A Causal Marriage between VLM and IRM from Understanding to Reasoning (Appendix Included Version)

Ziliang Chen[1], Tianang Xiao[2], Jusheng Zhang[3], Yongsen Zheng[4], Yang Liu[3], Zhao-rong Lai[5], Liang Lin[1,3*]

[1]Research Institute of Multiple Agents and Embodied Intelligence, Peng Cheng Laboratory, [2]Hong Kong University of Science and Technology (Guangzhou), [3]Sun Yat-sen University, [4]Nanyang Technological University, [5]Jinan University

Abstract

Vision-Language Models (VLMs) like CLIP exhibit extraordinary out-of-distribution (OOD) generalization, while the theoretical foundations underlying this robustness remain largely unexplored. This work establishes a connection between CLIP and Invariant Risk Minimization (IRM), the principled paradigm to overcome OOD problems, through token-level causal representation learning. Our key insight is that CLIP's contrastive objective, when optimally trained, recovers modality-invariant causal factors at the word-and-phrase granularity. By decomposing text prompts into class-specific tokens (causal factors) and class-agnostic context tokens (environmental factors), we prove that a vocabulary-constrained InfoNCE objective becomes formally equivalent to IRM's invariance criterion. Grounded in this equivalence, we propose a mid-training paradigm aiming to inject invariant learning signals into pre-trained CLIP without architectural modification, yielding CLIP-IRM with superior OOD performance. We further extend this causal alignment to multimodal reasoning via using CLIP-IRM's invariant alignment scores as process-level rewards in reinforcement learning, effectively transplanting IRM's guarantees to robust sequential decision-making in Multimodal Large Language Models. Extensive experiments validate our theoretical framework and present substantial improvements in both multimodal OOD understanding and reasoning tasks.

1. Introduction

Vision-Language Models (VLMs), particularly those built upon the Contrastive Language-Image Pre-training (CLIP) [64] framework, have demonstrated astounding generalization performances with the excellence at a wide array of downstream applications in zero-shot and few-shot settings without task-specific training. This ability to transfer knowledge across diverse tasks suggests that they learn robust and widely applicable representations from vast quantities of image-text data. However, despite their empirical success and widespread adoption, the principles underlying their strong out-of-distribution (OOD) robustness remain largely unexplained. The current understanding is predominantly phenomenological, lacking a rigorous theoretical foundation that formally accounts for why these models generalize so effectively to novel data distributions. This gap between empirical performance and theoretical comprehension presents a significant challenge and opportunity to predictably analyze and improve them.

To build a principled understanding of VLM generalization, it is natural to turn to established frameworks designed explicitly for this purpose. Invariant Risk Minimization (IRM) [2] offers a rigorous paradigm for OOD generalization by learning predictors that rely on features causally linked to the outcome, while remaining invariant to spurious correlations across different environments. The objective of IRM conceptually aligns with the observed robustness of models like CLIP, motivating a formal connection between them. However, The two paradigms differ fundamentally: CLIP employs a dual-encoder architecture with a contrastive learning objective on unstructured data, whereas IRM is typically formulated as a bi-level optimization problem that requires explicit data environments. This architectural and objective mismatch presents a significant challenge to formalizing their relationship.

This paper bridges the conceptual and technical gap between CLIP and IRM from a review of token-level causal representation [14]. We posit that the semantic alignment in image-text pairs is governed by an underlying causal structure where shared, modality-invariant variables give rise to the content. Our key theoretical contribution demonstrates that an optimally trained CLIP model, through its contrastive objective, effectively learns to recover these invariant causal factors at the granularity of individual words and phrases. This token-aware perspective is the critical insight that enables the causal connection between the two frameworks: by decomposing text prompts into class-specific tokens (the causal factors) and class-agnostic context tokens

*indicate corresponding author;

(the environmental factors), we can reframe CLIP's prompt-based probing, then prove that a constrained variant of the InfoNCE objective becomes formally equivalent to the IRM objective to achieve explanable OOD generalization.

Grounded in this theoretical equivalence, we propose a practical and effective mid-training paradigm designed to explicitly inject invariance into pre-trained CLIP models. This approach does not alter the model's fundamental two-tower architecture but instead reconfigures the training data and supervision signals. By curating class and environment vocabularies from large-scale datasets, we construct training batches that guide the model to align representations based on class-relevant tokens while simultaneously discouraging reliance on spurious environmental context. This process effectively implements the IRM objective within the standard contrastive learning setup, yielding a new model, CLIP-IRM, with demonstrably superior OOD performance on a suite of challenging generalization benchmarks for multimodal understanding.

With regards to IRM and its reinforcement learning (RL) variant IPO (Invariant Policy Optimization [75]), we extend the benefits of this causal alignment from understanding to the more complex domain of multimodal RL-based reasoning. We leverage the causally robust CLIP-IRM model as a source of guidance for training Multimodal Large Language Models (MLLMs). Specifically, we use the invariant alignment score computed by our mid-trained model as a process-level reward signal within a GRPO [25] framework. This reward encourages the MLLM to generate reasoning chains that are not only correct but also grounded in the invariant, class-relevant features of the input. By integrating the principles of IPO through this novel reward mechanism, we successfully transplant the OOD guarantees of IRM to the sequential decision-making process of generative reasoning, enhancing the robustness and reliability of MLLMs.

Extensive experiments justify our theoretical framework and demonstrate substantial improvements in both multimodal OOD understanding and reasoning tasks.

2. Related Work

In this section, we provide some literature about CLIP, IRM, and the advance of vision-language reasoning by Multimodal LLM (MLLM). Since they can be connected from a causal lens, we also encourage the readers to go through the background knowledge of causality in Appendix.A.

CLIP and its fine-tuning techniques. CLIP (Contrastive Language-Image Pre-training) and its variants [15, 64, 77, 79] transfer visual representations via language supervision, achieving strong generalization across diverse recognition tasks [3, 22, 93]. Its core is contrastive pre-training on large image-text pairs, enabling open-vocabulary prediction with prompt templates (e.g., "a photo of a [CLASS]"), where category names are encoded as class-specific weights. Beyond zero-shot use, partial fine-tuning on target data further boosts performance. Open-vocabulary fine-tuning spans three lines: (1) **adapter-based tuning** [20, 92, 95], inserting small trainable modules into frozen encoders; (2) **prompt-tuning** [13, 100, 101], optimizing context embeddings of the template; and (3) **name-tuning** [53, 60], directly adjusting category-specific parameters. Hybrid methods combine these to gain complementary benefits [31, 53].

Invariant learning. IRM [2] addresses covariate shift across environments by seeking features that yield stable predictors. Let Φ be a feature extractor and w a classifier trained by ERM to predict $y \in \mathcal{Y}$ from $x \in \mathcal{X}$ in environment $e \in \mathcal{E}$, with risk $\mathcal{R}(w \circ \Phi; e) = \frac{1}{n}\sum_{i=1}^{n} r_e\big(w \circ \Phi(x), y\big)$, where $w \circ \Phi : \mathcal{X} \to \mathcal{Y}$ and r_e measures loss on samples from e. ERM-learned Φ often fails to generalize, so IRM seeks an invariant extractor Φ_{inv} enabling an optimal predictor across $\mathcal{E}$ by penalizing variance across environments while preserving accuracy, encouraging w to rely on environment-invariant, truly predictive features in $\mathcal{Y}$. As the IRM objective (Eq. 3) is a difficult bi-level problem, numerous surrogates have been proposed, including IRMv1 [2], REx [37], Bayesian IRM [46], Sparse IRM [102], ZIN [47], TIVA [80], and EIIL [16], among others.

Multimodal reasoning by MLLM. MLLM-based reasoning extends CoT from text-only LLMs to multimodal inputs via supervised trajectories distilled from Best-of-N, beam search, and MCTS ([74]). Recently, RL has emerged as a scalable alternative that optimizes reasoning paths with minimal annotations, driven by the R1 paradigm ([25]) and value-model-free or value-model-based algorithms such as GRPO ([18]), DAPO ([52]), VC-PPO ([21]), and VAPO ([28]). R1-style training with accuracy+format rewards has been adapted to MLLMs for general and domain-specific tasks, including medical VQA ([12]), visual reasoning ([45]), and two-stage text→vision transfer ([86]). Process- and step-wise rewards further enhance structural coherence in vision-language reasoning ([49]). Curriculum designs stabilize training and improve data efficiency in multimodal RL ([99]). In multimodal math, RL pipelines demonstrate strong gains without dense CoT supervision ([98]).

3. Preliminaries

In this section, we briefly introduce Contrastive Language-Image Pre-training (CLIP), including its learning objective and the routine to prompt its well-trained encoders for classification. Then we go through the formulation of Invariant Risk Minimization (IRM) [2] and its reinforcement learning variant, *i.e.*, Invariant Policy Optimization (IPO) [75]. In the rest of the paper, we are going to elaborate how the token-level causal representation bridge CLIP and IRM for multimodal understanding, also bridge CLIP-derived Large Vision Language Model and IPO for multimodal reasoning.

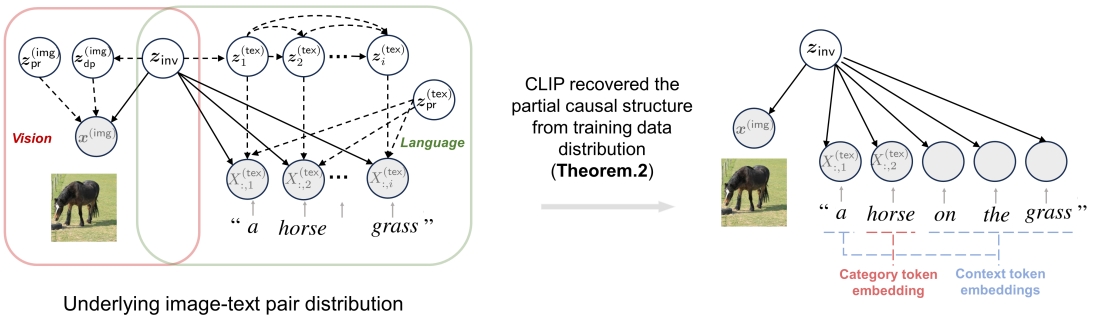


Figure 1. The comparison between **(a)** existing multimodal CRL theory [17] and **(b)** our CRL defined by Theorem.2 and Corollary.3. Our framework allows the analysis to CLIP with the word-and-phrase granularity, leading to connect CLIP's paradigm with IRM model.

3.1. Contrastive Language-Image Pre-training

The CLIP family [15, 30, 64] leverages semantically aligned image–text data via contrastive pre-training [26, 58]. Suppose $\langle x^{(\mathsf{img})}, x^{(\mathsf{tex})} \rangle \sim p_{\mathsf{mm}}\big(\boldsymbol{x}^{(\mathsf{img})}, \boldsymbol{x}^{(\mathsf{tex})}\big)$ denotes an image–text pair sampled from the multimodal joint distribution p_{mm} (that is, p_{mm} serves as the measure of cross-modal semantic coupling). CLIP employs an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$ to produce normalized features $f(x^{(\mathsf{img})})$ and $g(x^{(\mathsf{tex})})$, which are then used to form the InfoNCE objectives.

$$\min_{f,g} \mathbb{E}_{\mathcal{D}^{(K)} \sim p_{\mathsf{mm}}} \Big[\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) + \mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big)\Big]$$

$$\text{s.t. } \mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) = \sum_{i=1}^{K} -\log \frac{e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{j=1}^{K} e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_j^{(\mathsf{tex})})/\gamma\right)}},$$

$$\mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) = \sum_{i=1}^{K} -\log \frac{e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{j=1}^{K} e^{\left(f(x_j^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}} \tag{1}$$

where $\mathcal{D}^{(K)} = \{\langle x_i^{(\mathsf{img})}, x_i^{(\mathsf{tex})} \rangle\}_{i=1}^{K}$ denotes a training mini-batch of K image–text pairs, and $\{x_i^{(\mathsf{img})}, x_i^{(\mathsf{tex})}\}_{i=1}^{K}$ are the K pairs in the batch sampled from the joint density p_{mm}. With the encoders $f^*(\cdot)$ and $g^*(\cdot)$ trained, a text query $\mathbf{T}_y$ (i.e., a text prompt) corresponding to a class name y can be used to refer to any image x belonging to that class, *i.e.*,

$$P^{V}_{f^*,g^*}\big(\mathsf{class} = y|x\big) := \frac{\exp\Big(\mathsf{sim}\big(f^*(x), g^*(\mathbf{T}_y)\big)\Big)}{\sum_{\forall y' \in V} \exp\Big(\mathsf{sim}\big(f^*(x), g^*(\mathbf{T}_{y'})\big)\Big)}. \tag{2}$$

where $\mathsf{sim}(\cdot,\cdot)$ quantifies similarity between image and text features, and $P^{V}_{f^*,\mathbf{g}^*}\big(\mathsf{class} = y \mid x\big)$ denotes the probability that $x^{(\mathsf{img})}$ is assigned to class y, whose name appears in the open-vocabulary list V.

3.2. A Review of Invariant Predictors and Policy

IRM. Invariant Risk Minimization (IRM) [2] is a rigorous paradigm to formulate OOD generalization problem. As illustrated in Fig.2 (a), suppose that the latent variable $\boldsymbol{z}$ that generate each image $\boldsymbol{x}^{(\mathsf{img})}$ can be disentangled into the environment partition $\boldsymbol{z}^{(\mathsf{env})}$ simultaneously associated with the environment variable $\boldsymbol{e}$ and the category variable $\boldsymbol{y}$, and the causal partition $\boldsymbol{z}^{(\mathsf{cls})}$ only connected with the category variable $\boldsymbol{y}$. To this, IRM achieves the OOD class prediction by joint learning the feature extractor Φ the classifier w:

$$\min_{w,\Phi} \mathcal{R}_{\mathsf{IRM}}(w,\Phi) = \sum_{e\in\mathcal{E}} R^{(e)}(w\circ\Phi),$$
$$\text{s.t.}..w \in \arg\min_{\hat{w}} R^{(e)}(\hat{w}\circ\Phi),\ \forall e \in \mathcal{E} \tag{3}$$

where

$$R^{(e)}(w\circ\Phi) = \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, y\rangle \\ \sim P_e}}\left(-\log \frac{\exp\left(w_y^\top \Phi(x^{(\mathsf{img})})/\gamma\right)}{\sum_{y'\in\mathcal{Y}} \exp\left(w_{y'}^\top \Phi(x^{(\mathsf{img})})/\gamma\right)}\right)$$

represents the classification risk on samples from environment e, where examples are drawn from the environment-specific image distribution P_e. Here, $\mathcal{Y}$ and $\mathcal{E}$ are the supports of the distributions for the variables $\boldsymbol{y}$ and $\boldsymbol{e}$, respectively, and w_y denotes the classifier's weights for category y for all $y \in \mathcal{Y}$. The IRM solution (w^*, Φ^*) obtained from Eq. 3 yields class predictions that are deconfounded from variations across environment partitions for all $e \in \mathcal{E}$, providing a causal rationale for OOD generalization.

IPO. Invariant Policy Optimization (IPO) [75] transports the IRM principles from classification to reinforcement learning by replacing class prediction with action selection. Analogous to IRM, which learns (w, Φ) so that w is simultaneously optimal across environments $e \in \mathcal{E}$ given a shared representation Φ, IPO learns a representation $\Phi : \mathcal{O} \to \mathcal{H}$ together with an action-predictor (policy) $\pi : \mathcal{H} \to \mathcal{A}$ optimal across training domains $\{1, \ldots, n_d\}$:

$$\max_{\Phi,\pi}\ \mathcal{R}_{\mathsf{IPO}}(\pi,\Phi) = \sum_{d=1}^{n_d} R^d\big(\pi\circ\Phi\big)$$
$$\text{s.t.}\quad \pi \in \arg\max_{\bar{\pi}:\mathcal{H}\to\mathcal{A}} R^d\big(\bar{\pi}\circ\Phi\big),\ \forall d \in \{1,\ldots,n_d\}. \tag{4}$$

Comparing (4) with IRM in Eq. 3, the role of the classifier w and risk $R^{(e)}$ can equivalently refer to minimize the negative return $-R^d$ with regards to the policy π, respectively, while the shared representation Φ is constrained so

that the same π is optimal across domains. This invariance criterion encourages Φ to discard domain-specific, spurious observational factors (analogous to $\boldsymbol{z}^{(\mathsf{env})}$) and retain causal task features (analogous to $\boldsymbol{z}^{(\mathsf{cls})}$), thereby promoting OOD generalization of policies.

