Image Difference Captioning via Adversarial Preference Optimization

Zihan Huang[1]*, Junda Wu[1]*, Rohan Surana[1], Tong Yu[2], David Arbour[2], Ritwik Sinha[2], Julian McAuley[1]
[1]UC San Diego, [2]Adobe Research,
{zih043, juw069, rsurana, jmcauley}@ucsd.edu, {tyu, arbour, risinha}@adobe.com

Abstract

Image Difference Captioning (IDC) aims to generate natural language descriptions that highlight subtle differences between two visually similar images. While recent advances leverage pre-trained vision-language models to align fine-grained visual differences with textual semantics, existing supervised approaches often overly focus on dataset-specific language patterns and fail to capture fine-grained and context-aware preferences on IDC, due to limited annotation diversity and a lack of semantically informative negative examples during training, To address these limitations, we propose an *adversarial direct preference optimization* (ADPO) framework for IDC, which formulates IDC as a preference optimization problem under the Bradley-Terry-Luce model, directly aligning the captioning policy with pairwise difference preferences via Direct Preference Optimization (DPO). To model more accurate and diverse IDC preferences, we introduce an adversarially trained hard negative retriever that selects counterfactual captions, This results in a minimax optimization problem, which we solve via policy-gradient reinforcement learning, enabling the policy and retriever to improve jointly. By dynamically generating semantically challenging negatives, our method reduces reliance on dataset-specific patterns. Experiments on benchmark IDC datasets show that our approach outperforms existing baselines, especially in generating fine-grained and accurate difference descriptions.

1 Introduction

Image Difference Captioning (IDC) requires a model to generate natural-language descriptions that highlight salient differences between a pair of visually similar images. It serves as a core capability in applications such as visual quality inspection (Jhamtani and Berg-Kirkpatrick, 2018), image editing feedback (Tan et al., 2019), and fine-grained visual understanding for multimodal agents (Zhang et al., 2024; Wu et al., 2025b; Wang et al., 2025). Unlike generic image captioning (Xu et al., 2015; Stefanini et al., 2022), which focuses on holistic scene descriptions, IDC requires the model to precisely isolate and describe subtle, localized changes while preserving shared content.

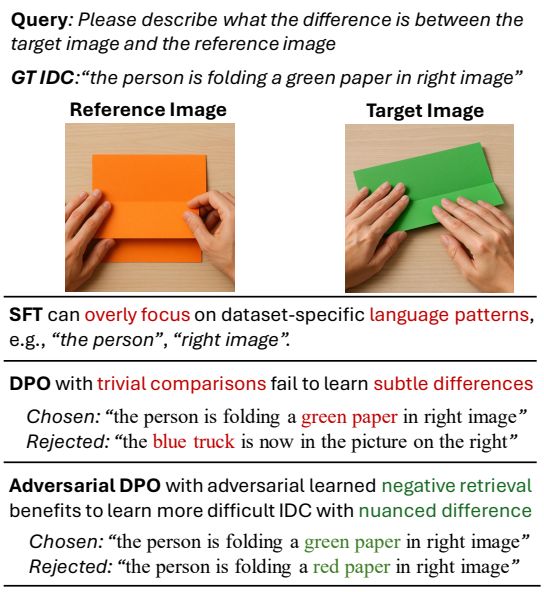


Figure 1: Comparison among supervised fine-tuning (SFT), conventional direct preference optimization (DPO), and our proposed adversarial DPO (ADPO). The paper-folding images are AI-generated for illustrative purposes.

Existing IDC systems are mostly trained under supervised fine-tuning over human captions (Jhamtani and Berg-Kirkpatrick, 2018; Park et al., 2019a), advanced through multimodal instruction tuning (Liu et al., 2023; Wu et al., 2024a), alignment (Zhang et al., 2025; Wu et al., 2024b, 2025c), and test-time inference (Chen et al., 2024; Wu et al., 2024d). Recent approaches incorporate vision–language pre-training and contrastive ob-

*These authors contributed equally.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33747–33759
November 4-9, 2025

jectives (Hao et al., 2022; Liu et al., 2023; Yan et al., 2024). However, existing methods potentially suffer from the challenges of (i) overly focus on dataset-specific language patterns, (ii) confusion between global scene semantics and local edits, and (ii) a lack of exposure to fine-grained and context-aware differences. Collecting extra human annotations to repair these deficiencies is prohibitively expensive. Adapted from Souček et al. (2022), Figure 1 illustrates these issues through an example with two images[1]: SFT tends to overly focus on dataset-specific expressions like "people standing" or "right image", rather than the underlying semantic change. Conventional DPO, when trained with semantically irrelevant negatives (e.g., "blue truck"), fails to capture subtle yet meaningful differences in count or spatial context. In contrast, our proposed adversarial DPO introduces more targeted hard negatives (e.g., "less people in parking lot") that challenge the model to disambiguate nuanced edits, leading to better alignment with ground-truth IDC.

To overcome these challenges, we formulate IDC as a direct preference optimization (DPO) problem (Rafailov et al., 2023; Wu et al., 2025a; Xie et al., 2025). Specifically in IDC, given two candidate captions, preference signals capture the visual difference and provide feedback with pairwise comparison, even when an absolute numeric reward is unavailable. DPO converts such comparisons into a likelihood objective under the Bradley-Terry-Luce (BTL) model, avoiding unstable reward models typical of RLHF pipelines (Ouyang et al., 2022a). However, DPO alone inherits the quality of negative captions supplied during training.

