Dynamics-Aware Preference Optimization for Vision-Language Models

Jusheng Zhang[1] Kaitong Cai[1] Jing Yang[1] Jian Wang[2] Keze Wang[1]
[1]Sun Yat-sen University [2]Snap Inc.

Abstract

Preference-based finetuning of vision-language models (VLMs) is notoriously unstable, i.e., trivially wrong negatives inject uninformative gradients that distort optimization and degrade calibration. This work revisits this issue through the lens of ***learning dynamics*** *and identifies a core pathology, i.e., the* squeezing effect*, where easy negatives retain large, misaligned gradients despite negligible loss. To address this, we propose* ***Cooling-Weighted Direct Preference Optimization (CW-DPO)****, a two-stage framework that smooths and then stabilizes the alignment process.* ***Stage 1*** *employs a constrained SFT phase with low-weight "gentle negatives" to regularize overconfident distributions and flatten the loss landscape.* ***Stage 2*** *introduces a competence-aware cooling weight that adaptively scales negative gradients according to the model's average per-token log-probability, suppressing uninformative updates while emphasizing hard, on-policy contrasts. This dynamics-aware weighting effectively mitigates the squeezing effect and enables smoother convergence. Extensive and comprehensive results on the mainstream benchmarks, i.e., COCO, Flickr30k, NoCaps, MMMU, and MMBench1.1, our CW-DPO achieves state-of-the-art performance, e.g., $+3.4$ CIDEr over PPO and $+2.4\%$ absolute accuracy on MMMU, while improving calibration and halving convergence steps. This justifies that* ***smoothing before cooling*** *constitutes a simple yet general principle for robust VLM preference optimization. https://github.com/jushengzhang/Dynamics-Aware-Preference-Optimization*

1. Introduction

The finetuning of vision-language models (VLMs) involves intricate learning dynamics that pose significant challenges for stable optimization [10, 19, 47, 48, 53, 55, 58]. VLMs process multimodal inputs, encoding textual and visual components as high-dimensional sequences, where the visual stream introduces complex state dependencies, such as pixel embeddings and spatial metadata, that tightly couple gradient updates across tokens [17, 28, 51, 57]. Prominent finetuning methods, including supervised finetuning (SFT) [27, 38] and direct preference optimization (DPO) [29], employ diverse loss geometries and supervision signals, necessitating a unified analytical framework to unravel their behavioral foundations, especially in preference-based alignment aimed at prioritizing human-preferred outputs [31]. Preference-based finetuning is essential for aligning VLMs with human intent [6, 20, 28, 45, 52, 54, 56], yet it suffers from notorious instability in practice. Alignment datasets often contain static or mis-specified negative examples (trivially incorrect or off-distribution) that inject uninformative gradients [5, 14, 35, 46, 49]. These gradients disrupt optimization, degrade calibration, and produce overconfident, peaky posteriors. Off-policy methods exacerbate this by penalizing unlikely responses, while even naive on-policy approaches struggle with gradient spikes from dominant "easy negatives" [7, 14, 50]. This points to a common flaw: alignment is often treated as a static optimization task, ignoring the dynamic evolution of the model's learning trajectory [8, 14, 31].

In this work, we adopt a learning-dynamics perspective, reframing alignment to explicitly model and harness how the model's beliefs evolve during finetuning [32]. We introduce Cooling-Weighted Direct Preference Optimization (CW-DPO), a two-stage strategy that aligns with this evolution. The first stage smooths the loss landscape to enhance stability, while the second applies a competence-aware preference optimization to refine training, as depicted in Figure 2. Specifically, Stage 1 enhances SFT by incorporating "gentle negatives", introducing low-weight smoothed supervision to reduce overconfidence around negative responses without harsh penalties. We define the per-token average log-probability as $\bar{\ell}_\theta(y \mid \chi) = \frac{1}{L}\sum_{l=1}^{L} \log \pi_\theta(y_l \mid \chi_{\le l})$, measuring the model's average confidence per token on any response y given sample χ (elaborated in §3.2), with y_l (loser) and y_w (winner) specifying roles in Stage 2. The objective, formalized as a constrained optimization (detailed in §4), is: $\min_\theta \mathbb{E}_{(x,y^+)\sim\mathcal{D}}[-\log \pi_\theta(y^+ \mid x)] + \eta\mathcal{R}_{\text{smooth}}(\theta; x, y^-), \quad 0 < \eta \ll 1$, where $\mathcal{R}_{\text{smooth}}$ (e.g., entropy smoothing or a ReLU-based soft constraint) regularizes the negative trajectory y^-. This "smooth-

before-optimize" approach de-peaks distributions and flattens sharp loss regions, reducing noise in subsequent contrastive learning, as motivated by the peaking pitfalls in §3.1. In Stage 2, we transition to preference pairs y_w (winner) and y_l (loser), as detailed in §4. Stage 2 advances with a novel DPO-style objective featuring competence-aware reweighting [29]. Vanilla DPO minimizes $-\log\sigma(\beta(\Delta_w - \Delta_l))$, where $\Delta_w = \log\pi_\theta(y_w \mid x) - \log\pi_{\text{ref}}(y_w \mid x)$ and $\Delta_l = \log\pi_\theta(y_l \mid x) - \log\pi_{\text{ref}}(y_l \mid x)$. We enhance it with a cooling weight: $w_c(\theta; y_l, \chi) = \sigma\left(\frac{\bar{\ell}_\theta(y_l|\chi) - \ell_{\text{floor}}}{\tau}\right)$, which down-weights y_l with low probabilities (indicating "easy" negatives), steering optimization toward hard negatives where uncertainty lingers. The resulting loss is: $\mathcal{L}_{\text{CW-DPO}} = -\mathbb{E}\big[\log\sigma(\beta(\Delta_w - w_c(\theta; y_l, \chi)\cdot\Delta_l))\big]$, where ℓ_{floor} sets an easiness baseline and τ adjusts the cooling schedule's sharpness. Negatives are primarily on-policy, with optional dataset-negative mixing to keep contrast fresh. Across both stages, $\Delta\log p$ probes on a held-out set monitor learning dynamics, providing a low-cost signal for early stopping and curriculum design. This endogenous curriculum adapts to model competence. Extensive and comprehensive experimental evaluations in §5 demonstrate that our CW-DPO surpasses SFT-only and vanilla DPO in stability, efficiency, calibration, and win-rates across visual QA, binary judgments, and open-ended tasks.