4. A Causal Connection across CLIP and IRM

In this section, we first provide the warm-up study to token-level causal representation [14] and how it associates with the optimal CLIP, then we show how the partial SCM recovered for this causal representation can be aligned with IRM. It provide the unified view to mid-train CLIP as IRM.

4.1. Warmup: Understanding CLIP via Token-level Causal Representation Identifiability

The CLIP-IRM connection rises from [14]. Concretely, with the language prior defined in [88], it assumes a general distribution behind image-text pairs trained and test by CLIP:

Assumption 1. (**Token-aware SCM of image-text data generation**, Fig.1.a)[14] The joint semantics between image-text pairs are derived via $z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}}$; given z_{inv}, the image-private partition $z_{\mathsf{pr}}^{(\mathsf{img})}$ and text-private partition $z_{\mathsf{pr}}^{(\mathsf{tex})}$ are drawn by $z_{\mathsf{pr}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{img})}}$, $z_{\mathsf{pr}}^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{tex})}}$; and the image-dependent partition is obtained by $z_{\mathsf{dp}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{img})}}(\cdot|z_{\mathsf{inv}})$. Suppose $z_i^{(\mathsf{tex})}$ as the token-dependent partition of the i^{th} token, and each of them is recursively sampled via $z_i^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_i^{(\mathsf{tex})}}(\cdot|z_{\mathsf{inv}}, \{z_j^{(\mathsf{tex})}\}_{j=1}^{i-1})$; then each image-text pair $\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}\rangle$ is generated through the nonlinear mixing functions $\mathbf{f}, \{\mathbf{g}_i\}_{i=1}^{\mathsf{k}_{\max}}$ to specify p_{mm}

$$\begin{aligned} x^{(\mathsf{img})} &:= \mathbf{f}\big(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{img})}, z_{\mathsf{pr}}^{(\mathsf{img})}\big); \\ X_{:,i}^{(\mathsf{tex})} &:= \mathbf{g}_i\big(z_{\mathsf{inv}}, \{z_j^{(\mathsf{tex})}\}_{j=1}^{i}, z_{\mathsf{pr}}^{(\mathsf{tex})}\big). \end{aligned} \tag{5}$$

where the nonparametric functions extend the text from a vector $x^{(\mathsf{tex})} \sim p_{\boldsymbol{x}^{(\mathsf{tex})}}$ to a k-column matrix $X^{(\mathsf{tex},\mathsf{k})} \sim p_{\mathbf{X}^{(\mathsf{tex},\mathsf{k})}}$, where $\forall \mathsf{k} \in \{1, \cdots, \mathsf{k}_{\max}\}$ indicates the sentence length and the i^{th} column $X_{:,i}^{(\mathsf{tex},\mathsf{k})}$ indicates the i^{th} token embedding. The sampling stops at k^{th} step if $\mathsf{k}=\mathsf{k}_{\max}$ or $X_{:,k}^{(\mathsf{tex})}$ reaches the embedding of [EOF].

Derived from the token-level understanding to p_{mm}, the block identifiability result is hold below:

Theorem 2. ***(Block-Identified Modal-invariant Alignment (Token-aware)*** *Fig.1.b) [14] Consider the image-text pairs generated by Assumption.1 . If their densities and mappings meet:* ***1).*** *$\mathbf{f}$ and $\mathbf{g}_i$ $(\forall i \in \{1, \cdots, \mathsf{k}_{\max}\})$ are diffeomorphisms;* ***2).*** *$\boldsymbol{z}^{(\mathsf{img})}$, $\boldsymbol{z}_i^{(\mathsf{tex})}$ $(\forall i \in \{1, \cdots, \mathsf{k}_{\max}\})$ are smooth and with continuous distributions $p_{\boldsymbol{z}^{(\mathsf{img})}} > 0$, $p_{\boldsymbol{z}_i^{(\mathsf{tex})}} > 0$ almost everywhere. Consider $f : \mathcal{X}_{\mathsf{img}} \to (0,1)^{n_{\mathsf{inv}}}$ and $g : \cup_i^{\mathsf{k}_{\max}} \mathcal{X}_{\mathsf{tex}}^{(i)} \to (0,1)^{n_{\mathsf{inv}}}$ as smooth functions that are trained to jointly minimize the functionals,*

$$\begin{aligned} \mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})} := \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}\rangle \\ \sim p_{\mathsf{mm}}}} \Big[||f(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})})|| \Big] \\ - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{X}^{(\mathsf{tex})})), \end{aligned} \tag{6}$$

where $H(\cdot)$ denotes the differential entropy of the random variables $f(\boldsymbol{x}^{(\mathsf{img})})$ and $g(\boldsymbol{X}^{(\mathsf{tex})})$ taking value in $(0,1)^{n_{\mathsf{inv}}}$. Then given the optimal image encoder f^* and the text encoder g^*, there exist invertible functions h_f and h_g satisfying the following decompositions, respectively:

$$f^* = h_f \circ \mathbf{f}_{1:n_{\mathsf{inv}}}^{-1}, \; g^* = h_g \circ \mathbf{g}_{1:n_{\mathsf{inv}}}^{-1} \tag{7}$$

Corollary 3. ***(CLIP are optimally aligned as token-level casual representation)*** *[14] The optimal encoders f^*, g^* in Theorem.2 are obtained if and only if $(f^*, g^*) = \arg\min_{f,g} \mathcal{L}_{\mathsf{InfoNCE}}^{(\mathsf{img}\to\mathsf{tex})} + \mathcal{L}_{\mathsf{InfoNCE}}^{(\mathsf{tex}\to\mathsf{img})}$ with infinite training pairs.*

Interpretation. Theorem 1 shows CLIP's optimal encoders recover, up to invertible maps, the modality-invariant block of a token-aware SCM, aligning images and text in shared causal coordinates at the word and phrase level. With InfoNCE, this targets class-relevant, environment-agnostic factors when prompts respect token roles. It enables CLIP's prompt probing as IRM: under vocabulary/context conditions, aligning to class tokens while excluding environment tokens yields a mid-training objective equivalent to IRM.

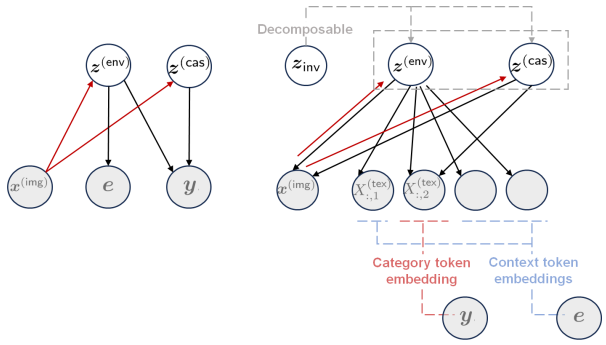
4.2. Bridging CLIP and IRM by A Partially Recovered SCM

Concretely, the nuanced connection between CLIP and IRM refers to their invariant prediction pipelines (Fig.2).

Definition 4. **(Necessary Conditions for Bridging CLIP and IRM)** If the prompt-based probing is consistent with the OOD class predictor in IRM, following rules are hold:

1. **Class-set consistency**. The class set $\mathcal{Y}$ should be consistent with the vocabulary V used in Eq.2, *i.e.*, $\mathcal{Y} = V$.
2. **Class-agnostic context.** Given a textual prompt $\mathbf{T}_{(y)}$ with its class token with respect to $\forall y \in \mathcal{Y}$, its context tokens $\forall t_{\mathsf{ctx}} \in \mathsf{set}(\mathbf{T}_{(\cdot)})$ satisfy $\{t_{\mathsf{ctx}}\} \cap \mathcal{Y} = \emptyset$.
3. **V-specific decomposibility of $\boldsymbol{z}_{\mathsf{inv}}$.** $\forall V = \mathcal{Y}$, there exists a decomposition of $\boldsymbol{z}_{\mathsf{inv}}$, *i.e.*, $\boldsymbol{z}_{\mathsf{inv}} = (\boldsymbol{z}^{(\mathsf{env})}, \boldsymbol{z}^{(\mathsf{cls})})$, such that if the i-th token $X_{:,i}^{(\mathsf{tex})}$ in Assumption.1 is a context token embedding, its generation process refers to
$$X_{:,i}^{(\mathsf{tex})} := \mathbf{g}_i\big(z^{(\mathsf{env})}, \{z_j^{(\mathsf{tex})}\}_{j=1}^{i}, z_{\mathsf{pr}}^{(\mathsf{tex})}\big). \tag{8}$$

Note that the rule 1,2 are straightforward and can be obviously fulfilled by the prompt-based probing in Eq.2. The rule 3 is fundamental for the variable-based feature recovery in OOD generalization: without its satisfaction, each element in the representation recovered from $\boldsymbol{z}_{\mathsf{inv}}$ would involve the environment information, leading to the failure



(a) OOD class prediction by IRM (b) CLIP's prompt-based probing (Eq.2) from a view of IRM

Figure 2. **(a)** SCM behind IRM to achieve OOD generalization **(b)** the CLIP-base prompting derived from Theorem.2 and Corollary.3. Suppose that the red arrow "→" in (a),(b) denote their image feature extractors (encoder) that reverse the generation direction in SCMs, then if Definition.9 is satisfied, the prompting holds the consistent classifying pipeline with IRM. In Theorem.5, we justify that a proper reconfiguration of InfoNCE can promise CLIP aligned with IRM through mid-training.

of OOD generalization regardless of environmental knowledge. Definition.4 paves the way to comprehend CLIP from an IRM viewpoint, whereas the necessary conditions do not promise the prompt-based classifier derived from CLIP has to be a OOD class predictor obtained by IRM. Beyond this, CLIP is built with a two-tower architecture therefore its learning objective is significantly different from IRM in Eq.3. So it is infeasible to employ existing IRM algorithms [39, 50] to train the encoders in CLIP.

To this end, we elaborate how to reframe CLIP as IRM in the mid-training stage. It is simply a vocabulary-supervised contrastive learning paradigm derived from Theorem.2:

Theorem 5. ***(CLIP Modality-Aligned as IRM)*** *Consider the image-text pairs generated by Assumption.1 and follow the condition rules in Definition.4, so that provided with the vocabulary V, we could identify the category name $y \in V$ and the environment name $e \in \mathcal{E}$ for each image-text pair,* i.e., $\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y, e\rangle \sim p_{\mathsf{mm}}$. *Suppose that the nonlinear mixing functions* $\mathbf{f}$, $\mathbf{g}_i$ $(\forall i \in \{1, \cdots, \mathsf{k}_{\mathsf{max}}\})$, *the latent variables* $\boldsymbol{z}^{(\mathsf{img})}$, $\boldsymbol{z}_i^{(\mathsf{tex})}$ $(\forall i \in \{1, \cdots, \mathsf{k}_{\mathsf{max}}\})$, *and encoder functions f, g are consistent with those in Theorem.2, and we define the functional* $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}$ *as the vocabulary-supervised learning objective derived from* $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}$:

$$\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}(f,g;V,\mathcal{E}) := \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y, e\rangle \\ \sim p_{\mathsf{mm}}}} \Big[\|f(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})}_{y/e})\|\Big] - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{X}^{(\mathsf{tex})}_{V/\mathcal{E}})), \tag{9}$$

where $X^{(\mathsf{tex})}_{y/e}$ *denotes the token sequence that contains y yet eliminates the environment token e from* $X^{(\mathsf{tex})}$, *and* $\boldsymbol{X}^{(\mathsf{tex})}_{V/\mathcal{E}}$ *indicates the random variable responds to their distribution, then the constrained modal-invariant alignment objective*

$$\min_{f,g} \mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}(f,g;V,\mathcal{E}),\ \mathsf{s.t.}\ f,g \in \min_{\hat{f},\hat{g}} \mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(\hat{f},\hat{g}) \tag{10}$$

is equivalent with the IRM objective in Eq.3.

Intuition. In brief, Theorem.5 implies that a simple data reconfiguration of InfoNCE losses ($\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}$ is consistent with $\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}$ and $\mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}$ according to Corollary.3 so $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}$ can also be derived from them), the bi-level InfoNCE objective consists of the vocabulary-supervised learning (Eq.9) with regards to the self-supervised learning can be treated as the IRM variant.

5. Mid-training CLIP for OOD Understanding

Grounded in the causal equivalence between CLIP and IRM in Theorem 5, we propose a mid-training paradigm that injects invariant learning signals into CLIP without altering its two-tower architecture. The key insight is to retain InfoNCE but restructure supervision and batches so that the encoders align along class-relevant, environment-agnostic causal coordinates. This is realized by vocabulary-supervised modal alignment and environment-token pruning, thereby the token-aware SCM ensures that the learned embeddings invertibly recover the invariant block.

We instantiate this paradigm by curating large-scale vocabularies over LAION: a class vocabulary $\mathcal{V}$ and an environment vocabulary $\mathcal{E}$ (Definition 4, rules 1–2). We extract class tokens, remove environment tokens, and concurrently synthesize environment-invariant pairs by swapping captions that share the same class but contain no environment tokens. These yield augmented batches $\mathcal{D}^{(K)}_{\mathcal{V}}$ that embody the constrained, token-level matching $X^{(\mathsf{tex})}_{y/e}$, enforcing alignment on the causal class subspace and discouraging reliance on spurious context. To unify standard CLIP pre-training with IRM-style constraints, we mid-train CLIP with the following objective, which directly implements the constrained alignment program in Eq. (10):

$$\begin{aligned}
&\min_{f,g}\ \mathbb{E}_{\mathcal{D}^{(K)},\mathcal{D}^{(K)}_{\mathcal{V}}\sim p_{\mathsf{mm}}}\Big[\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}_{\mathcal{V}}\Big) + \mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}_{\mathcal{V}}\Big)\Big] \\
&+\lambda\Big[\mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big) + \mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}\Big)\Big] \\
&\mathsf{s.t.}\ \mathcal{L}^{(\mathsf{img}\to\mathsf{tex})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}_{\mathcal{V}}\Big) = \\
&\sum_{i=1}^{K} -\log \frac{\sum_{<x_i^{(\mathsf{img})},x_i^{(\mathsf{tex})}>\sim w\in\mathcal{V}} e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{<x_i^{(\mathsf{img})},x_i^{(\mathsf{tex})}>\sim w\in\mathcal{V}}\sum_{j=1}^{K} e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_j^{(\mathsf{tex})})/\gamma\right)}}, \\
&\mathcal{L}^{(\mathsf{tex}\to\mathsf{img})}_{\mathsf{InfoNCE}}\Big(\mathcal{D}^{(K)}_{\mathcal{V}}\Big) = \\
&\sum_{i=1}^{K} -\log \frac{\sum_{<x_i^{(\mathsf{img})},x_i^{(\mathsf{tex})}>\sim w\in\mathcal{V}} e^{\left(f(x_i^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}{\sum_{<x_i^{(\mathsf{img})},x_i^{(\mathsf{tex})}>\sim w\in\mathcal{V}}\sum_{j=1}^{K} e^{\left(f(x_j^{(\mathsf{img})})^\top g(x_i^{(\mathsf{tex})})/\gamma\right)}}
\end{aligned} \tag{11}$$

where the first expectation term trains on $\mathcal{D}_{\mathcal{V}}^{(K)}$ to align image–text pairs in an environment-agnostic manner, thus realizing the invariant decomposition of Theorem 2 at token level. The second term (weighted by λ) preserves the coverage and diversity of standard CLIP pre-training on $\mathcal{D}^{(K)}$, stabilizing optimization and mitigating over-constraint. Under the conditions of Theorem 5, this single-stage program is equivalent to IRM's objective, avoiding bi-level optimization while preserving invariant prediction guarantees. Practically, we mine $\mathcal{V}$ and $\mathcal{E}$ at scale, apply token-level pruning that respects phrase granularity, and interleave augmented and original pairs during training. This mid-training stage can be inserted after initial pre-training on LAION, requiring only dataset curation and batch reconfiguration. The implementation details can be found in Appendix.C.