To further enable robust IDC preference alignment, we introduce **A**dversarial **D**irect **P**reference **O**ptimization (**ADPO**), a minimax framework that couples a captioning policy with an *adversarial hard-negative retriever*. At each iteration, the retriever proposes *counterfactual* (Liu et al., 2025; Wu et al., 2024c) captions that are semantically close to the ground truth yet subtly incorrect. The policy then learns via the DPO loss, to prefer positives over these informative negatives. We solve the resulting game using an efficient policy-gradient algorithm that alternates between closed-form DPO updates for the policy and REINFORCE updates for the retriever.

We summarize our contributions as follows,

- We formulate IDC as pairwise preference learning by a novel preference model.
- We propose *ADPO*, a minimax learning framework that jointly trains a captioning policy and an adversarial hard-negative retriever.
- We derive a minimax learning objective with closed-form DPO policy updates with policy-gradient retriever updates.

2 Related Work

Image Difference Captioning Image Difference Captioning (IDC) aims to generate descriptions of subtle distinctions between similar images. Early supervised methods such as CLIP4IDC (Guo et al., 2022) and IDC-PCL (Yao et al., 2022a) leverage pretrain-finetune frameworks to bridge domain gaps and improve IDC performance. Souček et al. (2022) introduce a dataset of untrimmed web videos for object states and state-modifying actions, and address their temporal localization with minimal supervision. Recent work expands IDC capabilities with new frameworks and data strategies. VisDiff (Dunlap et al., 2024) highlights set-level reasoning over multiple images, while BLIP2IDC (Evennou et al., 2025) uses synthetic augmentation to mitigate data scarcity. DIRL (Tu et al., 2024) improves robustness against distractors, and FINER-MLLM (Zhang et al., 2024) applies LoRA tuning to enhance change captioning. OneDiff (Hu et al., 2024) adopts a siamese encoder with a visual delta module for fine-grained difference detection. While these advancements have significantly improved IDC methodologies, challenges remain, particularly in capturing fine-grained distinctions and maintaining data diversity. Addressing these issues will require a concerted effort toward developing sophisticated models that leverage cross-modal learning and scalable dataset generation techniques, as highlighted in previous studies (Yao et al., 2022b).

Direct Policy Optimization Direct Policy Optimization (DPO) has emerged as a pivotal framework for aligning models with human preferences, streamlining the optimization process compared to conventional reinforcement learning approaches. Rafailov et al. (2023) introduced DPO

[1]In Figures 1 and 2, the paper-folding images were generated using OpenAI's ChatGPT (GPT-5, using the image generation feature, September 2025). The prompts were: "Two hands folding an orange paper." and "Two hands folding a green paper." These images are solely for illustrative purposes.

as a binary classification task, which simplifies fine-tuning large language models (LLMs) without the complexities of sampling and hyperparameter tuning. This framework has been effectively extended to multimodal contexts. For instance, CHiP (Fu et al., 2025) integrates visual and textual preferences, enhancing the model's ability to discern hallucinations from accurate descriptions. MDPO (Wang et al., 2024) addresses the unconditional preference problem by optimizing both image and language preferences, significantly improving performance in multimodal scenarios. Furthermore, S-VCO (Wu et al., 2025d) introduces a finetuning objective that aligns visual details with text, thereby reducing hallucinations. DAMA (Lu et al., 2025) employs dynamic adjustments based on data hardness and model responsiveness, resulting in enhanced alignment performance across benchmarks. The advancements underscore the critical role of DPO in refining multimodal alignment and preference learning.

3 Preliminary

3.1 Direct Preference Alignment

Direct Preference Optimization (Rafailov et al., 2023) provides a method for reinforcement learning from human feedback (Ouyang et al., 2022b) that does not require explicit reward modeling. DPO builds on the Bradley-Terry-Luce (BTL) model, which defines the probability of preferring one response over another using a sigmoid function applied to their respective reward differences:

$$p^*(y_1 \succ y_2 \mid x) = \sigma\left(r(x, y_1) - r(x, y_2)\right), \quad (1)$$

where $\sigma(z) = 1/(1 + \exp[-z])$.

Instead of modeling rewards directly, DPO implicitly aligns the policy with pairwise human preferences by maximizing the likelihood of preferred responses under the BTL model. The resulting DPO loss is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_1,y_2)\sim D}[\log \sigma(\beta \log \frac{\pi_\theta(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)} - \beta \log \frac{\pi_\theta(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)})],$$

where the normalization constants cancel out. This objective directly optimizes the target policy π_θ to reflect human preference data, bypassing the need for reward regression.

3.2 Image Difference Captioning

Image Difference Captioning (IDC) aims to generate a natural language description that captures the subtle differences between two similar images. Formally, given an image pair (X_1, X_2), where $X_1, X_2 \in \mathbb{R}^{H\times W\times C}$, the objective is to learn a mapping

$$\pi : (X_1, X_2) \to T,$$

where $T = \{t_1, t_2, \ldots, t_n\}$ is a sequence of tokens forming a descriptive caption that highlights the differences between the images.

4 Method

IDC is uniquely challenging due to the need to capture subtle, localized changes rather than general scene descriptions. This challenge is underscored by our analysis of the IDC dataset, as illustrated in Figure 3, which reveals that the vast majority of image pairs exhibit high similarity. We assessed the degree of difference for 100 randomly selected image pairs on a scale from *1 (completely different)* to *5 (almost identical)*, using both human evaluators and GPT-4o. Over 88% of the pairs were rated as *4 (very similar)* or *5 (almost identical)* by both humans and GPT-4o, with average scores of 4.12 and 4.58 respectively. The prevalence of such highly similar pairs underscores the difficulty of capturing subtle, localized variations within the IDC data. Existing methods often miss these fine distinctions and focus on general language patterns. To address this, we first formulate the task as IDC preference optimization (Section 4.1). Then, we introduce an adversarial preference optimization framework that leverages hard negative retrieval (Section 4.2 and 4.3), explicitly pushing the model to focus on nuanced visual differences. To solve the minimax learning problem, we propose policy-gradient optimization strategy to enable joint optimization of both policies (Section 4.4). We illustrate our framework in Figure 2.