2. Related Work

Supervised and RLHF-based Alignment. The alignment of vision-language models (VLMs) with human preferences has become a central topic, following the success of large language model (LLM) alignment techniques [1, 13, 19, 36]. Early efforts such as supervised finetuning (SFT) [20, 27] relied on curated image-text pairs, achieving strong descriptive ability but often suffering from overconfidence and limited generalization to diverse intents. To improve alignment fidelity, reinforcement learning from human feedback (RLHF) [9, 14] is extended to multimodal settings using proximal policy optimization (PPO) [33, 37]. However, RLHF pipelines are notoriously unstable and computationally expensive, requiring online rollouts, a separate reward model, and careful tuning [30, 59].

Direct Preference Optimization. Direct Preference Optimization (DPO) [29, 39, 40, 42] simplified RLHF into a reward-free objective that directly optimizes on preference pairs. Its simplicity and efficiency motivated numerous multimodal extensions. V-DPO [39] incorporates visual preference cues. GRPO [34] introduces group regularization to reduce overfitting, and OPA-DPO [42] enhances data efficiency through online preference augmentation. Other works, such as Task Preference Optimization (TPO) [41], Structured Preference Optimization (SPO), and Calibrated Multi-Preference Optimization (CaPO) [16], extend the paradigm toward task-specific, structured, or calibration-aware formulations. Despite their effectiveness, these methods largely treat all preference pairs uniformly, overlooking how gradient signals evolve during training.

Method	No Reward Model	Visual-Aware	Online Efficient	Regularized Gradients	Dynamics-Aware	Stable Calibration
SFT [27]	✓	✓	✓	×	×	×
RLHF [7]	×	×	×	×	×	×
DPO [29]	✓	×	✓	×	×	×
V-DPO [40]	✓	✓	✓	×	×	×
GRPO [34]	✓	✓	×	✓	×	×
OPA-DPO [42]	✓	✓	✓	×	×	×
CW-DPO (ours)	✓	✓	✓	✓	✓	✓

Table 1. **Comparison with representative preference optimization methods.** CW-DPO uniquely integrates learning dynamics modeling and competence-aware weighting for stable VLM alignment.

Learning Dynamics and Stable Alignment. Recent studies [11, 26, 31] reveal that alignment follows distinct learning phases—early generalization followed by memorization—yet uninformative gradients from easy or off-policy negatives can distort this trajectory. This leads to the *squeezing effect*, where probability mass collapses toward dominant modes, harming calibration and diversity. Our proposed **Cooling-Weighted Direct Preference Optimization (CW-DPO)** explicitly models these dynamics via a two-stage scheme: (1) a *constrained SFT* phase that smooths the loss landscape through gentle negatives, and (2) a *competence-aware cooling mechanism* that adaptively scales negative gradients according to the model's per-token log-probability. By jointly smoothing and cooling the learning trajectory, CW-DPO mitigates gradient imbalance, improves calibration, and stabilizes convergence.

Comparison with Prior Methods. Table 1 summarizes representative preference optimization methods and highlights how CW-DPO uniquely combines learning dynamics modeling and adaptive weighting for robust multimodal alignment. As shown in Table Table 1, only CW-DPO simultaneously satisfies all six desirable properties, offering both theoretical interpretability and empirical stability across diverse multimodal benchmarks.

3. Problem Formulation: The Unstable Dynamics of VLM Finetuning

We systematically dissect the core instabilities afflicting VLM alignment, i.e., **a fundamental dilemma in preference-based learning, manifesting as the "squeezing effect,"** in §3.1. This effect underscores **a perilous decoupling between a sample's loss-based informativeness**

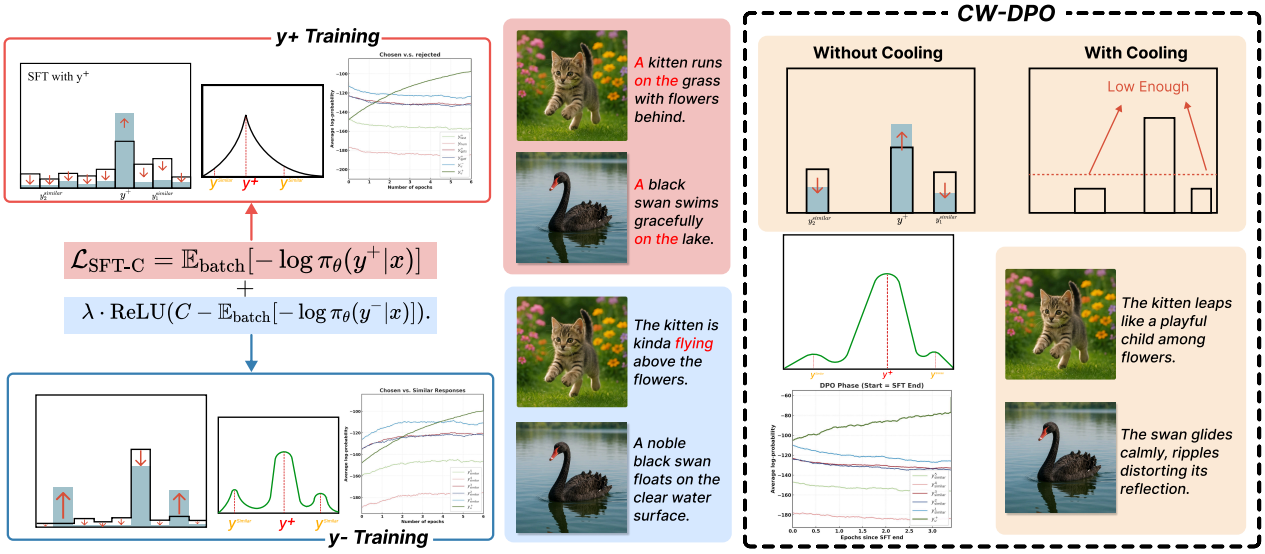

Figure 1. Two-stage optimization process of CW-DPO. Stage 1 (y^+ Training) leverages positive supervision for stability but yields overly uniform language styles (e.g., "A . . . on the . . . "). Stage 2 (y^- Training) introduces negative contrast for variation but risks errors (e.g., a running kitten as "flying"). CW-DPO's cooling-weighted mechanism dynamically attenuates uninformative negatives while amplifying hard ones, mitigating error propagation, and enhancing stylistic diversity.

and its gradient-based influence during training. Subsequently, in §3.2, we develop a formal analytical lens rooted in learning dynamics to diagnose this issue. This framework not only elucidates the root causes of instability but also **yields a principled blueprint for our dynamics-aware solution**.