Few-shot Prompt-tuning. Beyond OOD robustness, the mid-trained model provides a stronger initialization for few-shot prompt tuning. Because class and environment factors are disentangled during mid-training, prompt templates adapt more reliably across environments, improving base-to-new transfer. In cross-dataset settings, the learned invariances reduce sensitivity to background, style, and source biases, yielding more stable calibration of image–text similarities and lower sample complexity in adaptation. Empirically, this leads to consistent improvements in zero-shot OOD performance and amplified gains.

6. CLIP-IRM as Process-Reward Guidance for Multimodal OOD Reasoning

Building on the mid-training paradigm that causally aligns CLIP with IRM (Theorem 5), we extend the invariant alignment principle to guide multimodal reasoning in reinforcement learning. Our goal is to use CLIP's token-aware invariant supervision as a self-supervised process-level reward [40] that shapes policy learning in MLLMs, thereby transplanting the OOD guarantees of IRM to RL policy over language tokens and visual observations.

From IPO to CLIP-IRM-guided policy. Reinforcement learning (RL) has become a primary avenue for improving multimodal reasoning, where a policy generates chain-of-thought and actions conditioned on images and textual context. IPO is equivalent with IRM as RL by enforcing that a shared representation Φ admits a single policy π that is optimal across domains (Eq.4). Leveraging the CLIP–IRM equivalence (Theorem5) and our mid-training strategy (Eq. 11), it is possible to instantiate an IPO-style framework in which the invariant representation is realized by CLIP's environment-agnostic token alignment, and the policy is realized by an MLLM decoder.

Despite so, in the CLIP-IRM view, the "policy" corresponds to the text encoder $g(\cdot)$, which barely support strong reasoning or long-horizon credit assignment due to the limitation of its bidirectional encoding and pretraining regime. Conversely, replacing the policy with a general MLLM decoder π_θ disconnects from the CLIP–IRM alignment of Theorem 5, since the decoder lacks the invariant supervision channel. This gap motivates a coupled architecture that preserves the IRM-aligned supervision while endowing the policy with generative reasoning capacity.

Coupled decoder–encoder to harvest process reward. We compose an MLLM text decoder with the CLIP text encoder via a sliding-window interface. Let π_θ autoregressively produce a token sequence $t_{1:T}$ conditioned on $\boldsymbol{x}^{(\mathsf{img})}$. At step k, we extract a window $t_{k-w+1:k}$ and feed it to the CLIP text encoder g to obtain $h_k^{(\mathsf{tex})} = g(t_{k-w+1:k})$. We pair $h_k^{(\mathsf{tex})}$ with image features $v^{(\mathsf{img})} = f(\boldsymbol{x}^{(\mathsf{img})})$ and compute a token-aware, vocabulary-constrained InfoNCE score using the mid-training batches $\mathcal{D}_{\mathcal{V}}^{(K)}$:

$$r_k^{(\mathsf{proc})} \triangleq \ell_{\mathsf{InfoNCE}}\big(v^{(\mathsf{img})}, h_k^{(\mathsf{tex})}; \mathcal{V}\big) \;-\; \alpha\, \ell_{\mathsf{env}}(t_{k-w+1:k}; \mathcal{E}), \tag{12}$$

where ℓ_{InfoNCE} follows Eq. (1) over the class vocabulary $\mathcal{V}$, and ℓ_{env} penalizes overlaps with the environment vocabulary $\mathcal{E}$. Because f, g are mid-trained by Eq. (11) to satisfy Theorem 5, maximizing $r_k^{(\mathsf{proc})}$ pushes π_θ to generate token trajectories whose induced representations align with the invariant class subspace so suits the IPO invariance criterion.

Data curation and batch construction. To train the process-reward model, we consider the principles to build the mid-training dataset: (i) curate a multimodal reasoning corpus with image–rationale–answer triplets; (ii) extract phrase-level class vocabulary $\mathcal{V}$ and environment vocabulary $\mathcal{E}$, prune environment phrases, and synthesize caption/rationale swaps within the same class to form $\mathcal{D}_{\mathcal{V}}^{(K)}$ in parallel to $\mathcal{D}^{(K)}$; (iii) for visual grounding, select high-confidence image patches via a calibrated proposal network and compute patch features $\{v_m\}$; then augment the process reward by local grounding

$$\begin{aligned} r_k^{(\mathsf{patch})} &= \max_{m \in \mathcal{M}} \; \mathrm{sim}\big(v_m,\; h_k^{(\mathsf{tex})}\big), \\ r_k^{(\mathsf{proc})} &\leftarrow r_k^{(\mathsf{proc})} + \beta\, r_k^{(\mathsf{patch})}. \end{aligned} \tag{13}$$

GRPO-based optimization. We adopt GRPO [25] as the policy optimizer to integrate the process rewards:

$$\begin{aligned} R(\tau) &= \sum_{k=1}^{T} \Big(r_k^{(\mathsf{task})} + \lambda_{\mathsf{proc}}\, r_k^{(\mathsf{proc})} \Big), \\ \mathcal{J}(\theta) &= \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_k \mathrm{GRPO}\big(A_k(\tau)\big) \right] - \beta_{\mathsf{KL}}\, \mathrm{KL}\big(\pi_\theta || \pi_{\theta_0}\big). \end{aligned} \tag{14}$$

where $r_k^{(\mathsf{task})}$ is measured by the correctness of answer/step and the format, and π_{θ_0} is a supervised reference. For stability: (a) normalize $r_k^{(\mathsf{proc})}$ per batch; (b) anneal λ_{proc}; (c) freeze f, g early, then optionally fine-tune them at a small

Table 1. Zero-shot generalization results of state-of-the-art OOD generalization baselines, competitive MLLM baselines, and CLIP-IRM in the evaluation protocol [91].

	PACS	VLCS	OfficeHome	NICO++	DomainNet	Avg
			OOD generalization models			
ERM	85.8	78.4	68.0	79.6	47.4	71.8
SWAD	88.1	79.0	71.4	80.8	49.6	73.8
RSC	86.8	78.2	67.9	79.7	47.3	72.0
GroupDRO	85.6	78.2	68.0	79.7	44.9	71.3
Fishr	86.6	78.0	67.7	79.6	47.2	71.8
CORAL	86.7	78.1	67.8	79.5	47.5	71.9
MMD	86.4	64.4	67.2	69.7	47.3	67.0
SagNet	85.6	78.0	66.9	79.2	46.5	71.2
IRM	84.7	78.1	68.2	79.7	47.3	71.6
Mixup	83.4	78.2	70.0	79.8	48.0	71.9
			Multimodal foundation models			
CLIP	97.7	73.4	85.4	88.7	76.7	83.4
BLIP-2	**100**	93.7	52.7	67.3	50.8	72.9
QWEN-VL	96.4	94.3	63.5	76.3	36.5	73.4
LLaVa	98.0	**97.5**	73.6	84.9	48.0	80.4
GPT-4V	96.9	87.2	84.8	88.0	74.8	86.3
Gemini	98.7	83.2	**89.7**	89.7	75.9	87.4
			Ours			
CLIP-IRMv1	95.1	78.8	83.9	87.7	72.7	83.6
CLIP-IRMv2	98.6	83.3	88.3	**91.8**	**78.1**	**88.0**

learning rate on interleaved $\mathcal{D}_{\mathcal{V}}^{(K)}/\mathcal{D}^{(K)}$ minibatches. The implementation details are found in SM.

7. Experiments

In this section, we aim to validate our CLIP-IRM connection on its technical contributions from multimodal understanding to reasoning. Main results are proposed in this section, and the ablation and analysis are in Appendix.D.

7.1. Multimodal Understanding by CLIP

Here we consider two evaluation setups to justify our mid-training strategy to CLIP. First, we achieve the mid-training to obtain the new CLIP model, *i.e.*, CLIP-IRM, then evaluated by its zero-shot inference on remarkable OOD generalization benchmarks. It helps verify whether our mid-training strategy can reap the theoretical merit from IRM. Second, we fine-tune CLIP-IRM with diverse post-tuning approaches, then observe whether their feasibility can be extended on top of our methodology.

7.1.1. Out-of-Distribution Generalization

Benchmarks and Baselines: PACS [42], VLCS [34], OfficeHOME [84], DomainNet [62], NICO++ [97]. In order to justify our theorems, we compare CLIP-IRM with some state-of-the-art OOD generalization baselines [91], *i.e.*, ERM [82], SWAD [10], RSC [27], GroupDRO [67], Fishr [65], CORAL [78], MMD [43], SagNet [56], IRM [2, 39], Mixup [89]; and also with various multimodal foundation models including CLIP [64], BLIP-2 [44], QWEN-VL [5], LLaVa [48], GPT4-V [90], and Gemini [81]. The OOD generalization baselines are trained with the training split in the OOD generalization benchmarks to facilitate their evaluation, while the foundation model baselines employ prompt to facilitate their zero-shot inference of the classes in the benchmarks. To ensure a fair comparison, we propose a variant *CLIP-IRMv1* aligned with all OOD generalization baselines in their image feature extractors using Vit-B/16; then the other variant *CLIP-IRMv2* with a stronger image-encoder Vit-L/14 backbone to compare with the more powerful foundation model baselines.

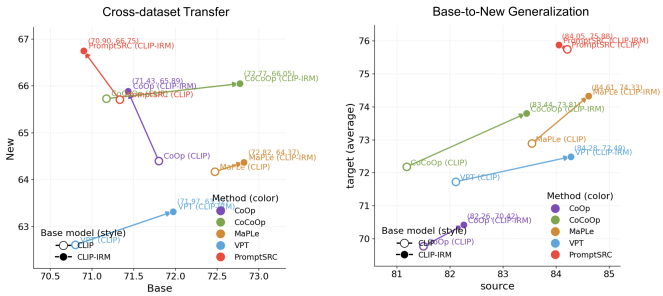


Figure 3. The prompt-tuning results based on CLIP and CLIP-IRM in Base-to-New generalization and Cross-dataset transfer.

Results. The accuracy are observed in Table.2 where the top-1 and top-3 baselines in each benchmark have been highlighted with boldface and underline. The baselines present diverse behaviors across the benchmarks. Specifically, traditional OOD baselines almost fall behind foundation models, particularly in PACS and VLCS. CLIP-IRMv1 employs the same backbone with those baselines while thanks to scaling laws on LAION data, it outperforms them across all domains with large margins. Open-source foundation models present a large variance across the benchmarks: despite the impressive results in PACS and VLCS, their performances significantly drop in OfficeHome and DomainNet. Instead, the proprietary foundation models, *i.e.*, GPT4-V and Gemini, more robustly perform and win the top-3 in the average score. Despite so, CLIP-IRMv2 outperform all baselines, particularly, exceeding the non-proprietary models with the leap over 4.6%, and also beats down the best proprietary foundation models in the most challenging OOD scenarios, NICO++ and DomainNet benchmarks.

7.1.2. Prompt-tuning Generalization

Benchmarks and Baselines: Given the CLIP-IRM model mid-trained by our method, we are interested to observe whether the post-training, in particular, prompt-tuning with few-shot examples can further benefited from it. Specifically, we evaluate CLIP-IRM and its original version with their performances in Base2New generalization and Cross-dataset generalization benchmarks. Beyond this, we also employ three single-modality prompt-tuning variants (*i.e.*, CoOp [101], CoCoOp [100], VPT [4]), two multimodal

prompt-tuning approach (Maple [31], PromptSRC [32]). Among the single-modality baselines, CoOp and CoCoOp are designed for tuning textual prompts yet VPT focuses on visual prompt. They combine with the multmodal prompt-tuning baselines to comprehensively validate the model transfer of CLIP-IRM and CLIP.

Main results. In Fig.3, across both settings—Base-to-New generalization (11 sub-domains) and cross-dataset source→target transfer (6 sub-datasets)—replacing CLIP with CLIP-IRM consistently shifts every prompt-tuning method to a better accuracy frontier. In Base-to-New, CLIP-IRM improves new-class performance for all five baselines and improves base-class performance for 3/5. In cross-dataset transfer, CLIP-IRM wins on target domains for all 5 baselines and on source domains for 4/5 (details in Appendix D). More specifically, two observations justify CLIP-IRM as a stronger post-training substrate. First, the paired points (CLIP vs. CLIP-IRM) move predominantly north-east, indicating new-domain gains without sacrificing base accuracy. The largest new-class jump appears with PromptSRC, but the pattern holds for CoOp, CoCoOp, MaPLe, and VPT, suggesting the effect is architecture-agnostic. Second, target-domain gains exceed source-domain gains, implying better robustness to genuine distribution shifts rather than improved memorization.

7.2. Vision-Language Reasoning by CLIP-IRM-Guided Process-Rewarded Policy

In this section, we aim to evaluate that MLLM trained by process-rewarding approach armed with GRPO, can yield more powerful and robust multimodal reasoning baselines.

Training Dataset and Evaluation: We used the EasyR1[1] framework built on verL [73], then adopt Qwen2.5-VL-7B-Instruct [6] as our MLLM base for training. Our training set was sourced from Geometry3K [59], with 2.1K samples focused on geometric problems, and MMK12 [19], with 6.4K samples that cover diverse K-12 math topics. To prevent reward hacking and model guessing, all questions were converted from multiple-choice to a free-form format. Our evaluation methodology assessed performance across two key dimensions. First, we measured out-of-domain generalization on 5 benchmarks: 4 for OOD multimodal reasoning (MathVerse [96], MathVision [87], MathVista [51], and WeMath [63]) and the rest for evaluate the MLLM robustness on hallucination (HallusionBench [24]). We evaluated in-domain performance by comparing our method, against the vanilla GRPO baseline on the Geometry3K test set. To ensure consistent assessment, we employed greedy decoding for generating responses and utilized Gemini-2.5-pro as a judge model to parse the outputs.

Main results. After trained on Geometry3K (Fig.4), GRPO augmented by our CLIP-IRM process reward delivers consistent OOD gains across MathVista, WeMath, MathVerse, and MathVision: the red curves steadily overtake GRPO and the margins grow with training, especially on the harder MathVerse / MathVision, evidencing stronger transfer under distribution shift. Robustness extends to faithfulness: on HallusionBench our method maintains higher accuracy throughout, indicating reduced hallucination via visually grounded, environment-pruned rewards. Importantly, these gains come without sacrificing in-domain performance: on Geometry3K (Acc reward), our approach matches early learning but surpasses the baseline at scale, yielding higher final reward and better sample efficiency. Mechanistically, converting mid-trained CLIP-IRM's IRM-equivalent alignment into step-wise process rewards (with optional patch grounding) effectively couples IPO-style invariance with GRPO, guiding trajectories that remain optimal across domains while staying image-grounded. Overall, the curves substantiate that GRPO+CLIP-IRM improves OOD generalization, curbs hallucination, and boosts ID reward—without changing the policy architecture.

In Appendix.D, we provide more through experimental statistics to offer deeper insights to our methodology.

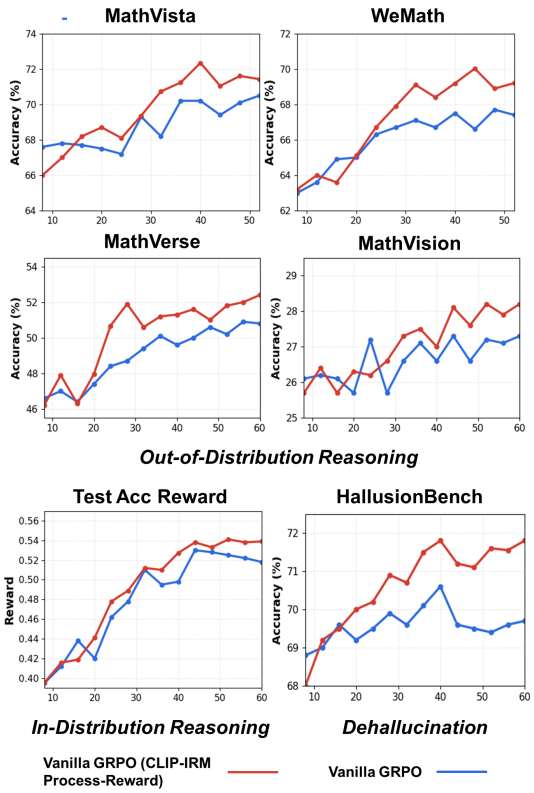


Figure 4. The vision-language reasoning results based on 4 Out-of-Distribution (OOD) reasoning benchmarks, 1 dehallucination benchmark, and the test set of Geometry3K, *i.e.*, In-Distribution (ID) reasoning results on Acc reward.