4.1 IDC Preference Optimization Objective

We aim to optimize an image difference captioning (IDC) policy that generates captions describing the salient differences between a given image pair $Q = (X_1, X_2)$. To this end, we formulate the task using the Bradley-Terry-Luce (BTL) model, which compares two textual responses T and T' conditioned on the same image pair. The probability that

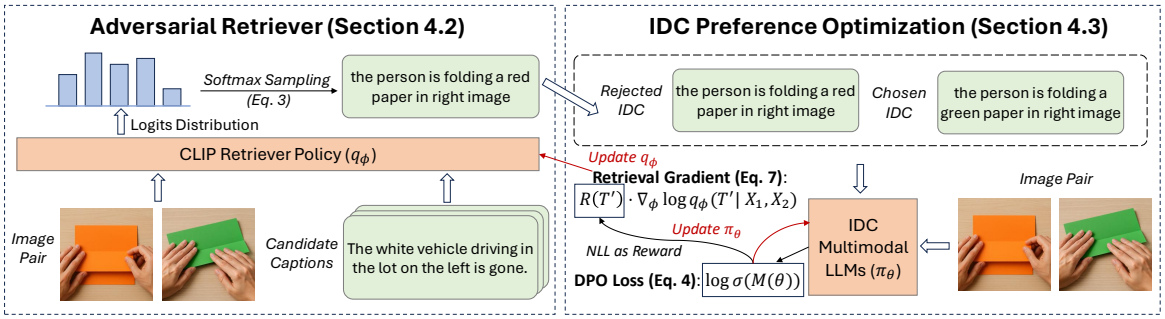


Figure 2: Overview and illustration of our proposed ADPO framework.

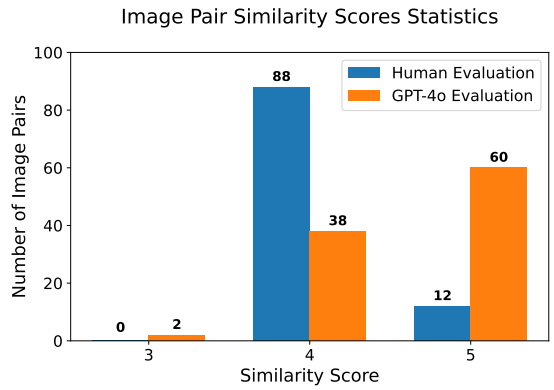


Figure 3: Distribution of similarity scores for 100 random image pairs from the IDC (Spot-the-Diff) dataset, as evaluated by humans and GPT-4o. A score of 5 indicates "Almost Identical" images. The prevalence of high scores (4 and 5) shows that discerning subtle differences is common in IDC task.

T is preferred over T' is defined as:

$$p^*(T \succ T' \mid X_1, X_2) = \sigma\Big(r(X_1, X_2, T) - r(X_1, X_2, T')\Big),$$

where σ denotes the sigmoid function, and $r(X_1, X_2, T)$ is a reward model that scores the quality or informativeness of caption T given the image pair. Here, T' is an alternative caption (i.e., a negative sample) for the same image pair.

We seek to learn a policy π_θ that assigns higher probabilities to better captions under this pairwise preference. To encourage divergence from a fixed reference policy π_{ref} while aligning with human preferences, we define the margin-based score:

$$M(\theta) = \beta \log \frac{\pi_\theta(T \mid X_1, X_2)}{\pi_{\text{ref}}(T \mid X_1, X_2)} - \beta \log \frac{\pi_\theta(T' \mid X_1, X_2)}{\pi_{\text{ref}}(T' \mid X_1, X_2)}, \quad (2)$$

where β is a temperature scaling parameter controlling the sharpness of the preference.

Using this margin, we define the IDC policy learning objective as:

$$\mathcal{L}_{\text{IDC}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(X_1, X_2, T, T') \sim D}\big[\log \sigma(M(\theta))\big],$$

where D is a dataset of quadruples containing image pairs and corresponding positive and negative captions. This objective encourages π_θ to favor captions aligned with the implicit reward function learned from preferences.

4.2 IDC Negative Sampling

The effectiveness of the IDC training objective hinges on the quality of negative samples T', which ideally should be similar to T but less informative or accurate. Poor or irrelevant negatives can make the learning signal weak or noisy.

To address this, we adopt a hard negative sampling strategy inspired by CLIP2IDC. Specifically, we introduce a retriever model q_ϕ that selects semantically similar but suboptimal captions:

$$T'(\phi) \sim q_\phi(\cdot \mid X_1, X_2). \quad (3)$$

The retriever q_ϕ is parameterized to select high-quality negative captions from a candidate pool conditioned on the input image pair. The goal is to find challenging negatives that force the policy to learn more discriminative representations of image differences. Once a negative caption T' is selected using q_ϕ, it is paired with the positive caption T from the dataset to construct training tuples.