3.1. A Core Dilemma: The "Squeezing Effect"

The **fundamental dilemma** of preference finetuning is that aligning with human intent requires penalizing a vast space of undesirable responses (y^-) [14]. As learning progresses, most undesirable responses are gradually converted into "easy negatives", i.e., sequences assigned near-zero probability by the model. This engenders **a destructive feedback loop**, wherein optimization expends disproportionate gradient bandwidth on these uninformative samples. As shown in Figure 1, the consequence is the **squeezing effect**, i.e., a decoupling where a sample's low loss (indicating minimal informativeness) belies its potentially large, misdirected gradient [31]. Although the loss from an easy negative $\pi_\theta(y^-|x) \to 0$ is negligible, its gradient can remain substantial and poorly aligned. This misalignment induces an undesirable redistribution of probability mass: instead of fostering a calibrated spread across viable alternatives, updates "squeeze" mass toward the dominant mode, typically $y^* = \arg\max_y \pi_\theta(y|x)$, which may correspond to a preferred response y_w in later optimization stages. **This engenders a "rich-get-richer" dynamic**, amplifying overconfidence, curtailing linguistic diversity, and impairing calibration.

Remark (Insufficiency of DPO's Implicit Regularization). DPO implicitly counters this via regularization: the negative-term gradient is modulated by $\beta(1-a)$, where $a = \sigma(\beta(\Delta_w - \Delta_l))$ is the sigmoid-transformed margin. For extremely easy negatives, Δ_l drives $a \to 1$, attenuating the gradient. Theoretically elegant, this falters in practice due to a wide "vulnerable region" for **moderately easy negatives**, where $\log \pi_\theta(y^-)$ is low but a (e.g., $\in [0.8, 0.99]$) insufficiently suppresses the residual gradient $\beta(1-a)$, especially at high β [31]. This perpetuates instability and the squeezing effect. (See Appendix 18 for a formal analysis).

3.2. Analytical Lens: Per-Step Influence Decomposition

To transcend empirical observations and rigorously diagnose the squeezing effect, we adopt a learning-dynamics perspective [12, 15] to enable precise tracing of how a single gradient update impacts global model behavior. Define $y = (y_1, \ldots, y_L)$ as a sequence of length L, with logits $z = (z_1, \ldots, z_L)$, each $z_l \in \mathbb{R}^{|V|}$ ($|V|$ denotes the vocabulary size). Gradients are w.r.t. the concatenated z, denoted ∇_z. A pivotal query: How does an update on "updating" sample $\chi_u = (x_u, y_w, y_l)$ alter confidence on "observing" sample χ_o? Confidence is quantified via average per-token log-probability: $\bar{\ell}_\theta(y \mid \chi) = \frac{1}{L}\sum_{l=1}^{L} \log \pi_\theta(y_l \mid \chi_{<l})$. A first-order Taylor expansion of $\bar{\ell}_\theta$ post-update $\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t; \chi_u)$ yields:

$$\Delta \bar{\ell}_t(y|\chi_o) = -\eta (\nabla_\theta \bar{\ell}_{\theta_t})^\top (\nabla_\theta \mathcal{L}(\theta_t)) + \mathcal{O}(\eta^2). \tag{3.1}$$

Linearizing logits $z(\theta; \chi)$ around θ_t decomposes this into interpretable factors.

Proposition 3.1 (Sequence-Aware One-Step Influence)**.** *The log-likelihood change on χ_o post-update on χ_u (rate*

η) *approximates:*

$$\Delta\bar{\ell}_t(y \mid \chi_o) \approx -\eta \left\langle \underbrace{\nabla_z \bar{\ell}_{\theta_t}(y \mid \chi_o)}_{A_t:\,\textit{Belief Geometry}}, \underbrace{K_t(\chi_o, \chi_u)}_{\textit{eNTK Kernel}} \underbrace{\nabla_z \mathcal{L}(\theta_t; \chi_u)}_{G_t:\,\textit{Loss Residual}} \right\rangle. \tag{3.2}$$

Key Elements: ***Belief Geometry*** *(A_t) encodes predictive sensitivity to logit perturbations, capturing belief-landscape curvature.* ***eNTK Kernel*** *($K_t = J_o J_u^\top$) ($J = \nabla_\theta z(\theta_t; \chi)$: Jacobian) propagates updates parametrically.* ***Loss Residual*** *(G_t) directs logit adjustments via $\nabla_z \mathcal{L}$.*

Decomposing the DPO Gradient. The power of this decomposition becomes evident when we specify the Loss Residual G_t for the DPO objective. For DPO, $G_t = \nabla_z \mathcal{L}_{\text{DPO}}$ (derived in Appendix 18), whose full form is given in Eq. 4.3 and can be broken down into components related to the winner y_w and the loser y_l: $G_t = \beta(1-a)(G_t^w - G_t^l)$, where G_t^w and G_t^l are the gradient components for the winning and losing responses, respectively. As discussed in Remark 3.1, the squeezing effect occurs precisely when y_l is an "easy negative." In this scenario, while the loss itself is small, DPO's implicit regularization $(1-a)$ is often insufficient to fully suppress the gradient, leaving the loser component G_t^l disproportionately large and noisy. This oversized residual from uninformative samples is the direct source of instability.

Implication for Algorithm Design. This analysis transcends explanation: **it isolates the instability's source to the oversized, destabilizing "loser" component (G_t^l) of the loss residual from negative examples** y_l. The squeezing effect, therefore, emerges not from an inherent flaw in preference optimization but from an unregulated G_t^l. **This mandates a surgical solution**: instead of heuristically regularizing the entire loss, a principled algorithm must directly temper this specific residual component. This diagnosis is the analytical foundation for our method, detailed in the next section 4.

4. Dynamics-Aware Cooling-Weighted DPO

Grounded in the principled insights of our diagnostic analysis (§3), our CW-DPO in Figure 2 provides a dynamics-aware manner to align VLMs.