[1] https://github.com/hiyouga/EasyR1

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, and Kush R Varshney. Learning identifiable and interpretable latent models of high-dimensional data. *arXiv preprint arXiv:2002.02893*, 2022. 14

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 1, 2, 3, 7

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 2

[4] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022. 7

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 7

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8

[7] Johann Brehmer, Julius Von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Weakly supervised causal representation learning. *arXiv preprint arXiv:2010.15794*, 2022. 14

[8] Simon Buchholz, Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Learning identifiable representations that support sample-efficient intervention. *arXiv preprint arXiv:2302.01828*, 2023. 14

[9] Colin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent concepts in language models with contrastive search. *arXiv preprint arXiv:2210.14922*, 2023. 14

[10] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34: 22405–22418, 2021. 7

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 14

[12] Xi Chen et al. Medvlm-r1: Reinforcement learning for medical visual qa reasoning. *arXiv preprint arXiv:2502.XXXX*, 2025. 2

[13] Ziliang Chen, Xin Huang, Quanlong Guan, Liang Lin, and Weiqi Luo. A retrospect to multi-prompt learning across vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22190–22201, 2023. 2

[14] Ziliang Chen, Tianang Xiao, Jusheng Zhang, Yongsen Zheng, and Xipeng Chen. Understanding hardness of vision-language compositionality from a token-level causal lens. *arXiv preprint arXiv:2510.26302*, 2025. 1, 4, 14

[15] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 2, 3

[16] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021. 2

[17] Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[18] Yuxiao Dong et al. Group relative policy optimization for reasoning llms. In *NeurIPS*, 2024. 2

[19] Zongkai Liu Zhixiang Zhou Quanfeng Lu Daocheng Fu Tiancheng Han Botian Shi Wenhai Wang Junjun He Kaipeng Zhang Ping Luo Yu Qiao Qiaosheng Zhang Wenqi Shao Fanqing Meng, Lingxiao Du. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2407.08739*, 2024. 8

[20] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 2

[21] Peng Gao et al. Value-calibrated ppo for long chain-of-thought. *arXiv preprint arXiv:2504.XXXX*, 2025. 2

[22] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 2

[23] Luigi Gresele, Julius von Kügelgen, Ricardo P Monti, Bernhard Schölkopf, and Kun Zhang. Causal discovery in a binary setting with interventions. *arXiv preprint arXiv:2010.14241*, 2021. 14

[24] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 8

[25] Daya Guo, Xiaoyu Liu, Shizhe Zhou, et al. Deepseek r1: Incentivizing reasoning in large language models via reinforcement learning. *arXiv preprint arXiv:2501.XXXX*, 2025. 2, 6

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[27] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *European conference on computer vision*, pages 124–140. Springer, 2020. 7

[28] Zhen Huang et al. Vapo: Value-model-based augmented ppo for long cot. *arXiv preprint arXiv:2504.XXXX*, 2025. 2

[29] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–438, 1999. 14

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[31] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 2, 8

[32] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023. 8

[33] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. 14

[34] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012. 7

[35] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *International conference on machine learning*, pages 2668–2677, 2018. 14

[36] Bohdan Kivva, Marc Vuffray, and Bryon Aragam. Identifiability of latent-variable models with arbitrarily many views. *arXiv preprint arXiv:2210.00063*, 2022. 14

[37] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021. 2

[38] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Disentanglement of grouped factors of variation by leveraging partial group supervision. *arXiv preprint arXiv:2010.08226*, 2021. 14

[39] Zhao-Rong Lai and Wei-Wen Wang. Invariant risk minimization is a total variation model. *arXiv preprint arXiv:2405.01389*, 2024. 5, 7

[40] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025. 6

[41] Florian Leeb, Julius von Kügelgen, Bernhard Schölkopf, and Michel Besserve. Causal concept embedding models. *Advances in Neural Information Processing Systems*, 35: 23668–23681, 2022. 14

[42] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 7

[43] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 7

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 7

[45] Wei Li et al. Vision-r1: Towards rl-enhanced visual reasoning. *arXiv preprint arXiv:2504.XXXX*, 2025. 2

[46] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030, 2022. 2

[47] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542, 2022. 2

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 7

[49] Ming Liu et al. R1-vl: Step-wise rl for vision-language reasoning. *arXiv preprint arXiv:2503.XXXX*, 2025. 2

[50] Chaochao Lu, Yuhuai Wu, Jose Miguel Hernandez-Lobato, and Bernhard Scholkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021. 5

[51] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 8

[52] Tian Luo et al. Dapo: Dynamic sampling policy optimization for reasoning. *arXiv preprint arXiv:2503.XXXX*, 2025. 2

[53] Fengmao Lv, Changru Nie, Jianyang Zhang, Guowu Yang, Guosheng Lin, Xiao Wu, and Tianrui Li. Rethinking the effect of uninformative class name in prompt learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8345–8354, 2024. 2

[54] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35: 17359–17372, 2022. 14

[55] Ricardo P Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with hidden confounders using independent component analysis. *arXiv preprint arXiv:1906.08773*, 2019. 14

[56] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8690–8699, 2021. 7

[57] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020. 14

[58] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[59] Shibiao Jiang Liang Qiu Siyuan Huang Xiaodan Liang Song-Chun Zhu Pan Lu, Ran Gong. Inter-gps interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2407.08739*, 2024. 8

[60] Sarah Parisot, Yongxin Yang, and Steven McDonagh. Learning to name classes for vision and language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23477–23486, 2023. 2

[61] Dong-Sub Park, Wilson Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Eduard Hovy, and Quoc V Le. Speech-t5: Unifying speech generation and speech recognition via a single t5-based model. *arXiv preprint arXiv:2110.07205*, 2021. 14

[62] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 7

[63] Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070, 2025. 8

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 7

[65] Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 18347–18377. PMLR, 2022. 7

[66] Aida Ravichander, Eduard Hovy, and Richard M Pang. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:2004.09384*, 2020. 14

[67] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 7

[68] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019. 14

[69] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 14

[70] Sebastian Schwettmann, Florian Leeb, Bernhard Schölkopf, and Michel Besserve. Concept embedding models: A case study in toxicology. *arXiv preprint arXiv:2301.11823*, 2023. 14

[71] Sebastian Schwettmann, Florian Leeb, Bernhard Schölkopf, and Michel Besserve. Towards a theoretical framework for concept discovery. *arXiv preprint arXiv:2305.18728*, 2023. 14

[72] Anna Seigal and Yuesong Shen. Identifiability of deep generative models with structural constraints. *arXiv preprint arXiv:2006.07899*, 2021. 14

[73] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025. 8

[74] Bowen Shi et al. Planning to think: Best-of-n, beam search and mcts for multimodal cot. In *ICLR*, 2024. 2

[75] Anoopkumar Sonar, Vincent Pacelli, and Anirudha Majumdar. Invariant policy optimization: Towards stronger generalization in reinforcement learning. In *Learning for Dynamics and Control*, pages 21–33. PMLR, 2021. 2, 3

[76] Chandler Squires, Yue Wu, Kun Zhang, and Bryon Aragam. Causal-learn: Causal discovery in python. *The Journal of Machine Learning Research*, 24(1):14781–14787, 2023. 14

[77] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024. 2

[78] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 7

[79] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[80] Xiaoyu Tan, Lin Yong, Shengyu Zhu, Chao Qu, Xihe Qiu, Xu Yinghui, Peng Cui, and Yuan Qi. Provably invariant learning without domain information. In *International Conference on Machine Learning*, pages 33563–33580. PMLR, 2023. 2

[81] Gemini Team, R Anil, S Borgeaud, Y Wu, JB Alayrac, J Yu, R Soricut, J Schalkwyk, AM Dai, A Hauth, et al. Gemini: A family of highly capable multimodal models, 2024. *arXiv preprint arXiv:2312.11805*, 10, 2024. 7

[82] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. 7

[83] Kush R Varshney. On the identifiability of nonlinear latent variable models. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2402–2406, 2017. 14

[84] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 7

[85] Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. Non-linear identifiability of causal representations from temporal sequences. *arXiv preprint arXiv:2006.15059*, 2021. 14

[86] Jian Wang et al. Lmm-r1: Two-stage rl for multimodal reasoning. *arXiv preprint arXiv:2502.XXXX*, 2025. 2

[87] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 8

[88] Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *Advances in Neural Information Processing Systems*, 34:16158–16170, 2021. 4

[89] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020. 7

[90] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 7

[91] Han Yu, Xingxuan Zhang, Renzhe Xu, Jiashuo Liu, Yue He, and Peng Cui. Rethinking the evaluation protocol of domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21897–21908, 2024. 7

[92] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 2

[93] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[94] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024. 17

[95] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2

[96] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 8

[97] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16036–16047, 2023. 7

[98] Yuxin Zhang et al. Mm-eureka: Rl for multimodal mathematical reasoning. *arXiv preprint arXiv:2503.XXXX*, 2025. 2

[99] Guanghao Zhou et al. Curr-reft: Curriculum rl for multimodal reasoning. *arXiv preprint arXiv:2503.XXXX*, 2025. 2

[100] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 7

[101] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 7

[102] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022. 2

A. Background of Causality

Structural Causal Models (SCMs). The concept of SCM pioneered by Judea Pearl, have become a cornerstone of modern causal inference. They provide a mathematical framework for representing causal relationships within a system. An SCM consists of a set of variables and a set of equations that describe how each variable is determined by others in the model. This framework allows us to not only model statistical associations but also to predict the effects of interventions and to reason about counterfactuals. At its core, an SCM is defined by a collection of endogenous (or child) variables, whose values are determined by other variables within the model, and exogenous (or parent) variables, which are external to the model and treated as random noise or unobserved influences. The relationships between these variables are specified by structural equations, which are deterministic functions that define how each endogenous variable is generated from its direct causes and an associated exogenous noise term. The power of SCMs lies in their ability to make the causal assumptions explicit. By defining the causal graph—a directed acyclic graph (DAG) where nodes represent variables and directed edges represent causal relationships—we can analyze the flow of causal influence and determine which variables are causes and which are effects. This explicit representation is crucial for tasks such as identifying causal effects from observational data, understanding confounding bias, and achieving robust predictions under distributional shifts.

To pave the way for understanding the specific assumption for multimodal data, let's first define a general SCM using a consistent LaTeX format. This will introduce the core components and notation, which are then specialized in the assumption you provided.

A Structural Causal Model (SCM) is formally defined as a tuple $\mathcal{M} := \langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u}) \rangle$, where:

$\boldsymbol{V} = V_1, \ldots, V_n$ is a set of endogenous variables. These are the variables whose values are determined by other variables within the model. In the context of your assumption, the observed data, such as an image $x^{(\mathsf{img})}$ and a text description $x^{(\mathsf{tex})}$, are considered endogenous.

$\boldsymbol{U} = U_1, \ldots, U_n$ is a set of exogenous variables. These are mutually independent random variables that represent unobserved background conditions or noise. They are the ultimate sources of randomness in the model. In your assumption, the latent variables z_{inv}, $z^{(\mathsf{img})}\mathsf{dp}$, $z^{(\mathsf{img})}\mathsf{pr}$, $z^{(\mathsf{tex})}\mathsf{dp}$, and $z^{(\mathsf{tex})}\mathsf{pr}$ can be thought of as being determined by exogenous sources of variation.

$\mathcal{F} = f_1, \ldots, f_n$ is a set of structural equations, where each function f_i assigns a value to the corresponding endogenous variable V_i based on its direct causes $\mathsf{pa}(V_i) \subseteq \boldsymbol{V} \setminus V_i$ and its associated exogenous variable U_i:

$$V_i := f_i(\mathsf{pa}(V_i), U_i) \tag{15}$$

This equation states that the value of V_i is causally determined by the function f_i of its parents $\mathsf{pa}(V_i)$ and the exogenous noise U_i. $P(\boldsymbol{u})$ is a probability distribution over the exogenous variables $\boldsymbol{U}$.

Now, let's connect this general definition to the variables in your specific SCM assumption for image-text data generation. The assumption posits a hierarchical generation process that can be mapped onto the SCM framework. In particular, Exogenous Variables: The fundamental sources of variation are the latent variables drawn from their respective prior distributions: $z_{\mathsf{inv}} \sim p_{\boldsymbol{z}_{\mathsf{inv}}}$: The modal-invariant feature. $z_{\mathsf{pr}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{img})}}$: The image-private feature. $z_{\mathsf{pr}}^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_{\mathsf{pr}}^{(\mathsf{tex})}}$: The text-private feature. The dependent partitions, $z_{\mathsf{dp}}^{(\mathsf{img})}$ and $z_{\mathsf{dp}}^{(\mathsf{tex})}$, are also influenced by exogenous noise, but their generation is conditioned on z_{inv}. Endogenous Variables: These are the variables whose values are generated within the model. This includes the dependent latent variables and the final observed data: $z_{\mathsf{dp}}^{(\mathsf{img})}$: The image-dependent partition, generated based on $z\mathsf{inv}$. $z_{\mathsf{dp}}^{(\mathsf{tex})}$: The text-dependent partition, generated based on $z\mathsf{inv}$. $x^{(\mathsf{img})}$: The generated image. $x^{(\mathsf{tex})}$: The generated text. Structural Equations: The assumption provides the structural equations for the final observed variables, $x^{(\mathsf{img})}$ and $x^{(\mathsf{tex})}$:

$$x^{(\mathsf{img})} := \mathbf{f}\big(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{img})}, z^{(\mathsf{img})}\mathsf{pr}\big); \quad x^{(\mathsf{tex})} := \mathbf{g}\big(z_{\mathsf{inv}}, z_{\mathsf{dp}}^{(\mathsf{tex})}, z^{(\mathsf{tex})}\mathsf{pr}\big), \tag{16}$$

There are also implicit structural equations for the dependent partitions:

$$z_{\mathsf{dp}}^{(\mathsf{img})} \sim p_{\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{img})}}(\cdot|z_{\mathsf{inv}}) \quad z_{\mathsf{dp}}^{(\mathsf{tex})} \sim p_{\boldsymbol{z}_{\mathsf{dp}}^{(\mathsf{tex})}}(\cdot|z_{\mathsf{inv}}) \tag{17}$$

These conditional distributions can be expressed as structural equations with their own exogenous noise terms. For example, $z_{\mathsf{dp}}^{(\mathsf{img})} := h_{\mathsf{img}}(z_{\mathsf{inv}}, U_{\mathsf{img_dp}})$, where $U_{\mathsf{img_dp}}$ is an exogenous noise variable.

By laying out the SCM in this manner, we can clearly see the causal dependencies. The modal-invariant feature z_{inv} is a common cause of both the image and the text, which is what creates the "mutual semantics" between them. The private features, $z^{(\mathsf{img})}\mathsf{pr}$ and $z^{(\mathsf{tex})}\mathsf{pr}$, account for the variability within each modality that is independent of the other. The dependent

partitions, $z^{(\mathsf{img})}\mathsf{dp}$ and $z^{(\mathsf{tex})}\mathsf{dp}$, represent stylistic or content variations that are specific to a modality but are still influenced by the core shared semantics. This detailed causal structure is what allows for a rigorous analysis of how a model like CLIP might be able to disentangle and recover the causally meaningful feature z_{inv}.