4.3 Adversarial Learning Objective

Rather than fixing the negative sample retriever, we propose a joint learning framework where the policy π_θ and the retriever q_ϕ are trained adversarially. In this setting, q_ϕ acts as an adversary trying

to find the hardest negatives that maximize the loss for the policy, improving the sample efficiency and robustness of the learning process. We redefine the margin with the learned negative sample $T'(\phi)$:

$$M(\theta, \phi) = \beta \log \frac{\pi_\theta(T \mid X_1, X_2)}{\pi_{\text{ref}}(T \mid X_1, X_2)} - \beta \log \frac{\pi_\theta(T'(\phi) \mid X_1, X_2)}{\pi_{\text{ref}}(T'(\phi) \mid X_1, X_2)}. \quad (4)$$

Plugging this into the IDC loss yields the adversarial learning objective:

$$\mathcal{L}_{\text{IDC-A}}(\pi_\theta; q_\phi) = - \mathbb{E}_{(X_1, X_2, T) \sim D,\, T' \sim q_\phi(\cdot \mid X_1, X_2)} \Big[\log \sigma(M(\theta, \phi)) \Big]. \quad (5)$$

This leads to a minimax optimization problem, where the policy minimizes the adversarial loss while the retriever maximizes it, encouraging the discovery of progressively harder negatives over training.

4.4 Minimax Optimization

The adversarial objective in IDC policy optimization is given by:

$$(\theta^*, \phi^*) = \arg \max_\phi \min_\theta \mathcal{L}_{\text{IDC-A}}(\pi_\theta; \pi_{\text{ref}}; q_\phi), \quad (6)$$

where the policy π_θ aims to minimize the IDC-A loss by increasing the preference score of the positive caption T over the sampled negative T', and the retriever q_ϕ learns to select hard negatives that maximize the loss.

Policy Update of θ. Given a mini-batch of training examples and corresponding negatives sampled from the retriever $T' \sim q_\phi(\cdot \mid X_1, X_2)$, we compute the policy gradient by backpropagating through the IDC-A loss in Eq. 5. The policy parameters θ are updated using gradient descent:

$$\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\text{IDC-A}}(\theta, \phi).$$

Retriever Update of ϕ via REINFORCE. Since T' is sampled from a discrete distribution $q_\phi(\cdot \mid X_1, X_2)$, we cannot backpropagate through the sample. Instead, we optimize ϕ using the REINFORCE estimator. Let $R(T') = -\log \sigma(M(\theta, \phi))$ be the reward signal. The gradient for retriever parameters is:

$$\nabla_\phi \mathbb{E}_{T' \sim q_\phi} \Big[R(T') \Big] = \mathbb{E}_{T' \sim q_\phi} \Big[R(T') \cdot \nabla_\phi \log q_\phi(T' \mid X_1, X_2) \Big]. \quad (7)$$

We approximate this expectation using a Monte Carlo sample for each training instance. Optionally, a baseline can be subtracted from the reward to reduce variance. We illustrate the algorithm in Algorithm 1.

Algorithm 1 Policy-gradient updates for joint captioning and retrieval policy learning

Require: Dataset D, initial policy θ_0, retriever ϕ_0, reference policy π_{ref}, learning rates η_θ, η_ϕ

for iteration $k = 0, 1, 2, \ldots$ until convergence **do**
  Sample mini-batch $\mathcal{B} \subset D$
  for each $(X_1, X_2, T) \in \mathcal{B}$ **do**
    Sample negative $T' \sim q_{\phi_k}(\cdot \mid X_1, X_2)$
    Compute margin: $M(\theta, \phi)$ in Eq. (4)
    Compute reward: $R(T') = -\log \sigma(M)$
    Accumulate policy loss:

$$\mathcal{L}_\theta \leftarrow \mathcal{L}_\theta + \log \sigma(M)$$

    Accumulate retriever gradient:

$$\nabla_\phi \mathcal{L}_\phi \leftarrow \nabla_\phi \mathcal{L}_\phi + R(T') \cdot \nabla_\phi \log q_{\phi_k}(T' \mid X_1, X_2)$$

  end for
  Update policy: $\theta_{k+1} \leftarrow \theta_k - \eta_\theta \nabla_\theta \mathcal{L}_\theta$
  Update retriever: $\phi_{k+1} \leftarrow \phi_k - \eta_\phi \nabla_\phi \mathcal{L}_\phi$
end for

5 Experiment

Dataset Following previous IDC works (Zhang et al., 2024; Guo et al., 2022; Hu et al., 2023), we evaluate the performance of models using 3 popular image difference captioning datasets, CLEVR-Change (Park et al., 2019b), Spot-the-Diff (Jhamtani and Berg-Kirkpatrick, 2018) and Image-Editing-Request (Tan et al., 2019).

Implementation Details We adopt Qwen2.5-VL-3B-Instruct (Yang et al., 2024; Team, 2025) as our baseline multimodal large language model (MLLM). We trained our model with Adam Optimizer and FSDPTrainer in a two-phase pipeline, SFT tuning and RL tuning, on 2 A6000 GPUs.

In SFT tuning, we tune Qwen2.5 in float32, with learning rate set as 5e-7 and warmup steps as 150. The batch size is set as 16, with a gradient accumulation step of 8. In the RL tuning, we further tune model based on the SFT fine-tuned model through traditional DPO and our proposed ADPO in bf16.

In this stage, the learning rate of Qwen is set within a range of 3e-6 to 5e-6 and ADPO retriever as 1e-4. A larger batch size range from 24 to 32 is applied in RL stage to improve training stability.

Evaluation Metrics Following previous IDC works (Huang et al., 2022; Guo et al., 2022), we use the standard evaluation protocol to evaluate the image difference captioning quality with metrics including BLEU-4 (B4) (Papineni et al., 2002), METEOR (M) (Banerjee and Lavie, 2005), ROUGE-L (R) (Lin, 2004) and CIDEr-D (C) (Vedantam et al., 2015).

Baselines We compare ADPO fine-tuned Qwen2.5 VL 3b with SFT (Ouyang et al., 2022a) fine-tuned Qwen2.5 VL 3b and zero-shot Qwen2.5 VL 3b baseline. Besides, we compare ADPO fine-tuned Qwen2.5 VL 3b with various IDC baselines which can be roughly categorized into five main groups.