4.1. Stage 1: Trajectory Priming via Constrained SFT

This stage prepares the learning trajectory of the model π_θ (θ denotes the model parameters) by curbing overconfidence, laying a smoother foundation for subsequent preference learning. Unlike standard SFT, which focuses solely on positive responses (y^+) and risks entrenching peaky distributions, we adopt a constrained optimization strategy. To mitigate overconfidence, we impose a constraint on the model's response to negatives, minimizing the negative log-likelihood (NLL) on positives while ensuring the NLL on negatives (y^-) remains above a threshold C to prevent their premature dismissal as:

$$\begin{aligned} &\min_\theta \mathbb{E}_{(x,y^+)\sim\mathcal{D}}\big[-\log \pi_\theta(y^+ \mid x)\big] \\ &\text{s.t. } \mathbb{E}_{(x,y^-)\sim\mathcal{D}}\big[-\log \pi_\theta(y^- \mid x)\big] \;\geq\; C \end{aligned} \tag{4.1}$$

Here, the objective $\mathbb{E}_{(x,y^+)\sim\mathcal{D}}[-\log \pi_\theta(y^+|x)]$ seeks to maximize the likelihood of positive responses y^+ drawn from dataset $\mathcal{D}$, while the constraint $\mathbb{E}_{(x,y^-)\sim\mathcal{D}}[-\log \pi_\theta(y^-|x)] \geq C$ ensures that the model assigns sufficient probability to negative examples y^-, preventing them from being overly suppressed. This dual focus promotes a more uniform allocation of probability mass, countering the peaking pitfalls outlined in §3.1, where overconfidence on easy negatives distorts the loss landscape. To solve this constrained problem practically, we apply a Lagrangian relaxation, introducing a penalty term to softly approximate the constraint. This leads to the Smoothed SFT loss:

$$\begin{aligned} \mathcal{L}_{\text{SFT-C}} = \mathbb{E}_{\text{batch}}\big[-\log \pi_\theta(y^+ \mid x)\big] \\ + \lambda\, \text{ReLU}\Big(C - \mathbb{E}_{\text{batch}}\big[-\log \pi_\theta(y^- \mid x)\big]\Big) \end{aligned} \tag{4.2}$$

The first term, $\mathbb{E}_{\text{batch}}[-\log \pi_\theta(y^+|x)]$, remains the standard NLL for positive examples, computed over mini-batches for efficiency. The second term, $\lambda \cdot \text{ReLU}(C - \mathbb{E}_{\text{batch}}[-\log \pi_\theta(y^-|x)])$, acts as a regularization: if the expected NLL of negatives falls below C, the ReLU activates, penalizing the model with a strength proportional to λ. This soft enforcement encourages the model to maintain a balanced response to negatives without rigid enforcement, approximating the original constraint stochastically. Here, mini-batch expectations provide practical approximations, and the ReLU term gently nudges the model toward a well-calibrated initialization. This process stabilizes the Belief Geometry (A_t in Prop. 3.1), setting the stage for the targeted adjustments in Stage 2 by smoothing the initial loss landscape.

4.2. Stage 2: Competence-Aware Preference Optimization

§3.2 reveals that instability stems from gradient updates for the negative (loser) sample y_l, particularly the loser component of the Loss Residual (G_t), which generates oversized and uninformative updates for easy negatives. By asymmetrically applying a cooling weight w_c to the loser's log-probability difference Δ_l, we achieve precise control over gradient influence. **Vanilla DPO Gradient.** Consider the DPO loss for a preference pair (y_w, y_l): $\mathcal{L}_{\text{DPO}} = -\log \sigma\big(\beta(\Delta_w - \Delta_l)\big)$, where $\Delta_{w/l} = \log \pi_\theta(y_{w/l}|x) - \log \pi_{\text{ref}}(y_{w/l}|x)$. The gradient with respect to the logits (the Loss Residual) is:

$$G_t^{\text{DPO}} = \nabla_z \mathcal{L}_{\text{DPO}} = \beta(1-a)\left((g_w - g_{\text{ref}}^w) - (g_l - g_{\text{ref}}^l)\right), \tag{4.3}$$