Causal representation learning (CRL) and concept discovery. In recent years, SCMs have found significant application in representation learning. In particular, causal representation learning (CRL) [68, 69] aims to learn the latent generative factors behind high-dimensional data. This exciting field has seen significant progress in the last few years [1, 7, 8, 23, 33, 36, 38, 41, 55, 72, 76, 83]. A fundamental perspective in this field is to ensure that the model parameters we attempt to recover are identifiable [29, 33, 85]. Concept discovery is an important sub-field of machine learning which extracts human-intepretable concepts from pre-trained models. We do not attempt to list the numerous works in this direction, see e.g., [9, 11, 35, 54, 57, 61, 66, 70, 71, 76].

B. Principled Certification

The proof of Theorem.2 and Corollary.3 refer to [14], and we only need to prove Theorem.5.

B.1. Proof of Theorem.5

Our proof can be divided to prove the necessary and sufficient conditions for the equivalence between Eq.10 and IRM. Combine them and we obtain the proof.

B.1.1. The necessary condition: IRM $\rightarrow$ Eq.10.

It implies that for all classifier w and feature extractor Φ in Eq.3,

$$\begin{aligned}\min_{w,\Phi} \mathcal{R}_{\mathsf{IRM}}(w,\Phi) = \sum_{e\in\mathcal{E}} R^{(e)}(w,\Phi),\\ \mathsf{s.t.}.. w \in \arg\min_{\hat{w}} R^{(e)}(\hat{w},\Phi),\ \forall e\in\mathcal{E}\end{aligned} \tag{18}$$

where

$$R^{(e)}(w,\Phi) = \mathbb{E}_{\substack{\langle x^{(\mathsf{img})},y\rangle \\ \sim P_e}}\left(-\log \frac{\exp\left(w_y^\top \Phi(x^{(\mathsf{img})})/\gamma\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(w_{y'}^\top \Phi(x^{(\mathsf{img})})/\gamma\right)}\right),$$

we can define $f(w,\Phi)=f$, $g(w,\Phi)=g$ such that IRM *i.e.*, w, Φ leading to Eq.10, *i.e.*,

$$\min_{f,g} \mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}(f,g;V,\mathcal{E}),\ \mathsf{s.t.}\ f,g\in\min_{\hat{f},\hat{g}} \mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(\hat{f},\hat{g}) \tag{19}$$

where

$$\begin{aligned}\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}(f,g;V,\mathcal{E}) &:= \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y,e\rangle \\ \sim p_{\mathsf{mm}}}}\left[||f(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})}_{y/e})||\right] - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{X}^{(\mathsf{tex})}_{V/\mathcal{E}}));\\ \mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}} &:= \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})}\rangle \\ \sim p_{\mathsf{mm}}}}\left[||f(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})})||\right] - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{X}^{(\mathsf{tex})})).\end{aligned} \tag{20}$$

From similar deduction in Corollary.3, we have

$$\begin{aligned}R^{(e)}(w,\Phi) &= \mathbb{E}_{\substack{\langle x^{(\mathsf{img})},y\rangle \\ \sim P_e}}\left(-\log \frac{\exp\left(w_y^\top \Phi(x^{(\mathsf{img})})/\gamma\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(w_{y'}^\top \Phi(x^{(\mathsf{img})})/\gamma\right)}\right)\\ \iff R^{(e)}(w,\Phi) &= \mathbb{E}_{\substack{\langle x^{(\mathsf{img})},y\rangle \\ \sim P_e}}\left[||\Phi(x^{(\mathsf{img})}) - w_y||\right] - H_{P_e}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{P_e}(\boldsymbol{w}_y).\end{aligned} \tag{21}$$

Under the conditions of labels and environments defined by Definition.4, we can connect P_e and p_{mm} that satisfy

$$P_e = \mathbb{E}_{\boldsymbol{X}^{\mathsf{text}}}\, p_{\mathsf{mm}}(\cdot|e) \tag{22}$$

then

$$\begin{aligned}R^{(e)}(w,\Phi) &= \mathbb{E}_{\substack{\langle x^{(\mathsf{img})},y\rangle \\ \sim P_e}}\left[||\Phi(x^{(\mathsf{img})}) - w_y||\right] - H_{P_e}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{P_e}(\boldsymbol{w}_y)\\ &= \mathbb{E}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}}\left[||\Phi(x^{(\mathsf{img})}) - w_y||\right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y).\end{aligned} \tag{23}$$

Let's define $f(w,\Phi) := \Phi$, *w.r.t.* $w^{(e)}(\Phi) = \{w_y^{(e)}(\Phi) : \forall y \in \mathcal{Y}\} \in \arg\min_{\hat{w}} R^{(e)}(\hat{w}, \Phi)\ \forall e \in \mathcal{E}$ with respect to Φ, and

$$g(w,\Phi) := \Bigg\{ g : \forall y \in \mathcal{Y} = \mathcal{V}, \forall e \in \mathcal{E}, \\ \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||\Phi(x^{(\mathsf{img})}) - g(X_{y/e}^{(\mathsf{tex})})|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{X}_{y/e}^{(\mathsf{tex})}) = \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||\Phi(x^{(\mathsf{img})}) - w_y^{(e)}|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)}) \Bigg\} \tag{24}$$

According to Eq.23,

$$\begin{aligned}
\sum_{e\in\mathcal{E}} R^{(e)}(w,\Phi) =& \sum_{e\in\mathcal{E}} \left(\mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||\Phi(x^{(\mathsf{img})}) - w_y|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(f^*(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y) \right) \\
\geq& \sum_{e\in\mathcal{E}} \left(\mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||\Phi(x^{(\mathsf{img})}) - w_y^{(e)}|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(f(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)}) \right) \\
=& \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y, e\rangle \\ \sim p_{\mathsf{mm}}}} \left[||\Phi(x^{(\mathsf{img})}) - g(X_{y/e}^{(\mathsf{tex})})|| \right] - H_{p_{\mathsf{mm}}}(f(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}}(\boldsymbol{X}_{V/\mathcal{E}}^{(\mathsf{tex})}) \\
=& \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y, e\rangle \\ \sim p_{\mathsf{mm}}}} \left[||f(x^{(\mathsf{img})}) - g(X_{y/e}^{(\mathsf{tex})})|| \right] - H_{p_{\mathsf{mm}}}(f(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}}(\boldsymbol{X}_{V/\mathcal{E}}^{(\mathsf{tex})}) \\
=& \mathcal{L}_{\mathsf{SMMAlign}}^{(\mathsf{img},\mathsf{tex})}(f,g;V,\mathcal{E}); \\
R^{(e)}(w^{(e)}(\Phi),\Phi) =& \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||\Phi(x^{(\mathsf{img})}) - w_y^{(e)}(\Phi)|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)}(\Phi)) \\
=& \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}, y\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||\Phi(x^{(\mathsf{img})}) - g(X_{y/e}^{(\mathsf{tex})})|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{X}_{y/e}^{(\mathsf{tex})}) \\
=& \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||\Phi(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})})|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{X}^{(\mathsf{tex})}) \quad \text{(Definition.9)} \\
=& \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}\rangle \\ \sim p_{\mathsf{mm}}(\cdot|e)}} \left[||f(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})})|| \right] - H_{p_{\mathsf{mm}}(\cdot|e)}(f(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{X}^{(\mathsf{tex})}); \\
\sum_{e\in\mathcal{E}} R^{(e)}(w^{(e)}(\Phi),\Phi) =& \mathbb{E}_{\substack{\langle x^{(\mathsf{img})}, X^{(\mathsf{tex})}\rangle \\ \sim p_{\mathsf{mm}}}} \left[||f(x^{(\mathsf{img})}) - g(X^{(\mathsf{tex})})|| \right] - H_{p_{\mathsf{mm}}}(f(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}}(\boldsymbol{X}^{(\mathsf{tex})}) \\
=& \mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f,g).
\end{aligned} \tag{25}$$

It is noteworthy that $\sum_{e\in\mathcal{E}} R^{(e)}(w,\Phi) \geq \mathcal{L}_{\mathsf{SMMAlign}}^{(\mathsf{img},\mathsf{tex})}(f(w,\Phi), g(w,\Phi); V, \mathcal{E})$ so that

$$\begin{aligned}
&\min_{w,\Phi} \sum_{e\in\mathcal{E}} R^{(e)}(w,\Phi) \ \ \mathsf{s.t.}\ w \in \arg\min_{\hat{w}} R^{(e)}(\hat{w},\Phi),\ \forall e \in \mathcal{E} \\
\to &\min_{w,\Phi} \mathcal{L}_{\mathsf{SMMAlign}}^{(\mathsf{img},\mathsf{tex})}(f(w,\Phi), g(w,\Phi); V, \mathcal{E}) \ \ \mathsf{s.t.}\ w \in \arg\min_{\hat{w}} R^{(e)}(\hat{w},\Phi),\ \forall e \in \mathcal{E} \\
\iff &\min_{w,\Phi} \mathcal{L}_{\mathsf{SMMAlign}}^{(\mathsf{img},\mathsf{tex})}(f(w,\Phi), g(w,\Phi); V, \mathcal{E}) \ \ \mathsf{s.t.}\ \{w^{(e)}\}_{\forall e\in\mathcal{E}} \in \arg\min_{\{\hat{w}^{(e)}\}_{\forall e\in\mathcal{E}}} \sum_{e\in\mathcal{E}} R^{(e)}(w^{(e)}(\Phi),\Phi) \\
\iff &\min_{w,\Phi} \mathcal{L}_{\mathsf{SMMAlign}}^{(\mathsf{img},\mathsf{tex})}(f(w,\Phi), g(w,\Phi); V, \mathcal{E}) \ \ \mathsf{s.t.}\ w \in \arg\min_{w} \mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(f(w,\Phi), g(w,\Phi)) \\
\iff &\min_{f,g} \mathcal{L}_{\mathsf{SMMAlign}}^{(\mathsf{img},\mathsf{tex})}(f,g;V,\mathcal{E}) \ \ \mathsf{s.t.}\ f,g \in \arg\min_{\hat{f},\hat{g}} \mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(\hat{f},\hat{g})
\end{aligned} \tag{26}$$

B.1.2. The sufficient condition: Eq.10 → IRM.

Given

$$\min_{f,g} \mathcal{L}_{\mathsf{SMMAlign}}^{(\mathsf{img},\mathsf{tex})}(f,g;V,\mathcal{E}),\ \mathsf{s.t.}\ f,g \in \min_{\hat{f},\hat{g}} \mathcal{L}_{\mathsf{MMAlign}}^{(\mathsf{img},\mathsf{tex})}(\hat{f},\hat{g}) \tag{27}$$

where

$$\begin{aligned}\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}(f,g;V,\mathcal{E}) &:= \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y,e\rangle\\ \sim p_{\mathsf{mm}}}}\left[||f(x^{(\mathsf{img})})-g(X^{(\mathsf{tex})}_{y/e})||\right] - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{X}^{(\mathsf{tex})}_{V/\mathcal{E}}));\\ \mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}} &:= \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})}\rangle\\ \sim p_{\mathsf{mm}}}}\left[||f(x^{(\mathsf{img})})-g(X^{(\mathsf{tex})})||\right] - H(f(\boldsymbol{x}^{(\mathsf{img})})) - H(g(\boldsymbol{X}^{(\mathsf{tex})})),\end{aligned} \tag{28}$$

then we prove it leading to IRM (Eq.3) with respect to Theorem.2.

Specifically, we consider $\Phi(f,g) := f$ with respect to (f,g) defined by Theorem.2, and

$$\begin{aligned}w^{(e)}(f,g) := \Big\{ & w_y^{(e)} : \forall y \in \mathcal{Y} = \mathcal{V},\\ & \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y\rangle\\ \sim p_{\mathsf{mm}}(\cdot|e)}}\left[||\Phi(x^{(\mathsf{img})})-g(X^{(\mathsf{tex})}_{y/e})||\right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{X}^{(\mathsf{tex})}_{y/e}) = \mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y\rangle\\ \sim p_{\mathsf{mm}}(\cdot|e)}}\left[||\Phi(x^{(\mathsf{img})}) - w_y^{(e)}||\right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)})\Big\}, \forall e \in \mathcal{E}.\end{aligned} \tag{29}$$

Observe that

$$\begin{aligned}&\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f,g;V,\mathcal{E}) = \sum_{e\in\mathcal{E}} \underbrace{\left(\mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y\rangle\\ \sim p_{\mathsf{mm}}(\cdot|e)}}\left[||\Phi(x^{(\mathsf{img})}) - w_y^{(e)}||\right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)})\right)}_{\geq 0}\\ \iff &\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f,g;V,\mathcal{E}) = \sum_{e\in\mathcal{E}} \underbrace{\left(\mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},y\rangle\\ \sim P(\cdot|e)}}\left[||\Phi(x^{(\mathsf{img})}) - w_y^{(e)}||\right] - H_{p_{\mathsf{mm}}(\cdot|e)}(\Phi(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)})\right)}_{\geq 0}\\ \iff &\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f,g;V,\mathcal{E}) = \sum_{e\in\mathcal{E}} \underbrace{\mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},y\rangle\\ \sim P_e}}\left(-\log\frac{\exp\left((w_y^{(e)})^\top\Phi(x^{(\mathsf{img})})/\gamma\right)}{\sum_{y'\in\mathcal{Y}}\exp\left((w_{y'}^{(e)})^\top\Phi(x^{(\mathsf{img})})/\gamma\right)}\right)}_{R^{(e)}(w,\Phi)}.\end{aligned} \tag{30}$$

From Theorem.2, we know that there is $(f^*,g^*) = \arg\min_{f,g}\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f,g;V,\mathcal{E})$, such that (f^*,g^*) satisfy $\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{MMAlign}}(f^*,g^*;V,\mathcal{E}) \to 0$. With such regards, we set $(w^{(e)})^* = w^{(e)}(f^*,g^*)$ and $\Phi^* = f^*$, and it obviously leads to

$$R^{(e)}((w^{(e)})^*,\Phi^*) \to 0,\ \forall e \in \mathcal{E}, \tag{31}$$

which implies that

$$(w^{(e)})^* \in \arg\min_{\hat{w}} R^{(e)}(\hat{w},\Phi^*),\ \forall e \in \mathcal{E}$$

Then we return to the IRM constraint, with $w^* \in R^{(e)}(\hat{w},\Phi^*), \forall e \in \mathcal{E}$, thus, $w^* \in \cap_{\forall e\in\mathcal{E}}\{w^{(e)} \in R^{(e)}(\hat{w},\Phi^*)\}$. Given this

$$\begin{aligned}&\mathcal{L}^{(\mathsf{img},\mathsf{tex})}_{\mathsf{SMMAlign}}(f^*,g^*;V,\mathcal{E}),\\ =&\sum_{e\in\mathcal{E}}\left(\mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y\rangle\\ \sim p_{\mathsf{mm}}(\cdot|e)}}\left[||\Phi^*(x^{(\mathsf{img})}) - (w_y^{(e)})^*||\right] - H_{p_{\mathsf{mm}}(\cdot|e)}(f(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)})\right)\\ \overset{Thereom.2}{\to}&\sum_{e\in\mathcal{E}}\left(\mathop{\mathbb{E}}_{\substack{\langle x^{(\mathsf{img})},X^{(\mathsf{tex})},y\rangle\\ \sim p_{\mathsf{mm}}(\cdot|e)}}\left[||\Phi^*(x^{(\mathsf{img})}) - w_y^*||\right] - H_{p_{\mathsf{mm}}(\cdot|e)}(f(\boldsymbol{x}^{(\mathsf{img})})) - H_{p_{\mathsf{mm}}(\cdot|e)}(\boldsymbol{w}_y^{(e)})\right)\\ &\qquad\qquad (w^* \in \cap_{\forall e\in\mathcal{E}}\{w^{(e)} \in R^{(e)}(\hat{w},\Phi^*)\})\\ =&\sum_{e\in\mathcal{E}} R^{(e)}(w^*,\Phi^*) \to 0.\end{aligned} \tag{32}$$

To this, we can derive the optimal classifier w^* and feature extractor Φ^* in IRM, based on f^*, g^* in Theorem.2.