- **Attention-based Methods:** dual attention approaches DUDA (Park et al., 2019b) and hybrid attention-reinforcement learning methods VAM (Shi et al., 2020), IFDC (Huang et al., 2022) and VACC (Kim et al., 2021a). These methods focus on learning spatial relationships with attention mechanisms.

- **Reinforcement Learning Approaches:** Comprising SRDRL (Tu et al., 2021b), R^3Net (Tu et al., 2021a), BiDiff (Sun et al., 2022), and SCORER (Tu et al., 2023c), which employ various reinforcement learning paradigms for decision-making in IDC.

- **Representation Learning Methods:** Including prototype-based IDC-PCL (Yao et al., 2022a), variational autoencoder-based VARD (Tu et al., 2023a), and noise-tolerant approaches NCT (Tu et al., 2023b) that focus on learning robust feature representations.

- **Pixel-level Alignment Methods:** Such as DDLA (Kim et al., 2021b), which perform clustering-based pixel alignment.

- **Pretrained Model-based Approaches:** Including CLIP with two-stage fine-tuning based CLIP4IDC (Guo et al., 2022) and MLLM-based FMLLM (Zhang et al., 2024), leveraging large pretrained models for IDC.

LLM Usage In this paper, LLMs are only used for refining the writing of natural language.

5.1 Results

Results on Image-Editing-Request. The Image-Editing-Request dataset focuses on generating captions for multiple differences between image pairs, requiring models to generate various captions to describe different variations between image pairs. We employ this dataset to evaluate a model's capability in distinguishing and describing multiple differences between image pairs. It is challenging for even state-of-the-art models to completely caption various differences in two images, particularly when dealing with visually subtle distinctions. As shown in Table 1, we evaluate the effectiveness of ADPO on this multi-difference captioning task. Our results demonstrate that ADPO achieves consistent improvements over both SFT and standard DPO, with a CIDEr score gain of 3.09% and 1.01%, respectively. Notably, the Qwen 2.5 VL 3B model fine-tuned with ADPO outperforms all other IDC baselines on this dataset. This improvement suggests that ADPO enhances model's ability to identify and caption multiple differences by exposing it to carefully constructed negative captions that highlight semantic details highly related to each image pairs, thereby encouraging more comprehensive analysis of subtle changes.

Result on CLEVR-change. CLEVR-Change consists of synthetic geometric image pairs containing subtle semantic variations in object color, shape, and position, along with visual distractors like perspective shifts. This dataset presents the unique challenge of determining whether visual differences correspond to actual semantic changes. Table 2 presents our evaluation of ADPO's ability to improve model performance on geometric shape difference captioning. ADPO consistently surpasses both SFT and DPO, achieving CIDEr improvements of 2.95% and 0.7%, respectively. Besides, ADPO-tuned Qwen 2.5 VL 3B outperforms all baselines on CLEVR-Change, achieving superior performance on more than half of the evaluation metrics. These results indicate that ADPO's retrieval of strategically confusing negative captions, such as those misattributing perspective changes to semantic differences, effectively enhances model performance on synthetic geometric IDC.

Result on Spot-the-Diff. Spot-the-Diff is a real-world IDC dataset containing temporal image pairs that may or may not contain semantic differences at specific locations. This dataset tests models' ability

Table 1: IDC evaluation results on Image-Editing-Request.

Method	B4	M	R	C
DUDA (Park et al., 2019b)	6.5	12.4	37.3	22.8
BiDiff (Sun et al., 2022)	6.9	14.6	38.5	27.7
CLIP4IDC (Guo et al., 2022)	8.2	14.6	40.4	32.2
NCT (Tu et al., 2023b)	8.1	15.0	38.8	34.2
VARD (Tu et al., 2023a)	10.0	14.8	39.0	35.7
SCORER (Tu et al., 2023c)	10.0	15.0	39.6	33.4
FMLLM (Zhang et al., 2024)	13.3	14.6	39.6	50.5
Qwen-2.5-VL	3.8	15.0	28.6	13.7
ADPO	**13.9**	17.4	**42.0**	**60.0**
w/o Preference Optimization	13.1	17.1	41.4	58.2
w/o Retrieval Adversarial	13.8	**17.5**	41.9	59.4

Table 2: IDC evaluation results on CLEVR-change.

Method	B4	M	R	C
DUDA (Park et al., 2019b)	47.3	33.9	-	112.3
VAM (Shi et al., 2020)	50.3	37.0	69.7	114.9
VAM+ (Shi et al., 2020)	51.3	37.8	70.4	115.8
IFDC (Huang et al., 2022)	49.2	32.5	69.1	118.7
DUDA+Aux	51.2	37.7	70.5	115.4
VACC (Kim et al., 2021a)	52.4	37.5	-	114.2
SRDRL (Tu et al., 2021b)	**54.9**	**40.2**	73.3	122.2
R^3Net (Tu et al., 2021a)	**54.7**	**39.8**	73.1	123.0
BiDiff (Sun et al., 2022)	54.2	38.3	-	118.1
IDC-PCL (Yao et al., 2022a)	51.2	36.2	71.7	128.9
CLIP4IDC (Guo et al., 2022)	**56.9**	38.4	76.4	150.7
NCT (Tu et al., 2023b)	**55.1**	**40.2**	73.8	124.1
VARD (Tu et al., 2023a)	**55.2**	**40.8**	74.1	124.1
SCORER (Tu et al., 2023c)	**56.3**	**41.2**	74.5	126.8
FMLLM (Zhang et al., 2024)	**55.6**	36.6	72.5	137.2
Qwen2.5-VL	4.5	17.1	50.7	90.5
ADPO	**54.3**	**38.5**	**78.4**	**153.6**
w/o Preference Optimization	54.2	37.6	76.4	149.2
w/o Retrieval Adversarial	53.0	38.4	78.2	152.5

to detect subtle semantic variations among diverse real-world objects. As shown in Table 3, ADPO demonstrates significant improvements over conventional SFT and DPO approaches, with CIDEr score increases of 7.47% and 1.82%, respectively. However, we observe that Qwen 2.5 VL fine-tuned with ADPO exhibits suboptimal performance to FMLLM on Spot-the-Diff. We attribute this to two factors: (1) FMLLM's parameter advantage (7B vs. 3B) may enhance its IDC performance on real-world IDC tasks, and (2) Spot-the-Diff's real-world image distribution aligns closely with Vicuna's pre-training data. Nevertheless, ADPO's substantial performance gains over other training paradigms highlight its effectiveness in improving representation learning efficiency for real-world IDC tasks during preference optimization.

