in DPO

The DPO loss for binary preference pairs is defined as:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \left(\beta \left[\log \frac{\pi_\theta(y_c \mid x)}{\pi_{\text{ref}}(y_c \mid x)} - \log \frac{\pi_\theta(y_r \mid x)}{\pi_{\text{ref}}(y_r \mid x)} \right] \right), \tag{15}$$

where π_θ is the fine-tuned policy and π_{ref} is a frozen reference model.

Let us denote:

$$s_c = \log \pi_\theta(y_c \mid x), \quad s_r = \log \pi_\theta(y_r \mid x), \quad \Delta s = s_c - s_r,$$

The gradient with respect to θ becomes:

$$\nabla_\theta \mathcal{L}_{\text{DPO}} \propto \sigma(-\beta \cdot \Delta s) \cdot (\nabla_\theta s_r - \nabla_\theta s_c). \tag{16}$$

If the supervision is correct, this gradient encourages π_θ to assign higher probability to y_c and lower to y_r. However, if the preference label is noisy and y_r is actually better than y_c, the gradient direction becomes misaligned with the true semantic preference. This creates a risk of "negative optimization," where the model is trained to reinforce hallucinated or lower-quality outputs, potentially worsening its performance.

A.2 Summary of Findings

Our theoretical analysis highlights a critical vulnerability of DPO in self-supervised settings where preference data may be noisy or semantically ambiguous. As shown by the gradient formulation, DPO's assumption of clear superiority between chosen and rejected responses makes it susceptible to incorrect supervision. A single flipped or unreliable preference pair can generate misaligned gradients, pushing the model away from the desired behavior. This inherent sensitivity to label noise strongly motivates our shift from standard binary preference optimization to the structured triplet alignment framework in Trident, which incorporates a stabilizing anchor and regularization to ensure more robust learning, as discussed in Section 3.

B Self-Summarization Prompt

To derive coherent and semantically enriched responses from multiple candidates, we employ a structured self-summarization prompt. The prompt is issued to the language model (typically the same SFT model) as follows:

```
Provide a one-sentence caption for the provided image. Note that you need to summarize a more appropriate answer based on the following candidates:
Candidate answer 1: {candidate_answer_1}
Candidate answer 2: {candidate_answer_2}
Candidate answer 3: {candidate_answer_3}
Candidate answer 4: {candidate_answer_4}
Output your summarized one-sentence caption for the provided image below:
```

This prompt is applied uniformly to both the chosen and rejected candidate response sets, ensuring a consolidated and high-signal representation of each semantic category.

C Efficiency Rationale.

Standard DPO on small, noisy datasets often suffers from overfitting to spurious patterns. By incorporating the anchor term $\mathcal{L}_{\text{tri}}$, Trident explicitly regularizes the policy towards the pretrained distribution. This provides a "stabilizing" signal that allows the model to learn robust preferences from as few as **4k triplets** without the catastrophic forgetting or instability typically associated with small-data regimes. This explains why Trident converges in a single epoch while binary DPO variants often require extensive training to mine useful signals.

D Quantitative results to validate y_c better than y_r

To further substantiate the claim that our semantic-preserving transformation set $\mathcal{T}_{\text{chosen}}$ reliably maintains visual semantics while the semantic-degrading set $\mathcal{T}_{\text{rejected}}$ introduces meaningful perturbations, we conducted an external evaluative study using automatically generated preference annotations. Specifically, for each image–question pair drawn from the **TextVQA** and **OCRVQA** datasets, we constructed tuples of the form $\{q, I, y_c, y_r\}$, where y_c and y_r denote the fused responses produced from semantic-preserving and semantic-degrading transformations, respectively, following

the self-summarization procedure described in Section B.

We employed the **GPT-4o-2025-03-26** model as an external judge to assess the relative semantic quality of the two candidate responses. The judge was prompted to select the superior answer according to accuracy, completeness, and relevance to the visual question. fig. 6 shows the exact prompt used in this evaluation.

```
You are a helpful and fair judge evaluating answers to visual questions. Given an image, a question, and two candidate answers – Answer A and Answer B – decide which answer is more accurate, complete, and relevant to the question. Respond with only: "Answer A", "Answer B", or "Tie".

Question: {q, I}
Answer A: {y_c}
Answer B: {y_r}
```

Figure 6: **Judging Prompt for Evaluating Semantic Preservation and Degradation.** The evaluation prompt used to compare fused responses from $\mathcal{T}_{\text{chosen}}$ and $\mathcal{T}_{\text{rejected}}$.