B.2. The Evidence of V-specific Decomposibility of $\boldsymbol{z}_{\mathsf{inv}}$ (Rule 3 in Definition.4)

Notice that Rule 3 in Definition.4 is particularly crucial to achieve the CLIP-IRM connection, whereas it is less apparent than Rule 1,2 in their observations. Unlike Rules 1–2, which are satisfied by prompt design (class vocabulary equals label space; context tokens exclude class names), Rule 3 asserts a representational property: the invariant latent $\boldsymbol{z}_{\mathsf{inv}}$ must admit a vocabulary-specific decomposition into $\boldsymbol{z}^{(\mathsf{cas})}$ (class-causal) and $\boldsymbol{z}^{(\mathsf{env})}$ (environment/context), so that class tokens align the former while environment tokens are excluded. This decomposition is what allows the constrained InfoNCE program to

become equivalent to IRM—hence its centrality. But because it concerns the internal geometry of the learned representation rather than token sets, it cannot be verified by prompt engineering alone.

The probing evidence from [94] lends concrete support for Rule 3's plausibility. Channel-Importance (CI) analysis shows that under standard prompt tuning (e.g., CoOp), most feature channels become biased toward base-specific knowledge: CI(Base) dominates CI(New) across the majority of channels, indicating collapse of task-shared, environment-invariant structure. In contrast, jointly tuning on base and new tasks produces much more consistent CI distributions across channels and markedly better zero-shot performance, implying that a substantial subset of channels (refer to $\boldsymbol{z}^{(\mathsf{env})}$ and $\boldsymbol{z}^{(\mathsf{cas})}$, respectively) can indeed carry class-causal information that is stable across environments when training dynamics preserve it.

Taken together, the channel probing in [94] provides precisely the kind of empirical evidence Rule 3 requires: CLIP's representation admits a channel-wise factorization compatible with the assumed decomposition, provided the training objective avoids collapsing shared channels into environment-specific usage. With such a decomposition in place, the vocabulary-constrained InfoNCE can target $\boldsymbol{z}^{(\mathsf{cas})}$ via class tokens while pruning $\boldsymbol{z}^{(\mathsf{env})}$ via environment-token constraints, thereby realizing IRM's invariance criterion during mid-training. This explains both the theoretical necessity of Rule 3 and its practical impact on OOD generalization observed by [94].

C. Implementation

In this section, we provide the implementation details to our mid-training paradigm and CLIP-IRM process reward modeling with the relevant policy optimization. Due to the privacy requirement of our endorsement to our project, the selected vocabulary and code will be released.

C.1. Vocabulary-Constraint Mid-training

Grounded in the causal equivalence between CLIP and IRM (Theorem 5), we realize an *invariance-injecting* mid-training regime that preserves CLIP's two-tower architecture and InfoNCE objective while restructuring supervision and batches to enforce token-level causal alignment. The key idea is to keep the standard contrastive stream over raw image–text pairs while adding a *vocabulary-constrained* stream that aligns images with environment-pruned, class-anchored text sequences. Under the token-aware SCM and Definition 4 (class-set consistency and class-agnostic context), this realizes the constrained alignment in Eqs. 9-10 and yields an objective equivalent to IRM (Eq. (3)). It leads to our methodology summarized by Eq.11, which transforms the notorious learning paradigm of IRM into the objective without bi-level optimization, thanks to the text encoders that simultaneously play the role of different classifiers by altering the vocabulary classes included in its prompt-based probing scheme (Eq.2).

Vocabulary construction. For the implementation of Eq.11, we curate two vocabularies: a class vocabulary $\mathcal{V}$ containing phrase-level class names aligned with downstream category spaces, which we define by all object categories extracted from the texts in LAION-400M (Only a single object category defined in each image-text pair, and we swap the texts by prompting them with LLaMa 3.0 [2]). An environment vocabulary $\mathcal{E}$ defined by the meta information identified in LAION-400M's texts via a similar prompt, in order to capture their backgrounds, domains, styles, and other class-agnostic context (e.g., *photo, cartoon, sketch, art painting, real, clipart, painting, infograph, quickdraw*, as well as phrases like *in the forest*, *oil painting*, *studio lighting*). We ensure $\mathcal{V} \cap \mathcal{E} = \varnothing$ and maintain phrase granularity to respect the token-level SCM (Assumption 1). This vocabulary construction enforces the $\mathcal{V}$-specific decomposability of $z_{\text{inv}} = (z_{\text{env}}, z_{\text{cls}})$ required by Definition 4, so that class phrases probe the causal subspace while environment phrases are excluded.

Object-Vocabulary-Construction Prompt:
You are given an image caption. Extract the single main object category mentioned, as a short noun or noun phrase (e.g., "dog", "fire truck", "coffee mug"). Follow these rules:
- Return exactly one category;
- If multiple objects appear, pick the most salient/main one;
- Do not include scene/style/context words (e.g., "photo", "painting", "in the forest", "studio lighting");
- Do not include attributes (colors, sizes, counts);
- If no clear object is present, return: unknown.

Caption: "⟨CAPTION_TEXT⟩"
Answer (one short noun phrase only):

[2] www.llama.com/models/llama-3/

Algorithm 1: Vocabulary Construction via LLaMA 3.0 Object/Background Screening

Input: LAION captions $\{\mathbf{x}^{(\mathrm{tex})}\}$; CLIP tokenizer; LLaMA 3.0 object prompt $\mathcal{P}_{\mathrm{obj}}$; LLaMA 3.0 background prompt $\mathcal{P}_{\mathrm{bg}}$; benchmark category spaces $\mathcal{Y}$; benchmark meta/domain labels (PACS, VLCS, OfficeHome, NICO++, DomainNet)
Output: Class vocabulary $\mathcal{V}$, environment vocabulary $\mathcal{E}$ with $\mathcal{V} \cap \mathcal{E} = \varnothing$

Initialize empty multisets $\mathcal{B}_{\mathrm{obj}} \leftarrow \varnothing$, $\mathcal{B}_{\mathrm{bg}} \leftarrow \varnothing$.
Seed environment list $\mathcal{E}_{\mathrm{seed}}$ from benchmark metas: {*photo, cartoon, sketch, art painting, real, clipart, painting, infograph, quickdraw*, phrases like *in the forest*, *oil painting*, *studio lighting*}.
for each caption $\mathbf{x}^{(\mathrm{tex})}$ **do**
    Query LLaMA 3.0 with $\mathcal{P}_{\mathrm{obj}}$ on $\mathbf{x}^{(\mathrm{tex})}$ to obtain object candidate $\hat{y}$ (single noun phrase or "unknown").
    Query LLaMA 3.0 with $\mathcal{P}_{\mathrm{bg}}$ on $\mathbf{x}^{(\mathrm{tex})}$ to obtain background candidate $\hat{e}$ (single noun phrase or "unknown").
    if $\hat{y} \neq$ "unknown" **then**
        Normalize $\hat{y}$: lowercase, trim, collapse whitespace; merge hyphenation/variants; keep multi-token phrase granularity (e.g., "fire truck").
        Append $\hat{y}$ to $\mathcal{B}_{\mathrm{obj}}$.
    end if
    if $\hat{e} \neq$ "unknown" **then**
        Normalize $\hat{e}$ as above; append to $\mathcal{B}_{\mathrm{bg}}$.
    end if
end for
Optional filtering by frequency: keep object phrases with count $\geq \tau_{\mathrm{obj}}$ in $\mathcal{B}_{\mathrm{obj}}$; keep background phrases with count $\geq \tau_{\mathrm{bg}}$ in $\mathcal{B}_{\mathrm{bg}}$.
Align objects to downstream categories: compute embedding similarity of each candidate in $\mathcal{B}_{\mathrm{obj}}$ to names/synonyms in $\mathcal{Y}$; retain candidates with similarity $\geq \delta_{\mathrm{obj}}$; map near-duplicates to canonical forms in $\mathcal{Y}$ when possible.
Initialize class vocabulary $\mathcal{V} \leftarrow$ unique phrases from filtered $\mathcal{B}_{\mathrm{obj}}$ (post-alignment).
Initialize environment vocabulary $\mathcal{E} \leftarrow$ unique phrases from filtered $\mathcal{B}_{\mathrm{bg}} \cup \mathcal{E}_{\mathrm{seed}}$.
Enforce disjointness $\mathcal{V} \cap \mathcal{E} = \varnothing$:
    For any phrase $p \in \mathcal{V} \cap \mathcal{E}$:
        Compute class-likeness score $s_{\mathrm{cls}}(p)$ (similarity to $\mathcal{Y}$, occurrence as object vs. background in LLaMA outputs).
        Compute environment-likeness score $s_{\mathrm{env}}(p)$ (similarity to $\mathcal{E}_{\mathrm{seed}}$, occurrence ratio as background).
        if $s_{\mathrm{cls}}(p) > s_{\mathrm{env}}(p) + m$ **then** keep p in $\mathcal{V}$ and remove from $\mathcal{E}$; **else** keep in $\mathcal{E}$ and remove from $\mathcal{V}$; break ties by dropping p.
Prune residual leakage:
    Remove from $\mathcal{V}$ any phrase matching known environment/style patterns (e.g., "in the *", "* painting", "* lighting", weather/season terms).
    Remove from $\mathcal{E}$ any phrase that is a concrete object hypernym/hyponym (e.g., "animal", "dog", "car") by WordNet/embedding heuristics.
Return $(\mathcal{V}, \mathcal{E})$.

Environment-Vocabulary-Construction Prompt:
You are given an image caption. Extract the single best background/environment noun or noun phrase that describes the scene, setting, domain, or style (e.g., "forest", "kitchen", "studio lighting", "oil painting", "cartoon", "clipart", "snowy street"). Follow these rules:
- Return exactly one background/environment term;
- Prioritize scene/place, medium/style, domain, weather/lighting, or general context;
- Exclude the main object (e.g., "dog", "person", "car") and its attributes;
- Exclude actions/verbs; use a noun or noun phrase;
- If no clear background term is present, return: unknown.

Caption: "⟨CAPTION_TEXT⟩"
Answer (one short noun phrase only):

With $\mathcal{V}$ and $\mathcal{E}$ in place, we form two batch streams per step. The first, $\mathcal{D}^{(K)}$, is a standard CLIP batch of K image–caption

pairs sampled from LAION, used to preserve coverage and stability. The second, $\mathcal{D}_{\mathcal{V}}^{(K)}$, is built by pruning each caption of environment phrases and keeping a single high-confidence class phrase $y \in \mathcal{V}$; if pruning yields degenerate text, we synthesize a minimal class-only prompt (e.g., *"a photo of a* $\{y\}$*"*). Optionally, we swap captions among images that share the same y and contain no environment tokens, creating environment-invariant pairs. We then compute within-batch contrastive negatives as in CLIP, but over the constrained sequences.

Training. The sum of InfoNCE (both directions) is built on the constrained stream $\mathcal{D}_{\mathcal{V}}^{(K)}$, plus a weighted ($\lambda = 1$) InfoNCE on the raw stream $\mathcal{D}^{(K)}$. The temperature γ (logit scale) remains trainable with a smaller learning rate. We initialize from pre-trained CLIP and train both encoders end-to-end without architectural changes. Following Sec. 7.1, we instantiate two variants: CLIP-IRMv1 (ViT-B/16) for comparisons with OOD baselines, and CLIP-IRMv2 (ViT-L/14) for comparisons with stronger foundation models. Optimization uses AdamW with CLIP-like weight decay, warmup (3–5%), cosine decay, gradient clipping, and mixed precision. We often set the text-encoder learning rate slightly lower than the image-encoder's to stabilize token-sensitive alignment; the logit-scale uses a smaller LR. The balance $\lambda \in [0.5, 0.7]$ reliably preserves general coverage while letting the invariant signal dominate sufficiently to realize OOD gains across PACS, VLCS, OfficeHome, NICO++, and DomainNet.

We implement phrase detection / removal at tokenizer-index level, combining exact / regex matches with embedding-thresholded fuzzy matches to reduce false positives. When captions lack a reliable class phrase, we back off to a minimal class-only template from $\mathcal{V}$. During validation, we always use standard zero-shot prompts (e.g., *"a photo of a [CLASS]"*) to ensure that observed gains reflect invariant feature learning rather than prompt overfitting. In few-shot prompt-tuning (CoOp, CoCoOp, VPT, MaPLe, PromptSRC), simply swapping CLIP with CLIP-IRM systematically shifts the accuracy frontier in Base-to-New and cross-dataset transfers, indicating that mid-training disentangles class and environment factors in a way downstream adapters can exploit.

C.2. CLIP-IRM as Process-Reward Guidance for Multimodal OOD Reasoning

C.2.1. Math-centric Mid-training (CLIP-IRM for Math Vision)

Our key idea is to teplace the LAION-oriented mid-training with a math-centric mid-training on Geometry3K and MMK12. Build math-specific class/environment vocabularies and apply token-level pruning to achieve invariant, diagram-grounded alignment via constrained InfoNCE.

Concretely, we construct a mid-training corpus from Geometry3K and MMK12, where images are math diagrams (geometry figures, graphs, charts) and texts include problem statements, rationales, and answers. This domain shift requires vocabularies and pruning rules that reflect *math-causal* versus *style/environment* factors.

Vocabulary design: class set V_{math} and environment set E_{math}. We employ the prompt by replacing the target objects and environments based on the list below,

- *Causal classes:* The class vocabulary V_{math} should contain phrase-level entities that are causally predictive for solving visual math:

$$\begin{aligned} V_{\text{math}} \supset \{&\text{point, line, segment, ray, angle, triangle, right triangle, isosceles triangle,} \\ &\text{square, rectangle, circle, arc, chord, tangent, radius, diameter,} \\ &\text{polygon, vertex, side, base, height, hypotenuse,} \\ &\text{grid, x-axis, y-axis, tick marks, origin, coordinates,} \\ &\text{parabola, function graph, bar chart, pie chart,} \\ &\text{length, area, perimeter, degree, radian, ratio, fraction,} \\ &\text{congruent, similar, } \perp, \parallel, \cong, \sim, \angle, \pi, {}^\circ\}. \end{aligned}$$

- *Environment factors:* The environment vocabulary E_{math} should contain rendering, layout, and dataset boilerplate that are not causal for the visual reasoning:

$$\begin{aligned} E_{\text{math}} \supset \{&\text{hand-drawn, scanned, photocopy, chalkboard, textbook illustration,} \\ &\text{low-resolution, watermark, page margin, header, footer,} \\ &\text{page number, problem index, box border, difficulty,} \\ &\text{choose one answer, multiple choice, time limit, URL}\}. \end{aligned}$$

We ensure $V_{\text{math}} \cap E_{\text{math}} = \varnothing$ by scoring phrases with a class-likeness versus environment-likeness criterion (embedding similarity to seed lists, usage statistics in Geometry3K/MMK12), keeping phrase granularity.

Algorithm 2: Mid-training Pipeline (Eq. (11)) for CLIP-IRM with Vocabulary-constrained Alignment

Input: Pre-trained CLIP encoders $f(\cdot)$ (image), $g(\cdot)$ (text); LAION pairs $\mathcal{D}$; vocabularies $(\mathcal{V}, \mathcal{E})$; batch size K; weight λ; temperature γ (trainable)

Output: Mid-trained CLIP-IRM encoders $f^\star, g^\star$

Initialize optimizer (AdamW), learning rates (image $\geq$ text; small LR for logit scale), warmup and cosine decay; enable mixed precision; set gradient clipping.

while training not converged **do**

  Sample a raw batch $\mathcal{D}^{(K)} = \{\langle x_i^{(\text{img})}, x_i^{(\text{tex})} \rangle\}_{i=1}^K$ from LAION.