Table 3: IDC evaluation results on Spot-the-Diff.

Method	B4	M	R	C
DDLA (Kim et al., 2021b)	8.5	12.0	28.6	32.8
DUDA (Park et al., 2019b)	8.1	11.8	29.1	32.5
VAM (Shi et al., 2020)	10.1	12.4	31.3	38.1
VAM+ (Shi et al., 2020)	**11.1**	12.9	33.2	42.5
IFDC (Huang et al., 2022)	8.7	11.7	30.2	37.0
DUDA+Aux	8.1	12.5	29.9	34.5
VACC (Kim et al., 2021a)	9.7	12.6	32.1	41.5
SRDRL (Tu et al., 2021b)	-	13.0	31.0	35.3
R^3Net (Tu et al., 2021a)	-	13.1	32.6	36.6
BiDiff (Sun et al., 2022)	6.6	10.6	29.5	42.2
CLIP4IDC (Guo et al., 2022)	**11.6**	14.2	**35.0**	47.4
VARD (Tu et al., 2023a)	-	12.5	29.3	30.3
SCORER (Tu et al., 2023c)	10.2	12.2	-	38.9
FMLLM (Zhang et al., 2024)	**12.9**	**14.7**	**35.5**	**61.8**
Qwen-2.5-VL	4.0	13.3	23.9	16.6
ADPO	**10.7**	**14.7**	**34.2**	**56.1**
w/o Preference Optimization	10.4	14.0	32.2	52.2
w/o Retrieval Adversarial	10.6	14.7	34.1	55.1

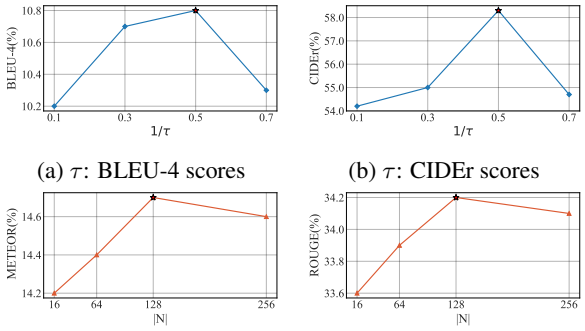
6 Analysis

6.1 Ablation Study

In this section, we study the ablation of ADPO retriever module and the influence of two key parameters, namely the temperature (τ) and the negative candidate set size $|N|$, on the ADPO performance, as are shown in Figure 4. τ is used to adjust the Logits Distribution (in Figure 2) of the similarity-based retriever, controlling the sharpness of the output probability distribution over candidate items, which in turn influence the randomness of deciding final retrieved negative caption. $|N|$ represents the size of a randomly selected candidate set from the dataset prior to calculating Logits Distribution by ADPO retriever, serving as a parameter that influences the randomness of the ADPO retriever's outputs. Smaller $|N|$ introduces higher randomness to the ADPO process.

Ablation Study on ADPO retriever. As is shown in Table 1, Table 2 and Table 3 ablating the modules leads to significant performance degradation across all evaluated datasets. Without SFT, zero-shot Qwen fails to follow captioning instructions. The introduction of SFT brings performance improvements, placing Qwen at a medium level among all IDC baselines. However, at this stage, the model still struggles with challenging IDC samples containing subtle image differences. The subsequent application of the DPO module yields further performance gains, demonstrating that conventional DPO enhances the model's capability to caption images with fine-grained visual distinctions. Never-

theless, Qwen remains inadequate for particularly difficult cases involving extremely subtle differences and visually ambiguous content. Finally, by incorporating ADPO, which emphasizes learning from high-quality, confusing positive-negative caption pairs, we observe substantial improvements, elevating Qwen's performance to state-of-the-art levels across almost all evaluated scenarios.



(a) τ: BLEU-4 scores (b) τ: CIDEr scores

(c) $|N|$: METEOR scores (d) $|N|$: ROUGE scores

Figure 4: Ablation results on two key factors of retriever: (a), (b) negative candidate set size; (c), (d) temperature.

Ablation Study on τ. As shown in Figure 4a and 4b, the IDC performance of ADPO-finetuned Qwen initially improves and then declines as $\frac{1}{\tau}$ increases. This suggests that maintaining an appropriate level of randomness in the ADPO retriever's negative caption selection is crucial. The observed trend can be explained by the trade-off in negative sample diversity. While a sharp distribution (low τ) intuitively enhances the likelihood of selecting the most semantically similar negative caption thereby contributing to the performance, it also reduces the variability of sampled negatives. Consequently, positive captions within a subset tend to be repeatedly paired with the same negative captions, thereby weakening preference optimization. Based on empirical results, we set $\tau \approx 2$ to balance semantic relevance and diversity in negative sample selection.