The aggregated win rates across the two datasets are reported in Table 5, demonstrating a strong and consistent preference for responses produced from semantic-preserving transformations. The results reinforce that y_c captures the salient semantic content of the original image–question pair, whereas y_r exhibits weakened or corrupted semantics consistent with our design goals.

Table 5: Win-rate comparison between y_c (semantic-preserving) and y_r (semantic-degrading), judged by GPT-4o-2025-03-26.

Dataset	y_c Win Rate	y_r Win Rate	Tie Rate
TextVQA	95.82%	1.87%	2.31%
OCRVQA	82.46%	5.72%	11.82%

These findings further corroborate the analyses presented in Appendix G. The consistently high preference for y_c verifies that $\mathcal{T}_{\text{chosen}}$ successfully produces semantically faithful variants, whereas $\mathcal{T}_{\text{rejected}}$ effectively degrades visual evidence to construct challenging but unambiguous negative examples. This separation is essential for the reliability of our self-supervised triplet construction pipeline and directly supports the motivation outlined in Section 3.

E Comparison Methods

In this section, we provide brief summaries of the key baseline algorithms discussed in the main paper. These methods represent recent state-of-the-art approaches that leverage preference-based learning, particularly variants of Reinforcement Learning from Human/AI Feedback (RLHF/RLAIF) and Direct Preference Optimization, to mitigate hallucinations in Large Vision-Language Models (LVLMs).

HALVA (Sarkar et al., 2024). Hallucination Attenuated Language and Vision Assistant (HALVA) employs Data-augmented Phrase-level Alignment (DPA) to fine-tune LVLMs. It constructs hallucinated/correct response pairs by modifying ground-truth information at the phrase level. The DPA loss encourages the model to assign lower likelihood to hallucinated segments, mitigating hallucinations while preserving general-purpose multimodal capabilities.

POVID (Zhou et al., 2024a). POVID frames hallucination as a modality alignment issue and applies preference optimization through AI-generated supervision. Preferred responses are derived from ground-truth annotations, while dispreferred responses are constructed by injecting plausible hallucinations using GPT-4V and by applying input image distortions. These synthetic preference pairs are used to fine-tune the model with DPO.

RLHF-V (Yu et al., 2024). RLHF-V enhances model trustworthiness by incorporating fine-grained human correctional feedback. Annotators provide segment-level edits to hallucinated outputs. These corrections are used to construct dense preference signals, enabling behavior alignment through direct preference optimization. Despite using only 1.4K annotated samples, RLHF-V achieves substantial reductions in hallucination rate across multiple benchmarks.

HA-DPO (Zhao et al., 2023b). Hallucination-Aware DPO (HA-DPO) recasts hallucination mitigation as a binary preference learning task. For each image, it constructs pairs of accurate and hallucinated responses and optimizes the model to favor the accurate ones. A pipeline is introduced for generating high-quality, style-consistent hallucination pairs. HA-DPO demonstrates strong gains across multiple MLLMs.

HSA-DPO (Xiao et al., 2025). Hallucination Severity-Aware DPO (HSA-DPO) introduces a

detect-then-rewrite strategy based on sentence-level hallucination detection. A hallucination classifier is trained using sentence-level annotations to localize errors across object, attribute, and relational dimensions. The model is then fine-tuned using preference pairs annotated with severity levels, enabling fine-grained mitigation by weighted preference optimization.

RLAIF-V (Yu et al., 2025). RLAIF-V presents a fully open-source AI feedback framework for hallucination reduction. It constructs both training-time preference data and inference-time feedback using only open-source MLLMs. This paradigm avoids the need for expensive proprietary models or human annotations. Experiments show that RLAIF-V substantially reduces hallucination and can outperform GPT-4V in trustworthiness under certain conditions.

mDPO (Wang et al., 2024). mDPO addresses the unconditional preference issue in multimodal DPO, where models may overfit to language-only preferences and ignore visual input. It introduces an image-aware objective that jointly optimizes language and visual alignment. Additionally, a reward anchor ensures that chosen responses are always positively reinforced, preventing degradation of desired outputs during preference optimization.

OPA-DPO (Yang et al., 2025). On-Policy Alignment DPO (OPA-DPO) highlights the importance of constructing preference pairs aligned with the initial policy distribution. It leverages expert feedback to correct hallucinated responses and ensures both the original and corrected responses remain within the support of the pretrained model. This approach reduces instability caused by KL divergence and achieves state-of-the-art hallucination mitigation with limited data.