  Build a constrained batch $\mathcal{D}_{\mathcal{V}}^{(K)}$:

  for each $\langle x^{(\text{img})}, x^{(\text{tex})} \rangle \in \mathcal{D}^{(K)}$ **do**

    Detect class phrase $y \in \mathcal{V}$ in $x^{(\text{tex})}$ (exact/regex + embedding similarity to canonical class list).

    if multiple or zero class phrases found **then**

      Use LLaMA 2 to *normalize* to a single class phrase $y \in \mathcal{V}$; if unresolved, resample another pair.

    end if

    Remove all environment phrases $e \in \mathcal{E}$ from tokenized $x^{(\text{tex})}$ to obtain $X_{y/e}^{(\text{tex})}$ (phrase-level pruning).

    if $X_{y/e}^{(\text{tex})}$ degenerate (empty/one token) **then**

      Set $X_{y/e}^{(\text{tex})} \leftarrow T(y)$, a minimal class-only prompt (e.g., "a photo of a {y}").

    end if

    (Optional) Swap environment-free captions among images with the same class y to create additional invariant pairs.

    Append $\langle x^{(\text{img})}, X_{y/e}^{(\text{tex})} \rangle$ to $\mathcal{D}_{\mathcal{V}}^{(K)}$.

  end for

  Encode images and texts:

    For both streams, compute $v_i = f(x_i^{(\text{img})})$.

    Compute $t_i^{\text{pruned}} = g(X_{y/e,i}^{(\text{tex})})$ for $\mathcal{D}_{\mathcal{V}}^{(K)}$; compute $t_i^{\text{raw}} = g(x_i^{(\text{tex})})$ for $\mathcal{D}^{(K)}$.

  Compute InfoNCE on the constrained batch (both directions), matching Eq. (1):

$$\mathcal{L}_{\text{img}\to\text{tex}}^{\mathcal{V}} = \sum_{i=1}^K -\log \frac{\exp\left(v_i^\top t_i^{\text{pruned}}/\gamma\right)}{\sum_{j=1}^K \exp\left(v_i^\top t_j^{\text{pruned}}/\gamma\right)}$$

$$\mathcal{L}_{\text{tex}\to\text{img}}^{\mathcal{V}} = \sum_{i=1}^K -\log \frac{\exp\left(v_i^\top t_i^{\text{pruned}}/\gamma\right)}{\sum_{j=1}^K \exp\left(v_j^\top t_i^{\text{pruned}}/\gamma\right)}$$

  Compute standard InfoNCE on the raw batch $\mathcal{D}^{(K)}$ (both directions), as in Eq. (1):

$$\mathcal{L}_{\text{img}\to\text{tex}}^{\text{raw}} = \sum_{i=1}^K -\log \frac{\exp\left(v_i^\top t_i^{\text{raw}}/\gamma\right)}{\sum_{j=1}^K \exp\left(v_i^\top t_j^{\text{raw}}/\gamma\right)}$$

$$\mathcal{L}_{\text{tex}\to\text{img}}^{\text{raw}} = \sum_{i=1}^K -\log \frac{\exp\left(v_i^\top t_i^{\text{raw}}/\gamma\right)}{\sum_{j=1}^K \exp\left(v_j^\top t_i^{\text{raw}}/\gamma\right)}$$

  Combine losses per Eq. (11):

$$\mathcal{L} = \left(\mathcal{L}_{\text{img}\to\text{tex}}^{\mathcal{V}} + \mathcal{L}_{\text{tex}\to\text{img}}^{\mathcal{V}}\right) + \lambda\left(\mathcal{L}_{\text{img}\to\text{tex}}^{\text{raw}} + \mathcal{L}_{\text{tex}\to\text{img}}^{\text{raw}}\right)$$

  Backpropagate $\nabla\mathcal{L}$; apply gradient clipping; update f, g, and logit scale $1/\gamma$ via AdamW.

  Periodically evaluate zero-shot OOD accuracy using standard prompts (e.g., "a photo of a [CLASS]"); adjust λ or batch ratios if OOD metrics plateau or regress.

end while

Return $f^\star, g^\star$ (CLIP-IRM weights).

Token-level pruning and class selection. For each sample $(x^{(\text{img})}, x^{(\text{tex})})$, detect a main class $y \in V_{\text{math}}$ (e.g., "triangle ABC", "x-axis") and remove all occurrences of tokens/phrases from E_{math} in the tokenized text to obtain the pruned sequence $X_{y/e_{\text{math}}}^{(\text{tex})}$. If multiple V_{math} phrases are present, prioritize the diagram-grounded one (via co-occurrence heuristics with detected

visual patches). If the pruned text becomes degenerate, synthesize a minimal prompt $T(y)$, e.g., "a diagram of a {y}".

Environment-invariant swaps. Within a minibatch, swap pruned texts among images sharing the same class y (and containing no E_{math} tokens) to create synthetic environment-invariant pairs, mirroring Sec. 5 but in the math domain.

Patch proposals for math grounding. Obtain M candidate patches via attention rollout or a lightweight proposal network; retain patches with high "mathness" (e.g., intersect with edge maps, suppress OCR-heavy paragraph regions while keeping diagram labels). These patches $\{v_m\}_{m=1}^{M}$ will support patch-grounded rewards later.

Math-centric mid-training objective. We define two streams (as in Eq. (11) of the paper), but on math data:

$$\mathcal{D}^{(K)}: \text{ raw math minibatch}, \quad \mathcal{D}^{(K)}_{V_{\text{math}}}: \text{ pruned/synthesized math minibatch},$$
$$v_i = f(x_i^{(\text{img})}), \quad t_i^{\text{pruned}} = g(X^{(\text{tex})}_{y/e_{\text{math}},i}), \quad t_i^{\text{raw}} = g(x_i^{(\text{tex})}).$$

The loss combines constrained InfoNCE on $\mathcal{D}^{(K)}_{V_{\text{math}}}$ and raw InfoNCE on $\mathcal{D}^{(K)}$:

$$\mathcal{L} = \Big(\mathcal{L}^{V}_{\text{img}\to\text{tex}} + \mathcal{L}^{V}_{\text{tex}\to\text{img}}\Big) + \lambda\Big(\mathcal{L}^{\text{raw}}_{\text{img}\to\text{tex}} + \mathcal{L}^{\text{raw}}_{\text{tex}\to\text{img}}\Big),$$

with the standard CLIP-style (temperatured) InfoNCE terms:

$$\mathcal{L}^{V}_{\text{img}\to\text{tex}} = \sum_{i=1}^{K} -\log \frac{\exp\left(\langle v_i, t_i^{\text{pruned}}\rangle/\tau\right)}{\sum_{j=1}^{K}\exp\left(\langle v_i, t_j^{\text{pruned}}\rangle/\tau\right)}, \quad \mathcal{L}^{V}_{\text{tex}\to\text{img}} = \sum_{i=1}^{K} -\log \frac{\exp\left(\langle v_i, t_i^{\text{pruned}}\rangle/\tau\right)}{\sum_{j=1}^{K}\exp\left(\langle v_j, t_i^{\text{pruned}}\rangle/\tau\right)},$$

and analogous definitions for the "raw" terms with t^{raw}. We use AdamW, warmup (3–5%), cosine decay, gradient clipping, and mixed precision. The text-encoder LR and logit-scale LR are set smaller than the image-encoder LR. We train ViT-B/16 (CLIP-IRM-math-B) and ViT-L/14 (CLIP-IRM-math-L).

Validation and sanity checks. We measure retrieval between math images and prompts "a diagram of a [CLASS]" on held-out data. Patch-level max-similarity should localize diagram elements rather than page artifacts. Improved alignment on graphs/geometry entities indicates successful math-causal invariance.

C.2.2. CLIP-IRM Process-level Reward Modeling for MLLM

After obtaining the CLIP-IRM math version, we convert math-IRM alignment into step-wise rewards that guide an MLLM policy (e.g., Qwen2.5-VL-7B-Instruct) via GRPO to produce invariant, image-grounded reasoning chains.

Coupled decoder–encoder interface. Let π_θ be the MLLM decoder and (f, g) the math mid-trained CLIP-IRM encoders. Given image $x^{(\text{img})}$ and an autoregressive token sequence $\{t_k\}$, for step k we form a window $t_{k-w+1:k}$, compute:

$$h_k^{(\text{tex})} = g\big(t_{k-w+1:k}\big), \quad v^{(\text{img})} = f\big(x^{(\text{img})}\big).$$

These drive a token-aware, math-constrained alignment score, and we set the window size as 64 in all our experiments.

Process alignment reward.

$$r_{\text{proc-align}}(k) = \underbrace{\text{InfoNCE}\big(v^{(\text{img})}, h_k^{(\text{tex})}; \mathcal{N}\big)}_{\text{alignment}} - \alpha \cdot \underbrace{\text{env\_overlap}\big(t_{k-w+1:k}; E_{\text{math}}\big)}_{\text{environment penalty}},$$

where $\mathcal{N}$ are negatives (other windows in the batch or cached class prototypes $g(T(y))$ for $y \in V_{\text{math}}$). The InfoNCE score may be implemented as a log-softmax over similarities with temperature τ, normalized to a convenient range. The env_overlap measures the presence or fraction of E_{math} phrases in the window (exact/fuzzy phrase matching). α schedules from larger to smaller (e.g., $0.3 \to 0.15$).

Encourage visual grounding: With patch embeddings $\{v_m\}_{m=1}^{M}$,

$$r_{\text{patch}}(k) = \max_{m\in\{1,\dots,M\}} \text{sim}\big(v_m, h_k^{(\text{tex})}\big), \qquad r_{\text{proc}}(k) = r_{\text{proc-align}}(k) + \beta \cdot r_{\text{patch}}(k),$$

with β annealed upward (e.g., $0.1 \to 0.3$) as patch proposals stabilize.

Anchor to main class y: If a main $y \in V_{\text{math}}$ is identified,

$$r_{\text{class}}(k) = \text{sim}\big(v^{(\text{img})}, g(T(y))\big) - \frac{1}{|V_{\text{math}}| - 1} \sum_{y' \neq y} \text{sim}\big(v^{(\text{img})}, g(T(y'))\big),$$

applied only when the window contains V_{math} phrases related to y to avoid over-shaping.

Stabilized, composite shaping signal:

$$r_{\text{proc-total}}(k) = \text{zscore}_{\text{batch}}[\, r_{\text{proc}}(k)\,] + \gamma \cdot r_{\text{class}}(k),$$

with small γ (e.g., 0.05–0.1). Batch-wise standardization mitigates scale drift across updates.

Combine task and process rewards: Let $r_{\text{task}}(k)$ include format/step rewards and a terminal correctness reward. The episodic return is:

$$R = \sum_k \Big(r_{\text{task}}(k) + \lambda_{\text{proc}}\, r_{\text{proc-total}}(k) \Big),$$

where λ_{proc} is annealed upward (e.g., $0.1 \to 0.4$) as policy stabilizes.

GRPO optimization. Here we use GRPO with a reference policy π_{ref},

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_k \text{GRPO}\big(A_k\big)\right] \; - \; \lambda_{\text{KL}}\, \text{KL}(\pi_\theta \,\|\, \pi_{\text{ref}})\,,$$

where advantages A_k are computed from the shaped return. Stabilization includes (i) batch z-scoring of r_{proc}, (ii) KL control via λ_{KL}, and (iii) initially freezing (f, g), with optional tiny-LR unfreezing later on interleaved $\mathcal{D}_{V_{\text{math}}}^{(K)}/\mathcal{D}^{(K)}$ steps to maintain alignment. We maintain a memory of recent $h^{(\text{tex})}$ windows for negatives; include prototypes $g(T(y))$ for $y \in V_{\text{math}}$. Use stride s (e.g., evaluate windows every s=8 tokens) to reduce compute. Downweight windows dominated by E_{math}. We penalize repeated V_{math} n-grams (diminishing returns), increase negative hardness if reward rises without V_{math} coverage, and apply sparsity penalties on excessive V_{math} term counts. Maintain a sufficiently strong λ_{KL}.

C.2.3. Practical Pipelines and Dataset-specific Notes

Key idea: Provide actionable pseudo-pipelines aligned with Sec. 7.2 (training on Geometry3K+MMK12; evaluating on MathVista, WeMath, MathVerse, MathVision, and HallusionBench; ID on Geometry3K).

Mid-training (CLIP-IRM-math) pipeline:

1. Sample a minibatch $\mathcal{D}^{(K)}$ of $(x^{(\text{img})}, x^{(\text{tex})})$ from Geometry3K/MMK12.
2. Build $\mathcal{D}_{V_{\text{math}}}^{(K)}$: detect $y \in V_{\text{math}}$; prune phrases from E_{math} to get $X_{y/e_{\text{math}}}^{(\text{tex})}$; if degenerate, set $X_{y/e_{\text{math}}}^{(\text{tex})} \leftarrow T(y)$; optionally swap texts across same-y images.
3. Encode $v_i = f(x_i^{(\text{img})})$, $t_i^{\text{pruned}} = g(X_{y/e_{\text{math}},i}^{(\text{tex})})$, and $t_i^{\text{raw}} = g(x_i^{(\text{tex})})$.
4. Compute $\mathcal{L}$ as defined above; update with AdamW, warmup+cosine, grad clip, mixed precision; smaller LR for text and temperature.
5. Periodically validate math retrieval and patch grounding.

GRPO with CLIP-IRM-math rewards:

1. Sample (image, prompt, GT) from Geometry3K/MMK12 (train).
2. Generate CoT with π_θ (greedy or low-temp).
3. For each output, compute $v^{(\text{img})} = f(x^{(\text{img})})$; for windows $t_{k-w+1:k}$ (stride s), compute $h_k^{(\text{tex})} = g(\cdot)$ and evaluate:

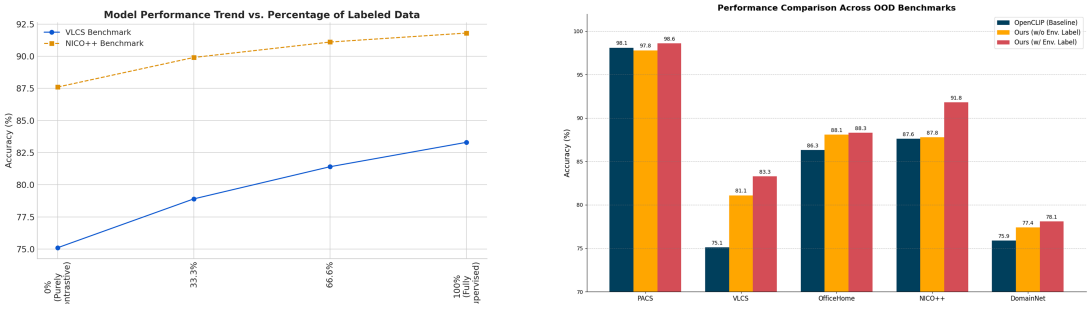
$$r_{\text{proc-align}}(k), \quad r_{\text{patch}}(k), \quad r_{\text{class}}(k), \quad r_{\text{proc-total}}(k).$$

4. Combine with $r_{\text{task}}(k)$ into $R = \sum_k \left(r_{\text{task}}(k) + \lambda_{\text{proc}} r_{\text{proc-total}}(k)\right)$; compute A_k and update π_θ with GRPO and KL regularization to π_{ref}.
5. Schedules: $\lambda_{\text{proc}} : 0.1 \to 0.4$, $\alpha : 0.3 \to 0.15$, $\beta : 0.1 \to 0.3$; optionally unfreeze (f, g) with tiny LR late in training on interleaved math batches.

Preprocessing and Evaluation Details. We run OCR to separate diagram labels (A,B,C, angle marks) from paragraph text; treat labels as potentially part of V_{math} for grounding; remove dataset boilerplate and non-causal metadata (in E_{math}). We train the policy model on Geometry3K + MMK12, then evaluate its OOD reasoning on MathVista, WeMath, MathVerse, MathVision; hallucination robustness on HallusionBench; ID on Geometry3K test. Use greedy decoding and an external judge to parse answers. Expect OOD gains via environment pruning and patch grounding, reduced hallucination, and maintained or improved ID accuracy with better sample efficiency.

D. More Experiments

Further analysis on OOD generalization



(a) (b)

Figure 5. (a) The comparisons across different proportions of the newly introduced category and domain information; (b)The analysis of environment knowledge.