Ablation Study on $|N|$. We further investigate the impact of the negative candidate set size $|N|$ on ADPO performance. As illustrated in Figure 4c and 4d, we evaluate $|N| \in \{16, 64, 128, 256\}$ and observe that IDC performance initially increase with larger $|N|$ and then decrease. This trend suggests that expanding the retriever's search space enhances its ability to retrieve more *highly-confusing* negative captions, thereby improving IDC preference optimization efficacy. Specifically, a larger $|N|$ allows the retriever to sample from a broader pool of candidates, increasing the likelihood of retrieving high-quality negatives that challenge the model's discriminative capabilities which is the key to ADPO's success. However, we find that excessively large candidate sets (e.g., $|N| = 256$) degrade performance compared to $|N| = 128$. We attribute this to a *diversity-accuracy* trade-off: while a larger $|N|$ improves the retriever's coverage, it also encourages retriever to select more globally optimal negatives across the dataset. This reduces the variability of training signals, ultimately hindering the model's ability to generalize. Thus, $|N| \approx 128$ strikes an optimal balance, providing sufficient diversity without sacrificing the precision of negative sample retrieval.

6.2 Case Study

In this section, we compare the negative samples employed by DPO with negative samples selected and employed by ADPO retriever through a case study in Figure 5. We observe that the negative captions retrieved by ADPO exhibit more semantically nuanced and challenging characteristics compared to the original DPO negatives, thereby enhancing the effectiveness of preference optimization.

In row (a) of Figure 5, the positive caption is focused on a missing white car on the left-top of the image. ADPO retriever select a negative sample that describes the location change of a noticeable red car on these images, which is a subtle perspective change of camera between these two images and is semantically incorrect but very likely to cause confusion in model's visual module and captioning process. Therefore, by adding such captions into negative sample, the model is encouraged to learn more semantic knowledge about captioning subtle image differences during the preference optimization process. While in DPO, a captioning describing a car moving downward is selected which is less correlated with the image pair and less semantically confusing, which is harder for model to learn the captioning of subtle changes between the given image pair. Similarly in row (b) of Figure 5, the retriever selects a caption that has at least two confusing semantic factors, the colour of the car and the temporal sequence of the image pair. Firstly, the leaving car is grey instead of black, which is a critical semantic factor while captioning the difference. Besides, distinguishing the temporal information of the images is crucial in IDC, where the ADPO retriever select car entering to encourage the learning of temporal information.

	Dataset Positive Caption	*ADPO Negative Caption*	*DPO Negative Caption*
(a)	"there is white car is missing"	"the red car is in a different location"	"a car is now traveling down the street"
(b)	"there s no longer a grey car next to the red car"	"the black car is entering the lot"	"the person in the before picture is standing next to a gray car while he is not next to any cars in the after picture"
(c)	"four people standing in the right image"	"more people in parking lot"	"the blue truck is now in the picture on the right"

Figure 5: Case study on negative samples retrieved by ADPO and DPO.

Table 4: Comparison of training efficiency between SFT, DPO, and ADPO processing each 208 samples.

Method	**Total Time (s)**	**Samples/s**	**Memory Usage**
SFT	133.45	1.56	28.6 GB
DPO	182.82	1.14	39.0 GB
ADPO	196.07	1.06	41.3 GB

While the DPO negative sample focus on person near grey car, which is less semantically related with the difference captioning of this image pair.

However, we also observe some cases where ADPO retriever may not be able to retrieve precise semantically difference. For example, in row (c) of Figure 5, the ADPO retriever selects a captioning that is actually semantically correct from a broader perspective. In this image pair, there is truly more people in the image, meaning that this captioning can also be viewed as a positive sample for this image pair. In our study, we have implemented a filtering to exclude annotated positive samples when constructing the negative candidate pool, thereby preventing the selection of high-quality positive samples for the specific image pair. This ensures that even if semantically correct samples are chosen, they are significantly less precise than the annotated positive samples, avoiding confusion during training. However, we note that more rigorous filtering mechanisms, such as utilizing MLLMs for semantic judgment, could potentially further improve precision, but could also require substantially more computational resources and time. How to rigorously filter out semantically correct caption candidates remains as a challenge and a valuable direction for future work.

6.3 Overhead Analysis

To evaluate computational overhead of ADPO, we present a detailed efficiency analysis in Table 4. Compared to the standard DPO, our proposed ADPO framework introduces only a marginal increase in computational cost, with a 7.2% longer training time and a 5.9% higher GPU memory footprint. This additional overhead, attributed to the adversarial minimax optimization, is a modest trade-off for the performance improvements demonstrated in our experiments. Crucially, this analysis shows that ADPO remains highly practical for real-world deployment, as its time and memory overhead are efficient on modern hardware.

7 Conclusion

We propose Adversarial Direct Preference Optimization (ADPO) for Image Difference Captioning, which aligns captioning policy with fine-grained human preferences via adversarial hard negative retrieval and direct preference optimization. By framing IDC as a pairwise preference problem with a novel preference model, our method overcomes limitations of supervised learning and enhances subtle difference localization. We derive a reinforcement learning objective that enables closed-form DPO policy updates and policy-gradient retriever updates. Experiments on standard IDC benchmarks show that ADPO delivers robust and discriminative captions, outperforming existing approaches and establishing a strong baseline for future work.

8 Limitation

Our approach focuses on caption-level descriptions, which are standard in IDC benchmarks and provide a well-defined setting for evaluating fine-grained visual differences. The framework is also compatible with more structured outputs, including region-level alignments or paragraph-level explanations, although additional effort may be required to support such representations in practice, depending on the needs of downstream applications. While our method is designed for natural image inputs, the underlying preference optimization framework is modality-agnostic and could in principle be adapted to structured visual formats such as HTML renderings or document layouts. These domains pose different types of reasoning challenges, and more efforts might be needed to adapt the framework effectively to such inputs.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonzalez, and Serena Yeung-Levy. 2024. Describing differences in image sets with natural language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24199–24208. IEEE.