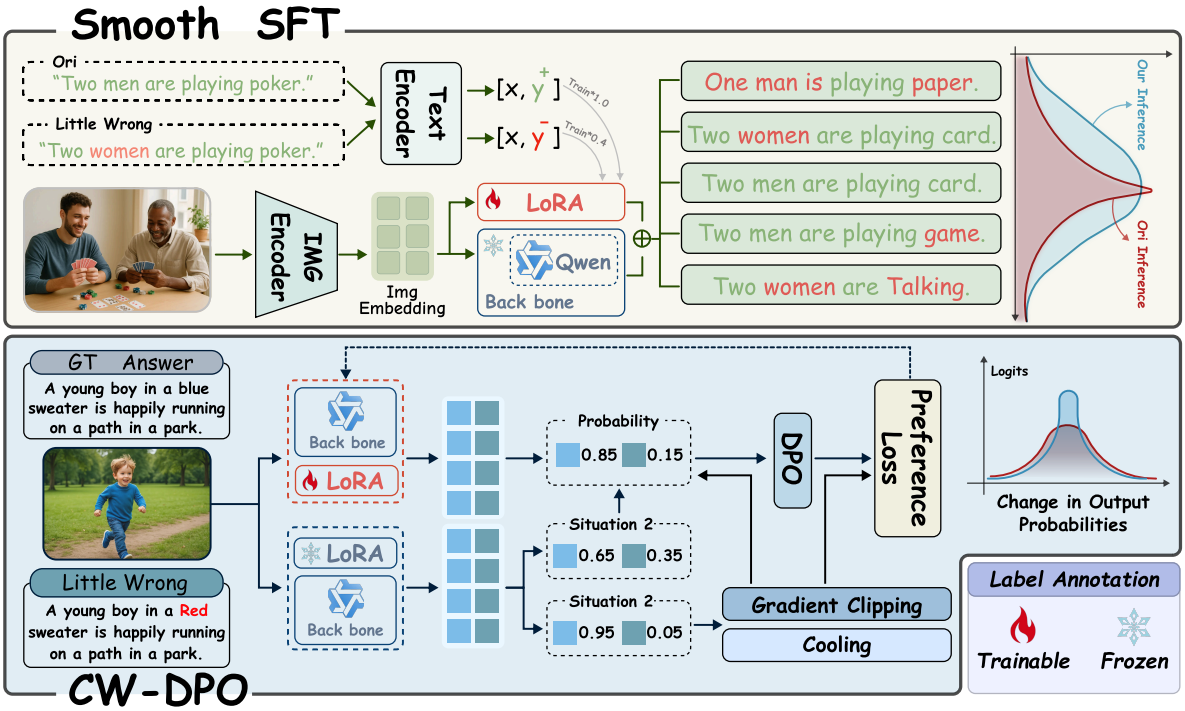

Figure 2. our CW-DPO is designed to balance **generalization** and **precision** through a two-stage optimization strategy. In Stage 1, *Smooth SFT* leverages positive samples together with negative samples containing minor errors to construct a smoothed supervision signal. This broadens the model's output probability distribution, thereby enhancing its generalization ability and robustness. In Stage 2, our CW-DPO employs preference pairs with fine-grained errors for DPO. By sharpening the probability distribution, this stage strengthens the model's capacity for precise discrimination of critical details.

Algorithm 1 The Two-Stage CW-DPO Finetuning Protocol

Require: Pretrained VLM θ_0, dataset $\mathcal{D}$, hyperparameters $\lambda, C, \beta, \tau, \ell_{\text{floor}}$, learning rate α.
Ensure: Finetuned VLM parameters θ.

Initialize: Policy model $\theta \leftarrow \theta_0$.

Stage 1: Trajectory Priming

for $t = 1, \ldots, T_1$ **do**
  Sample a mini-batch (x, y^+, y^-) from $\mathcal{D}$.
  Compute Smoothed SFT loss $\mathcal{L}_{\text{SFT-C}}$ using Eq. 4.2.
  Update parameters: $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{SFT-C}}$.
end for
Set reference model: $\pi_{\text{ref}} \leftarrow \pi_\theta$.

Stage 2: Cooled Preference Optimization

for $t = 1, \ldots, T_2$ **do**
  Sample a mini-batch of preferences (x, y_w, y_l) from $\mathcal{D}$.
  Compute cooling weight w_c for each sample using Eq. 4.4.
  Compute CW-DPO loss $\mathcal{L}_{\text{CW-DPO}}$ using Eq. 4.5.
  Update parameters: $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\text{CW-DPO}}$.
end for

return Finetuned parameters θ.

where $a = \sigma(\beta(\Delta_w - \Delta_l))$ and $g_{w/l} = \nabla_z \log \pi_\theta(y_{w/l}|x)$. Our decomposition pinpoints the squeezing effect to the loser term $(g_l - g^l_{\text{ref}})$, which drives instability for easy negatives [31].

Cooling Weight: Principled Modulator. To address this, we introduce the cooling weight w_c, which adjusts the negative-sample gradient based on real-time model confidence:

$$w_c(\theta; y_l, \chi) = \sigma\left(\frac{\bar{\ell}_\theta(y_l \mid \chi) - \ell_{\text{floor}}}{\tau}\right), \tag{4.4}$$

where $\bar{\ell}_\theta(y_l \mid \chi)$ is the average per-token log-probability (as defined in §3.2), ℓ_{floor} establishes an "easiness" baseline, and τ controls the transition sharpness, with higher values yielding a smoother weighting. For confidently rejected responses ($\bar{\ell}_\theta \ll \ell_{\text{floor}}$), $w_c \to 0$, nullifying the gradient; for uncertain hard negatives ($\bar{\ell}_\theta \geq \ell_{\text{floor}}$), $w_c \to 1$, preserving the learning signal.

Core Loss Function. We integrate w_c asymmetrically, dampening only Δ_l, to define our core loss:

$$\mathcal{L}_{\text{CW-DPO}} = -\log \sigma\left(\beta\left(\Delta_w - w_c(\theta; y_l, \chi) \cdot \Delta_l\right)\right), \tag{4.5}$$

Differentiating (treating w_c as locally constant) yields the cooled residual $G_t^{\text{CW}} = \nabla_z \mathcal{L}_{\text{CW-DPO}}$ as:

$$\begin{aligned}\nabla_z \mathcal{L}_{\text{CW-DPO}} &= \beta(1-a')\left(\nabla_z \Delta_w - w_c \nabla_z \Delta_l\right) \\ &= \beta(1-a')\left((\pi_\theta(\cdot \mid x) - y_w) - w_c\left(\pi_\theta(\cdot \mid x) - y_l\right)\right)\end{aligned} \tag{4.6}$$

where $a' = \sigma(\beta(\Delta_w - w_c \Delta_l))$. This G_t^{CW} ensures gradients from easy negatives are minimized, preserving positive updates, and resolves the squeezing effect for stable, superior alignment. See Algorithm 1 for the full protocol. In essence, CW-DPO stabilizes training by smoothing initial losses and refining preferences with competence-aware weights, as validated empirically in §5.

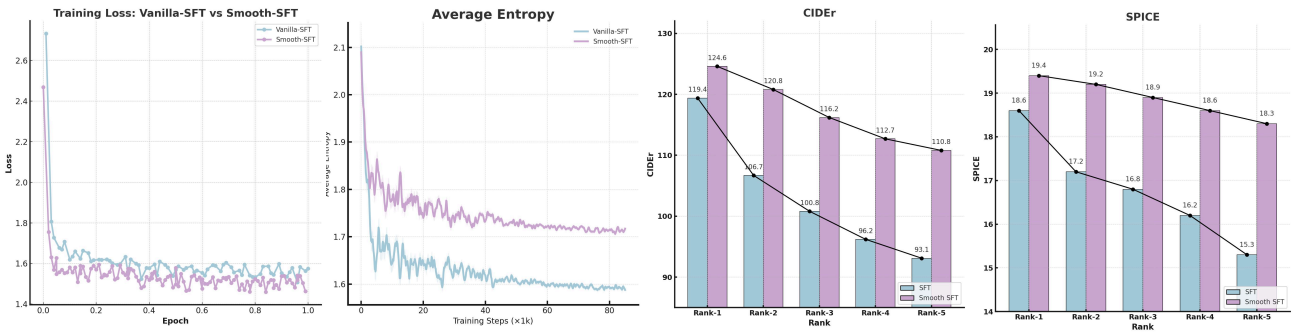

Figure 3. **Validation of Stage 1 Constrained SFT (SFT-C)** vs. standard SFT on: (1) loss; (2) entropy; (3) CIDEr; and (4) SPICE for Top-5 generations. SFT-C sustains higher entropy (less squeezing) and overall quality.