SeVa (Zhu et al., 2024). Self-Supervised Visual Preference Alignment (SeVa) introduces a novel unsupervised preference alignment method for Vision-Language Models (VLMs). This approach generates chosen and rejected responses based on both the original and augmented image pairs and conducts preference alignment using Direct Preference Optimization (DPO). SeVa's core idea is that properly designed image augmentations induce the model to generate false but challenging negative responses, enabling the model to learn from and produce more robust answers. The SeVa framework operates without supervision from GPT-4 or human involvement during alignment, making it highly efficient with minimal code. Despite using only 8k randomly sampled unsupervised data, SeVa achieves a 90% relative score to GPT-4 on complex reasoning in LLaVA-Bench and improves LLaVA-7B/13B by 6.7%/5.6% on the MM-Vet benchmark. Visualizations demonstrate its enhanced ability to align with user intentions, and ablation studies highlight its underlying mechanisms and potential for scaling.

F Detailed Training Configuration

To complement the description in Section 4.1, we provide here the full training configuration and hyperparameter setup used for Trident. All experiments were conducted on a server equipped with an Intel Xeon Platinum 8352V Processor and four NVIDIA RTX A6000 GPUs. All models were fine-tuned for a single epoch on 4k self-supervised triplets using DeepSpeed with ZeRO Stage-3 optimization. A detailed breakdown is provided in Table 6.

Table 6: Key hyperparameters for Trident training. These settings were used for both the LLaVA-1.5-7B and 13B models.

Category	Setting
Model Configuration	
Vision Encoder	CLIP ViT-L-336px
Max Sequence Length	2048
Optimizer & Scheduler	
Optimizer	AdamW
Learning Rate	2×10^{-6}
LR Scheduler	Cosine Decay
Weight Decay	0.0
Warmup Steps	0
Training Configuration	
Training Epochs	1
Per-Device Batch Size	16
Gradient Accumulation Steps	1
Total Batch Size	64
Precision	BF16
TF32 Enabled	True
Gradient Checkpointing	True
DPO Beta (β)	0.1
Loss Weight λ_{dpo}	1.0
Loss Weight λ_{tri}	0.05
LoRA Configuration	
LoRA Rank (r)	1024
LoRA Alpha (α)	2048

Table 7: Ablation study on individual data augmentation strategies for triplet construction. Each row shows the performance after removing a single augmentation from the full pipeline. The model is LLaVA-1.5-7B, evaluated on POPE (Adversarial). The results show that both diverse semantic-preserving and strong semantic-degrading augmentations are crucial for optimal performance.

Configuration	Acc.↑	Pre.↑
Ablating from Chosen Augmentations ($\mathcal{T}_{chosen}$)		
w/o Sharpening	84.95%	95.48%
w/o Small-Angle Rotation	85.00%	95.55%
w/o Contrast Enhancement	84.85%	95.31%
w/o Weak Cropping	84.80%	95.20%
Ablating from Rejected Augmentations ($\mathcal{T}_{rejected}$)		
w/o Gaussian Noise	84.65%	94.95%
w/o Horizontal Flipping	84.30%	95.11%
w/o Strong Cropping	84.10%	93.88%
w/o Image Removal	83.90%	93.65%
Full Trident	**85.10%**	**95.70%**

G Ablation Study on Data Augmentation Strategies

To provide a more granular understanding of our self-supervised triplet construction pipeline, we conducted a detailed ablation study on the individual augmentation techniques used to generate the chosen (y_c) and rejected (y_r) responses. This analysis isolates the contribution of each specific transformation, demonstrating that a diverse portfolio of both semantic-preserving and semantic-degrading augmentations is beneficial for creating high-quality training triplets. All experiments were performed on the LLaVA-1.5-7B model, trained on 4k self-generated triplets for one epoch, and evaluated on the POPE (Adversarial) benchmark.

The baseline is the **Full Trident** configuration, which includes all augmentations and the response summarization step. In each subsequent experiment, we remove exactly one augmentation from either the chosen set ($\mathcal{T}_{\text{chosen}}$) or the rejected set ($\mathcal{T}_{\text{rejected}}$) while keeping all other components fixed.

The results, presented in Table 7, show that nearly every augmentation contributes positively to the model's final performance.

For the **chosen responses**, removing any of the weak, semantic-preserving augmentations results in a minor but consistent drop in both accuracy and precision. This suggests that providing the model with varied, high-quality views of the image helps in generating a more robust and comprehensive "chosen" summary, which serves as a stronger positive anchor for alignment.

Table 8: Ablation study on the weight of the triplet regularization term (λ_{tri}) using the POPE (Adversarial) benchmark.