Gautier Evennou, Antoine Chaffin, Vivien Chappelier, and Ewa Kijak. 2025. Reframing image difference captioning with blip2idc and synthetic augmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1392–1402.

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

Zixin Guo, Tzu-Jui Julius Wang, and Jorma Laaksonen. 2022. CLIP4IDC: CLIP for image difference captioning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 33–42. Association for Computational Linguistics.

Zhiwei Hao, Jianyuan Guo, Ding Jia, Kai Han, Yehui Tang, Chao Zhang, Han Hu, and Yunhe Wang. 2022. Learning efficient vision transformers via fine-grained manifold distillation. *Advances in Neural Information Processing Systems*, 35:9164–9175.

Erdong Hu, Longteng Guo, Tongtian Yue, Zijia Zhao, Shuning Xue, and Jing Liu. 2024. Onediff: A generalist model for image difference captioning. In *Computer Vision - ACCV 2024 - 17th Asian Conference on Computer Vision, Hanoi, Vietnam, December 8-12, 2024, Proceedings, Part III*, volume 15474 of *Lecture Notes in Computer Science*, pages 114–130. Springer.

Jinhong Hu, Benqi Zhang, and Ying Chen. 2023. Clip-driven distinctive interactive transformer for image difference captioning. In *2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, pages 1232–1236.

Qingbao Huang, Yu Liang, Jielong Wei, Yi Cai, Hanyu Liang, Ho-fung Leung, and Qing Li. 2022. Image difference captioning with instance-level fine-grained feature representation. *IEEE Trans. Multim.*, 24:2004–2017.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4024–4034. Association for Computational Linguistics.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021a. Viewpoint-agnostic change captioning with cycle consistency. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2075–2084. IEEE.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021b. Viewpoint-agnostic change captioning with cycle consistency. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2075–2084.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, and 1 others. 2025. Large language models and causal inference in collaboration: A comprehensive survey. *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7668–7684.

Jinda Lu, Junkang Wu, Jinghan Li, Xiaojun Jia, Shuo Wang, YiFan Zhang, Junfeng Fang, Xiang Wang, and Xiangnan He. 2025. Dama: Data- and model-aware alignment of multi-modal llms. *Preprint*, arXiv:2502.01943.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019a. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019b. Robust change captioning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4623–4632. IEEE.

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.

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq R. Joty, and Jianfei Cai. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIV*, volume 12359 of *Lecture Notes in Computer Science*, pages 574–590. Springer.

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. 2022. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13956–13966.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

Yaoqi Sun, Liang Li, Tingting Yao, Tongyv Lu, Bolun Zheng, Chenggang Yan, Hua Zhang, Yongjun Bao, Guiguang Ding, and Gregory G. Slabaugh. 2022. Bidirectional difference locating and semantic consistency reasoning for change captioning. *Int. J. Intell. Syst.*, 37(5):2969–2987.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1873–1883. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5-vl.

Yunbin Tu, Liang Li, Li Su, Junping Du, Ke Lu, and Qingming Huang. 2023a. Viewpoint-adaptive representation disentanglement network for change captioning. *IEEE Trans. Image Process.*, 32:2620–2635.

Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang. 2023b. Neighborhood contrastive transformer for change captioning. *IEEE Trans. Multim.*, 25:9518–9529.

Yunbin Tu, Liang Li, Li Su, Chenggang Yan, and Qingming Huang. 2024. Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In *ECCV*, pages 311–328.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. 2023c. Self-supervised cross-view representation reconstruction for change captioning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2793–2803. IEEE.

Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. 2021a. R^3Net:relation-embedded representation reconstruction network for change captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9319–9329, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunbin Tu, Tingting Yao, Liang Li, Jiedong Lou, Shengxiang Gao, Zhengtao Yu, and Chenggang Yan. 2021b. Semantic relation-aware difference representation learning for change captioning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 63–73. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

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*.

Ruoyu Wang, Tong Yu, Junda Wu, Yao Liu, Julian McAuley, and Lina Yao. 2025. Weakly-supervised vlm-guided partial contrastive learning for visual language navigation. *arXiv preprint arXiv:2506.15757*.

Junda Wu, Xintong Li, Tong Yu, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Jingbo Shang, and Julian McAuley. 2024a. Commit: Coordinated instruction tuning for multimodal large language models. *arXiv preprint arXiv:2407.20454*.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, and 1 others. 2024b. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*.

Junda Wu, Rohan Surana, Zhouhang Xie, Yiran Shen, Yu Xia, Tong Yu, Ryan A Rossi, Prithviraj Ammanabrolu, and Julian McAuley. 2025a. In-context ranking preference optimization. *arXiv preprint arXiv:2504.15477*.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A Rossi, and 1 others. 2025b. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 67–78.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, and 1 others. 2025c. Mitigating visual knowledge forgetting in mllm instruction-tuning via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024c. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, and 1 others. 2024d. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. 2025d. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *Preprint*, arXiv:2502.13928.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, and 1 others. 2025. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, and 1 others. 2024. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Linli Yao, Weiying Wang, and Qin Jin. 2022a. Image difference captioning with pre-training and contrastive learning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3108–3116. AAAI Press.

Linli Yao, Weiying Wang, and Qin Jin. 2022b. Image difference captioning with pre-training and contrastive learning. In *AAAI Conference on Artificial Intelligence*.

Xian Zhang, Haokun Wen, Jianlong Wu, Pengda Qin, Hui Xue', and Liqiang Nie. 2024. Differential-perceptive and retrieval-augmented MLLM for change captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 4148–4157. ACM.

Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, and 1 others. 2025. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*.