Table 2. Performance comparison on vision-language benchmarks. For COCO, Flickr30k, and NoCaps, we report BLEU-4 (B@4), METEOR (M), CIDEr (C), and SPICE (S), with NoCaps split into In, Near, Out, and Entire. We also report accuracy on MMMU and MMBench1.1. Best results are in **bold**.

Method	COCO Test				Flickr30k Test		NoCaps Val				MMMU	MMBench
	B@4	M	C	S	C	S	In	Near	Out	Entire		
Qwen2.5-VL (Base)	31.2	26.2	128.8	23.8	78.9	17.2	115.6	113.7	117.6	116.2	70.2	84.9
SFT	35.2	28.4	136.5	24.3	83.2	17.5	121.2	118.5	120.1	120.4	71.8	86.2
DPO	33.5	28.0	136.9	24.0	86.5	18.0	119.5	117.2	119.8	118.9	71.1	84.9
PPO	34.9	28.7	139.2	24.7	82.1	17.7	120.2	118.9	120.0	119.7	71.4	85.8
V-DPO	36.6	28.7	138.3	24.8	86.3	18.2	122.5	119.0	121.6	121.0	72.9	86.8
GRPO	36.5	28.8	138.2	24.9	86.4	18.1	122.3	119.1	121.5	120.9	72.8	86.9
OPA-DPO	36.8	29.0	138.5	25.1	86.7	18.2	122.6	119.4	121.8	121.3	73.1	87.2
CW-DPO (Ours)	**39.6**	**30.4**	**142.6**	**25.8**	**89.2**	**18.6**	**125.6**	**121.3**	**123.7**	**123.6**	**74.6**	**89.6**

5. Experiments

5.1. Main Results on Standard Benchmarks

We evaluate our CW-DPO on three standard image captioning benchmarks: COCO [18], Flickr30k [43], and NoCaps [2] for generalization assessment, as well as two comprehensive multi-task evaluation benchmarks: MMMU [44] and MMBench [21]. For COCO and Flickr30k, we adopt the widely used Karpathy split. The backbone for our CW-DPO in all the experiments is **Qwen2.5-VL-72B** [4], and we compare it against a series of strong fine-tuning baselines, including SFT [27], vanilla DPO [29], PPO [33], and GRPO [34]. To ensure robustness, all reported results are averaged over **five independent runs**. As for **training protocol**, our CW-DPO follows a two-stage paradigm, i.e., Constrained SFT on 75% of the data and Preference Alignment on the remaining 25%. In Stage 2, preference pairs are built by synthesizing minimally perturbed alternatives y_l for each winning caption y_w via `GPT-4o`.

In Table 2, our CW-DPO consistently outperforms all compared methods, including recent DPO variants like V-DPO [40], GRPO, and OPA-DPO [42], across 5 mainstream vision-language benchmarks. On COCO Test, our CW-DPO achieves a new SOTA CIDEr score of **142.6**, surpassing the strongest baseline PPO by 3.4 points (+2.4%). It also yields a high BLEU-4 score of **39.6**, marking a substantial improvement of 2.8 points (+7.6%) over OPA-DPO, reflecting enhanced overall generation quality. On Flickr30k Test that evaluates cross-domain generalization, our CW-DPO continues to lead all baselines with a CIDEr score of **89.2**, 2.5 points higher than the next-best method, OPA-DPO. This suggests that the training stability introduced by our CW-DPO translates effectively into stronger generalization across distribution shifts. On the more challenging NoCaps, our CW-DPO achieves leading performance across all subsets with an overall score of **123.6**. Notably, the gain on the out-of-domain split (+1.9) does not come at the expense of in-domain performance (+3.0) when compared to the strongest baselines, indicating a favorable trade-off between generalization and retention of core knowledge. Furthermore, CW-DPO achieves strong adaptability on two multi-task evaluation suites by obtaining an accuracy of **74.6%** on MMMU, outperforming the strongest baseline OPA-DPO (73.1%) and attaining the highest accuracy of **89.6%** on MMBench. This confirms our CW-DPO extends beyond captioning to broader multimodal reasoning tasks.

Note that, vanilla DPO underperforms SFT on lexical metrics such as BLEU-4. This justifies our **core hypothesis** that naive preference optimization over easy negatives may induce over-penalization, thereby degrading generation quality.

5.2. Phase-One Smoothing Validation Experiment

To isolate and verify the effectiveness of our Constrained SFT (SFT-C) in mitigating the squeezing effect, we conduct a targeted validation experiment. We train two models on 75% of the COCO training data (∼85k samples): one with standard SFT and the other with SFT-C. During training, both models are periodically evaluated on a fixed *probe set* of 1,000 examples from the COCO validation split. To quantify the squeezing effect, we measure the **average entropy** of the model's predictive distribution over the probe set. Lower entropy signifies a sharper, less diverse ("squeezed") distribution. To ensure the increased smoothness does not degrade generation quality, we also compute the **CIDEr** and **SPICE** scores for the **Top-5** predictions of each model at every evaluation step. Figure 3 validates the advantage of our SFT-C. The standard SFT exhibits a rapid decrease in training loss, but this is coupled with a **precipitous drop in predictive entropy**. This confirms that standard SFT quickly develops an overconfident, peaky distribution, i.e., a clear indicator of the squeezing effect, when overfit to the training data's dominant patterns. In contrast, SFT-C successfully maintains a **significantly higher entropy** throughout training, preserving predictive diversity. The slightly higher training loss observed for SFT-C is not a sign of inferior learning but rather an indication that the model is actively avoiding collapse into a narrow mode, resulting in a smoother, more generalized distribution. Crucially, this enhanced smoothness directly translates to superior generation quality. The **sustained higher CIDEr and SPICE scores** for SFT-C (Figure 3, right panels) demonstrate that by preventing the distribution from becoming overly sharp, our CW-DPO explores a richer semantic space, consistently producing more accurate and diverse top-k candidates.