Model	λ_{tri}	POPE Adversarial	
		Accuracy↑	Precision↑
Trident-7B	0.0	84.60%	93.70%
	0.01	84.95%	94.80%
	0.05	**85.10%**	**95.70%**
	0.1	84.89%	95.20%
	0.5	84.72%	94.20%
Trident-13B	0.0	84.72%	94.20%
	0.01	85.25%	95.50%
	0.05	**85.51%**	**96.40%**
	0.1	85.19%	96.00%
	0.5	84.90%	95.00%

For the **rejected responses**, the impact is more pronounced, especially when removing the most severe transformations. Eliminating "Image Removal" or "Strong Cropping" causes the most significant performance degradation. This validates our core hypothesis that maximizing the semantic distance between chosen and rejected responses is critical for effective learning in a self-supervised setting. A clearly degraded and factually inconsistent rejected response provides a much stronger and less ambiguous training signal. Less severe degradations like "Gaussian Noise" and "Horizontal Flipping" also contribute, and their removal still leads to a noticeable decline in performance.

In summary, this granular ablation confirms that our curated sets of diverse augmentation strategies are not redundant. Each transformation plays a role in constructing effective triplets, and the synergy between them leads to the superior performance of the full Trident framework.

H Ablation of Loss Weighting

The final Trident objective function combines the standard DPO loss ($\mathcal{L}_{\text{DPO}}$) with our proposed Trident Preference Regularization ($\mathcal{L}_{\text{tri}}$) using two weighting hyperparameters, λ_{dpo} and λ_{tri}:

$$\mathcal{L}_{\text{Trident}} = \lambda_{\text{dpo}} \cdot \mathcal{L}_{\text{DPO}} + \lambda_{\text{tri}} \cdot \mathcal{L}_{\text{tri}}. \quad (17)$$

These weights are crucial as they balance the primary goal of learning preferences between chosen and rejected responses against the secondary

goal of maintaining structured semantic separation with respect to the anchor response. To investigate the sensitivity of our framework to these hyperparameters and validate our chosen configuration ($\lambda_{\text{dpo}} = 1.0, \lambda_{\text{tri}} = 0.05$), we conduct an ablation study by varying the weight of the triplet regularization term, λ_{tri}, while keeping the DPO weight fixed at $\lambda_{\text{dpo}} = 1.0$.

The results, presented in Table 8, demonstrate the impact of this balance on object-level hallucination metrics for both the 7B and 13B models on the POPE benchmark, while Figure 7 provides a visual illustration of the trend.

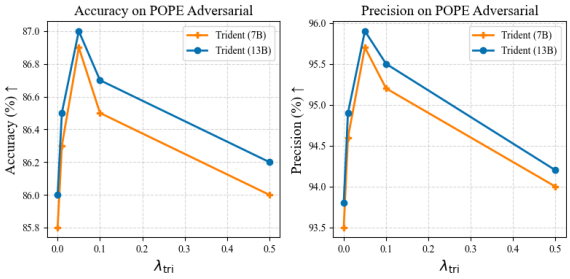

Figure 7: **Ablation study on the triplet regularization weight $\boldsymbol{\lambda_{\text{tri}}}$.** A clear peak in both accuracy and precision is observed at $\lambda_{tri} = 0.05$, indicating this is the optimal trade-off point. Performance declines with a higher weight, confirming the term's role as a regularizer.

Analysis. The results clearly indicate the importance of the triplet regularization term for mitigating object hallucination. When $\boldsymbol{\lambda_{\text{tri}} = 0.0}$, the model is equivalent to a standard DPO trained on our augmented triplet data (using only the chosen and rejected pair). While this baseline already improves upon the base SFT model, it yields the lowest accuracy and precision among the tested configurations. This finding suggests that simply enforcing a preference between an enriched and a degraded response is insufficient to achieve optimal factuality without the stabilizing influence of the anchor regularization.

As we introduce and increase the weight of $\mathcal{L}_{\text{tri}}$, we observe a consistent improvement in both accuracy and precision. Performance peaks at $\boldsymbol{\lambda_{\text{tri}} = 0.05}$, where the model achieves the highest scores. This setting appears to strike an optimal balance, allowing the DPO objective to effectively drive preference learning while the TPR term provides sufficient regularization to structure the semantic space, prevent drift, and ground the model's outputs in visual reality.

However, increasing the weight further to $\lambda_{\text{tri}} = 0.1$ and $\lambda_{\text{tri}} = 0.5$ leads to a degradation in performance. This suggests that an overly strong regularization penalty can overpower the primary DPO signal, potentially causing the model to become too conservative by adhering too closely to the anchor response, thereby hindering its ability to learn the fine-grained preference between the chosen and rejected examples.

Based on this analysis, we selected $\lambda_{\text{dpo}} = 1.0$ and $\lambda_{\text{tri}} = 0.05$ as the optimal configuration for all experiments reported in the main paper, as it consistently delivers the best performance across both model sizes and evaluation benchmarks.