5.3. Phase-Twp: Quantitative Analysis of Squeezing Effect Suppression

To evaluate the effectiveness of our CW-DPO in mitigating the "squeezing effect", we construct an experiment based on the COCO Caption dataset. Specifically, we sample 10,000 simple examples as the training set and an additional 1,000 examples as a fixed *probe set* to analyze distributional changes. Starting from a unified base model pretrained with **Smoothed SFT**, we apply standard DPO and CW-DPO on the same training split and compare their effects on the output distributions over the probe set. We compute the **Total Variation (TV)** and **Jensen-Shannon (JS)** distances between the pre- and post-finetuning output probabilities, and visualize changes in the **Top-5 token distributions** for representative samples to provide qualitative insights. To further assess whether CW-DPO alleviates overconfidence and calibration degradation caused by unstable gradients, we include the **Expected Calibration Error (ECE)** as an evaluation metric. We also report **CIDEr** and **SPICE** scores on the full COCO test set to comprehensively assess generation quality.

Figure 4 reveals substantial differences in optimization dynamics between standard DPO and CW-DPO. From a global standpoint, the first plot shows that standard DPO exhibits significantly higher TV and JS divergence, typically around 0.45 for TV and 0.30 for JS, indicating that its learning process is overly influenced by simple samples. This leads to drastic shifts in the output distribution relative to the initial SFT model, as the model aggressively reallocates probability mass in response to uninformative gradients from easy negatives. In contrast, CW-DPO achieves much smaller divergences (e.g., approximately 0.15 for TV and 0.10 for JS), suggesting that it performs more stable and conservative updates. By down-weighting easy negatives, CW-DPO preserves the model's distributional structure while aligning with preferences, mitigating squeezing, and reducing risks of forgetting or collapse. These differences are more pronounced at the micro level. As illustrated in the middle plot of Figure 4 (cross-referenced with Figures 2 and 3), vanilla DPO drives most probability mass onto the top token, often surpassing 80%, creating a peaked distribution that suppresses alternatives. This overconfidence raises ECE (e.g., $0.12 \rightarrow 0.25$), degrading calibration. By contrast, CW-DPO updates more smoothly, keeping the top-1 token around 50–60%, preserving entropy, and stabilizing ECE at 0.08–0.10. Such dynamic improvements also yield higher generation quality, as shown in the right plot: on the full COCO test set, CW-DPO achieves superior CIDEr (142.6 vs. 137.2) and SPICE (25.8 vs. 24.2), reflecting not only greater accuracy but also richer linguistic diversity and semantics.

5.4. Ablation Study

To evaluate the contributions of each component in our CW-DPO, we conduct a comprehensive ablation study covering both training stages under identical data splits and hyperparameters. **Besides ablating the core algorithmic modules, we provide a further study in Appendix 12 to analyze the model's robustness to different negative sampling strategies, thereby decoupling algorithmic gains from the data generation process.** In Stage 1 (SFT), we evaluate **w/o Smooth SFT**, directly applying CW-DPO on the pretrained model to assess the need for smoothed initialization; **w/o Negative Sampling**, removing negative-sample constraints and reducing to standard SFT; and **w/o Soft Penalty**

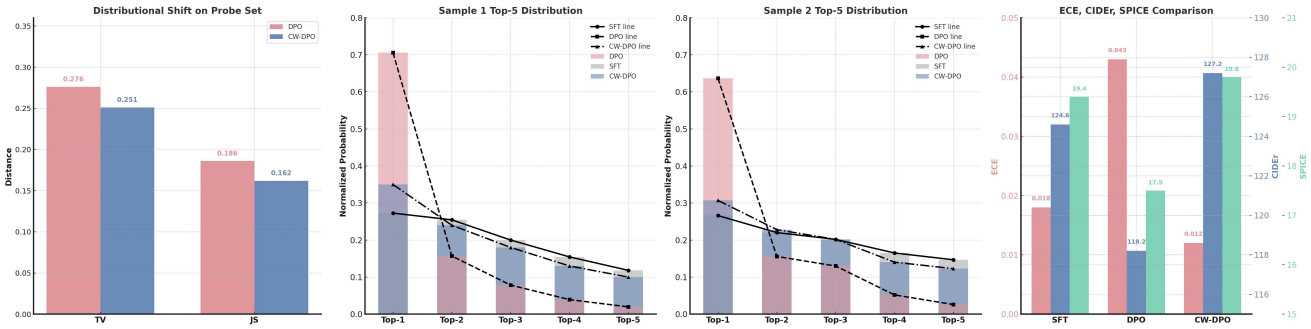

Figure 4. **CW-DPO alleviates the squeezing effect of vanilla DPO.** It yields smaller distribution shifts (left), smoother posteriors (middle), and improved generation quality with better calibration (right).

Table 3. Ablation study of **CW-DPO** on COCO Test, MMMU, and MMBench1.1.

Method	COCO Test				MMMU	MMBench1.1
	B@4	M	C	S	ACC	ACC
CW-DPO	**39.6**	**30.4**	**142.6**	**25.8**	**74.6**	**89.6**
Phase-One Ablation						
w/o Smooth SFT	34.6	28.4	137.6	24.4	71.8	86.3
w/o Negative Sampling	35.8	29.4	138.9	24.6	72.8	88.4
w/o Soft Penalty	36.2	29.7	139.2	24.8	73.2	88.7
Phase-Two Ablation						
w/o CW-DPO	36.7	28.8	140.7	24.7	72.9	86.7
w/o Cooling Weight	39.2	30.1	141.5	25.1	73.6	88.3
w/o Negative Filtering	36.1	27.9	137.4	24.3	73.4	87.4

(→ Hard Constraint), replacing the ReLU penalty with a hard constraint. In Stage 2 (DPO), we examine **w/o CW-DPO** (omitting the second-stage preference alignment), **w/o Cooling Weight** (fixing w_c to a constant (e.g., 0.7 or 1.0) instead of adaptive scaling), and **w/o Negative Filtering** (updating on all negatives with extremely easy ones, i.e., $\bar{\ell}_\theta \ll \ell_{\text{floor}}$).

Table 3 validates the independent contributions of each key component in **CW-DPO**. In Stage 1, removing Smooth SFT (w/o Smooth SFT) reduces CIDEr by about 5 points on COCO and also degrades performance on MMMU and MMBench1.1, indicating the importance of smoothed initialization for stable alignment. Further removing negative-sample constraints (w/o Negative Sampling) or replacing the soft ReLU penalty with a hard constraint (w/o Soft Penalty) also leads to consistent drops, showing that both negative-sample regularization and soft penalization are effective in alleviating overconfidence and improving generation quality. In Stage 2, omitting preference optimization (w/o CW-DPO) markedly reduces cross-task performance, confirming the need for competence-aware alignment. Using a fixed cooling weight (w/o Cooling Weight) achieves near CW-DPO CIDEr but lower MMMU and MMBench1.1 scores, underscoring the importance of adaptive scaling for generalization.

6. Limitations and Future Work

While **CW-DPO** demonstrates strong stability and generalization across diverse vision–language benchmarks, several limitations remain. First, our framework currently assumes access to paired preference data with reliable positive–negative supervision. Extending CW-DPO to fully unsupervised or weakly labeled settings, e.g., self-generated or noisy preferences, requires additional robustness mechanisms such as probabilistic confidence calibration or adaptive pseudo-label filtering. Second, although the cooling weight mitigates uninformative gradients, it introduces hyperparameters (τ, ℓ_{floor}) that may require tuning per dataset to maintain optimal gradient balance. A meta-learned or automatically scheduled variant could further improve efficiency. Third, our analysis focuses on the alignment dynamics of captioning-style VLMs; applying CW-DPO to interactive or long-horizon multimodal reasoning tasks (e.g., video QA or embodied agents) demands modeling temporal dependencies in learning dynamics. Future work will generalize CW-DPO to model-based alignment and dynamic preference settings for a unified multimodal alignment framework.

7. Conclusion

In this paper, we uncovered core instability issues in VLM preference-based finetuning via a fine-grained learning-dynamics perspective, focusing on the "squeezing effect" that causes uninformative gradients and unstable optimization. Our CW-DPO provides a principled two-stage solution, i.e., constrained SFT for loss landscape smoothing and competence-aware cooling weights to suppress easy negatives asymmetrically and adaptively. Extensive empirical results consistently and clearly demonstrate the strong superiority of our CW-DPO with faster convergence, stronger stability, and enhanced generalization.

8. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, and in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019. 2

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019. 6

[3] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. *On exact computation with an infinitely wide neural net*. Curran Associates Inc., Red Hook, NY, USA, 2019. 19

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 6

[5] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. Survey Certification, Featured Certification. 1

[6] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. 1

[7] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc. 1, 2

[8] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 1

[9] Kristian González Barman, Simon Lohse, and Henk W. de Regt. Reinforcement learning from human feedback in llms: Whose culture, whose values, whose perspectives? *Philosophy & Technology*, 38(2), 2025. 2

[10] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models, 2024. 1

[11] Shawn Im and Yixuan Li. Understanding the learning dynamics of alignment with human feedback, 2024. 2

[12] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc. 3, 19

[13] Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. A survey on human preference learning for large language models, 2024. 2

[14] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback, 2024. 1, 2, 3

[15] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1885–1894. JMLR.org, 2017. 3, 19

[16] Kyungmin Lee, Xiaohang Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models, 2025. 2

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 1

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296, 2024. 1, 2

[21] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Yike Yuan, Wangbo Zhao, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player?, 2024. 6

[22] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 13

[23] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

[24] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025.

[25] Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, et al. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*, 2025. 13

[26] Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 11241–11258. Association for Computational Linguistics, 2025. 2

[27] 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. Training language models to follow instructions with human feedback, 2022. 1, 2, 6

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[29] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 6, 19

[30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. 2

[31] Yi Ren and Danica J. Sutherland. Learning dynamics of llm finetuning, 2025. 1, 2, 3, 5, 19

[32] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org, 2020. 1

[33] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 2, 6

[34] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. 2, 6

[35] Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5311–5329, 2025. 1

[36] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. 2

[37] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. Reinforcement learning enhanced llms: A survey, 2025. 2

[38] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. 1

[39] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2

[40] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization, 2024. 2, 6

[41] Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Task preference optimization: Improving multimodal large language models with vision task alignment, 2025. 2

[42] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10610–10620, 2025. 2, 6

[43] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[44] Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024. 6

[45] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024. 1

[46] Jusheng Zhang, Kaitong Cai, Yijia Fan, Ningyuan Liu, and Keze Wang. MAT-agent: Adaptive multi-agent training optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[47] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. CF-VLM: Counterfactual vision-language fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[48] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm:counterfactual vision-language fine-tuning, 2025. 1

[49] Jusheng Zhang, Kaitong Cai, Xiaoyang Guo, Sidi Liu, Qinhan Lv, Ruiqi Chen, Jing Yang, Yijia Fan, Xiaofei Sun, Jian Wang, Ziliang Chen, Liang Lin, and Keze Wang. Mm-cot:a benchmark for probing visual chain-of-thought reasoning in multimodal models, 2025. 1

[50] Jusheng Zhang, Kaitong Cai, Qinglin Zeng, Ningyuan Liu, Stephen Fan, Ziliang Chen, and Keze Wang. Failure-driven workflow refinement, 2025. 1

[51] Jusheng Zhang, Yijia Fan, Kaitong Cai, Zimeng Huang, Xiaofei Sun, Jian Wang, Chengpei Tang, and Keze Wang. Drdiff: Dynamic routing diffusion with hierarchical attention for breaking the efficiency-quality trade-off, 2025. 1

[52] Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. GAM-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[53] Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. Gam-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning, 2025. 1

[54] Jusheng Zhang, Yijia Fan, Zimo Wen, Jian Wang, and Keze Wang. Tri-MARF: A tri-modal multi-agent responsive framework for comprehensive 3d object annotation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[55] Jusheng Zhang, Xiaoyang Guo, Kaitong Cai, Qinhan Lv, Yijia Fan, Wenhao Chai, Jian Wang, and Keze Wang. Hybridtoken-vlm: Hybrid token compression for vision-language models, 2025. 1

[56] Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. In *Forty-second International Conference on Machine Learning*, 2025. 1

[57] Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems. In *Forty-second International Conference on Machine Learning*, 2025. 1

[58] Jusheng Zhang, Yijia Fan, Kaitong Cai, Jing Yang, Jiawei Yao, Jian Wang, Guanlong Qu, Ziliang Chen, and Keze Wang. Why keep your doubts to yourself? trading visual uncertainties in multi-agent bandit systems. In *The Fourteenth International Conference on Learning Representations*, 2026. 1

[59] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023. 2