We provide the experiment of analyzing the importance of category-label and domain-label information for debaise. We vary the proportion of $\mathcal{D}_{\mathcal{V}}$ during pre-training from 0%, 33.3%, 66.6%, 100%, then we observe the pre-trained models' performances on VLCS and NICO++. The experimental results in Fig.5(a) provides a strong empirical validation for the proposed improved techniques. On both the VLCS and NICO++ benchmarks, model accuracy consistently increases with the percentage of labeled data used during pre-training. Performance on VLCS climbs from 75.1% to 83.3%, and on NICO++ from 87.6% to 91.8%. This directly confirms that incorporating the vocabulary-supervised, IRM-like objective effectively enhances out-of-distribution (OOD) generalization, as predicted by Theorem 10. Notably, the most significant performance gains occur early (from 0% to 33.3% labeled data), highlighting the efficiency of a hybrid training approach that blends self-supervision with targeted, invariance-promoting signals. The trend demonstrates that even partial supervision yields substantial robustness improvements.

Beyond this, we also provide a more thorough analysis caused by the environment knowledge impact in the experiments. Specifically, we re-trained the model that only consider the class-name information but do not shuttle the image-text pairs with the same class name across different environments. It leads to the comparison among, OpenCLIP (baseline), w/o environment label, and w environment label (ours). The chart in Fig.5(b) reveals how the effectiveness of the proposed techniques varies depending on the nature of the distribution shift in each benchmark.

- PACS: This benchmark shows very high accuracy across all models, with scores of 98.1%, 97.8%, and 98.6%. The baseline OpenCLIP is already extremely strong. Interestingly, our model without environment labels sees a minor dip in performance. However, with the addition of environment labels, our model achieves the highest score, suggesting that while the task is nearly saturated, the explicit environment signal helps fine-tune the model to achieve state-of-the-art performance.
- VLCS: This is where our proposed method shows the most dramatic improvement. The OpenCLIP baseline struggles with a 75.1% accuracy. Our model without environment labels achieves a massive +6 point gain to 81.1%. Adding the environment label provides a further +2.2 point boost to 83.3%. This strongly indicates that the spurious correlations present in VLCS (e.g., object style, background) are precisely the weakness that our invariance-focused training method is designed to correct.

Method	Source	Avg	Caltech101	OxfordPets	StanfCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101
CoOp	71.80	64.40	93.97	89.60	64.60	69.13	85.47	20.70	65.70	43.07	44.50	67.23
+CLIP-IRM	71.43	65.89	93.30	90.00	65.53	70.50	85.97	21.90	66.07	43.17	44.97	68.80
CoCoOp	71.17	65.73	94.30	90.80	65.53	71.80	86.13	22.83	67.73	45.57	43.47	69.10
+CLIP-IRM	72.77	66.05	94.10	90.63	66.23	72.17	86.27	22.90	67.30	45.50	44.17	69.53
MaPLe	72.47	64.17	92.97	90.20	63.97	70.03	84.83	23.23	66.00	43.23	40.03	67.23
+CLIP-IRM	72.82	64.37	92.53	90.10	64.60	70.10	85.57	23.63	66.40	45.03	40.13	67.53
VPT	70.80	62.61	91.67	90.03	62.47	66.03	81.70	24.07	65.27	44.27	35.77	64.83
+CLIP-IRM	71.97	63.32	91.30	90.03	62.63	66.77	83.03	23.73	65.57	44.57	37.03	65.40
PromptSRC	71.33	65.71	93.77	90.40	65.77	70.80	86.30	23.67	66.93	46.07	44.23	69.20
+CLIP-IRM	70.90	66.75	93.80	90.13	66.00	70.93	86.27	24.30	67.23	46.60	45.83	69.10

Table 2. Cross-dataset generalization performance of six baselines with or without our CLIP-IRM on 11 datasets. Source indicates ImageNet.

- OfficeHome: The trend on OfficeHome shows a clear, step-wise improvement. The baseline score of 86.3% is improved to 88.1% by our method without environment labels. The model with environment labels inches slightly higher to 88.3%. The small 0.2 point gap between our two models suggests that for this benchmark, the act of "scrubbing" the text of environmental cues captures nearly all the potential gains, and the explicit environment label provides only a marginal additional benefit.
- NICO++: This benchmark presents a unique and insightful result. The baseline OpenCLIP (87.6%) and our model without environment labels (87.8%) perform almost identically. However, our model with environment labels achieves a significant +4 point jump to 91.8%. This implies that for the types of distribution shifts in NICO++, simply removing environmental context from text is insufficient. The model requires the explicit signal from the environment label during training to learn the correct invariances and disentangle causal features from spurious ones.
- DomainNet: Performance on DomainNet follows a similar pattern to OfficeHome, but with lower overall scores, indicating it is a more challenging benchmark. There is a consistent improvement from the 75.9% baseline to 77.4% (without environment labels) and finally to 78.1% (with environment labels). These steady gains confirm the utility of our method, while the modest scale of improvement suggests the complexity of the domain shifts in DomainNet remains a significant challenge.

The complete experiments on prompt-tuning generalization. We have presented the complete evaluation results of base-2-new generalization (Table.3) and cross-dataset transfer (Table.2).

In Table.2, we observe that replacing CLIP with CLIP-IRM consistently lifts the target-domain accuracy across all five prompt-tuning families (CoOp, CoCoOp, MaPLe, VPT, PromptSRC), while keeping or slightly improving the ImageNet source accuracy. Gains are largest on texture/style/scene-sensitive targets such as DTD, EuroSAT, and SUN397, which aligns with our claim that mid-training prunes environment factors: for example, PromptSRC sees +1.04 Avg with noticeable rises on DTD (+0.53) and EuroSAT (+1.60), and VPT enjoys improvements on DTD (+0.30) and EuroSAT (+1.26) despite minimal source changes. The harder FGVCAircraft classifies fine-grained shapes under background/style variance; here CLIP-IRM yields the most striking jumps (e.g., CoOp: +7.08 on target, CoCoOp: +13.55 on target) but can trade a bit of source accuracy in a few cases, indicating that shape-centric invariances help more on the shifted target distributions. Overall, the systematic rightward (higher target) shift with small or neutral movement in source demonstrates that CLIP-IRM improves robustness to cross-dataset shifts without relying on source memorization, and that its benefits are architecture-agnostic—spanning text-only (CoOp/CoCoOp), visual (VPT), and multimodal prompt-tuning (MaPLe/PromptSRC).

In Table.3, CLIP-IRM systematically increases the New-class accuracy and the harmonic mean H across all baselines, with Base often unchanged or slightly improved—an important signal that mid-training recovers causal, class-relevant features rather than overfitting base classes. The biggest New/H gains appear where spurious style cues are known to dominate: DTD, EuroSAT, and SUN397 (e.g., PromptSRC+CLIP-IRM: New +1.50 on DTD, +0.50 on EuroSAT, +1.50 on SUN397; H improves accordingly), and on fine-grained FGVCAircraft where causal shape signals matter (e.g., MaPLe+CLIP-IRM: H +3.42; CoCoOp+CLIP-IRM: H +2.15). New-class improvements coexist with stable Base accuracy on ImageNet and Caltech101, reinforcing that the invariance signal does not erode in-distribution competence. Notably, the consistent H improvements across five distinct post-tuning paradigms confirm that CLIP-IRM serves as a stronger, more robust substrate for downstream prompt learning, pushing the Pareto frontier toward better performance on previously unseen classes without sacrificing the base domain.

Further analysis on OOD reasoning. We provide some ablation studies to analyze the OOD reasoning results achieved by GRPO trained with the CLIP-IRM process reward.

Method	Avg over 11 datasets			ImageNet			Caltech101			OxfordPets		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	81.50	69.77	75.18	76.57	69.97	73.12	98.17	94.83	96.47	95.57	97.53	96.54
+CLIP-IRM	82.26	70.42	75.82	77.13	70.10	73.45	98.33	94.33	96.29	94.70	97.63	96.14
CoCoOp	81.18	72.18	76.40	77.10	72.34	74.65	98.20	93.20	95.40	94.93	97.90	96.39
+CLIP-IRM	83.44	73.81	78.33	77.27	73.00	75.07	98.37	93.87	96.06	94.03	97.20	95.59
MaPLe	80.93	73.88	77.30	77.05	73.24	75.10	97.87	94.03	95.91	95.47	97.80	96.62
+CLIP-IRM	84.61	74.33	79.15	77.29	73.42	75.28	98.30	94.60	96.41	94.33	97.23	95.76
VPT	84.85	74.82	79.43	77.77	70.23	73.84	98.00	93.70	95.84	95.17	97.77	96.45
+CLIP-IRM	84.28	74.49	79.10	77.80	70.77	74.12	98.10	93.75	95.90	95.03	97.83	96.41
PromptSRC	84.21	75.75	79.72	77.80	73.90	75.80	98.10	93.87	95.94	95.27	97.23	96.24
+CLIP-IRM	84.05	74.88	79.23	78.20	73.70	75.88	98.57	94.10	96.28	95.43	97.33	96.37

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	74.30	72.10	73.18	97.07	74.33	84.19	90.43	90.97	90.70	31.70	17.30	22.38
+CLIP-IRM	79.67	72.40	75.80	98.90	83.08	90.43	91.43	93.23	90.88	42.53	22.53	29.46
CoCoOp	70.77	72.30	71.26	90.17	69.27	78.39	90.57	91.20	90.88	35.63	22.70	34.10
+CLIP-IRM	79.87	73.33	76.47	98.97	83.51	90.30	91.30	91.30	90.80	43.07	31.30	36.25
MaPLe	71.13	71.27	71.20	95.90	79.72	87.04	90.53	91.30	90.91	35.10	35.20	35.15
+CLIP-IRM	79.13	75.47	77.27	98.90	85.84	90.50	91.61	91.05	91.05	43.20	34.83	38.57
VPT	76.30	72.33	74.26	92.03	72.87	81.34	90.90	91.53	90.91	40.57	36.47	38.31
+CLIP-IRM	80.93	71.73	76.00	98.93	83.79	90.33	91.53	91.53	90.93	44.53	32.80	37.78
PromptSRC	82.13	72.17	76.23	98.17	82.71	89.80	90.87	91.20	91.03	35.90	30.37	32.90
+CLIP-IRM	80.80	75.00	77.79	98.40	83.86	90.27	91.03	90.65	90.65	45.30	31.87	37.41

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	81.13	76.07	78.52	79.33	49.70	61.11	89.35	57.30	69.82	83.87	69.80	76.19
+CLIP-IRM	83.44	73.81	78.57	83.23	60.53	69.60	90.07	66.27	75.80	85.43	72.17	78.24
CoCoOp	80.73	75.03	77.78	79.20	52.32	62.83	87.97	63.63	73.85	82.33	72.40	77.05
+CLIP-IRM	82.20	76.13	79.03	82.73	59.27	68.66	87.62	65.70	76.82	85.70	72.80	78.73
MaPLe	81.90	77.33	79.55	82.37	56.58	66.53	85.37	63.40	72.91	83.73	75.40	79.53
+CLIP-IRM	82.33	77.00	79.57	82.17	58.73	68.78	90.03	71.07	79.04	85.80	72.23	81.29
VPT	82.90	76.40	79.53	83.20	59.47	69.04	84.97	66.73	78.02	85.30	76.23	80.51
+CLIP-IRM	82.90	76.10	79.38	83.07	59.43	69.01	93.43	76.23	84.36	86.87	78.10	82.25
PromptSRC	83.03	77.47	80.15	83.75	60.53	70.65	92.87	77.40	84.78	84.63	72.90	78.40
+CLIP-IRM	83.27	78.97	81.06	84.80	62.23	71.49	93.23	77.90	84.84	87.73	77.70	82.46

Table 3. Base-to-new generalization performance of six baselines with or without our CLIP-IRM on 11 datasets. "H" is the harmonic mean of Base and New.

- **With/without environment-pruned mid-training**. We compare the encoder settings while keeping the GRPO policy, rewards, and schedules identical: (a) math mid-training on raw captions only (no Emath pruning, no same-class swaps) and (b) full environment-pruned mid-training (Vmath/Emath construction, phrase-level pruning, and environment-invariant swaps). The process rewards are computed with the corresponding encoders but otherwise identical (same windowing, negatives, β, α). We observe the results on OOD reasoning in Geometry3K and the average OOD performance on math reasoning benchmarks (MathVista/WeMath/MathVerse/MathVision) and HallusionBench: full pruning ¿ vanilla GRPO ¿ non-pruned. The gains correlate with two diagnostics: lower average environment-overlap penalties and higher fraction of windows with positive process reward. ID performance on Geometry3K is maintained or slightly improved with pruning,

Method	Geometry3K	OOD Avg.
Qwen2.5-VL-7B-Instruct	39.4	53.3
+ Vanilla GRPO	51.4	57.2
+ GRPO (CLIP-IRM Reward)	**52.7**	**58.8**
Ablation		
w.o. Environment-pruned	50.5	52.9
w.o. Fine V_{math}	46.8	48.7

Table 4. Ablation study on OOD reasoning by the base model (Qwen2.5-VL-7B-Instruct), GRPO, and our approach.

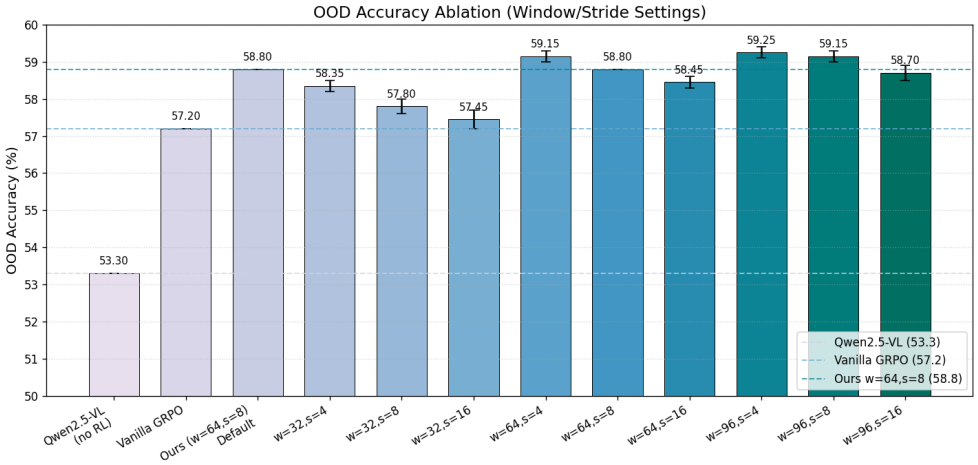


Figure 6. The ablation about the window size $w = 64$ and step size $s = 8$ in our experiment.

indicating the OOD benefit stems from better causal alignment rather than distribution-specific memorization. Qualitatively, the pruned encoders reduce reward "spikes" on boilerplate phrases and increase alignment to diagram entities, yielding more grounded chains of thought.

- **Coarse vs. fine mathematical vocabulary**. We hold $\mathcal{E}_{math}$ fixed and train two CLIP-IRM-math encoders that differ only in the granularity of the class vocabulary: a coarse V_{math} (20–30 high-level geometric/graph primitives) versus a fine V_{math} (120–200 phrase-level entities including compositional types like right triangle, tangent, *etc*. Using the same policy and reward settings (negatives = both; window size=64, step size=8; $\beta = 0.2$), the fine vocabulary consistently improves OOD math benchmarks and patch grounding metrics. The net effect suggests that phrase-level class granularity is important for capturing the causal subspace relevant to visual math reasoning.
- **Window size and Step size.** We also ablate the sliding-window interface that couples the MLLM decoder with the CLIP-IRM text encoder during process-reward computation. A window of recent output tokens with size w is fed to $g(\cdot)$ at every step, while a stride s determines how often we evaluate (i.e., we move the window by s tokens between reward calculations). Unless stated, other settings follow Appendix C: CLIP-IRM encoders are frozen, negatives include in-batch windows and class prototypes, and rewards combine alignment and environment-penalty terms with batch-wise z-scoring. We sweep $w \in \{32, 64, 96\}$ and $s \in \{4, 8, 16\}$, comparing against the base model (Qwen2.5-VL, 53.3%) and vanilla GRPO (57.2%). The default in the main text uses w=64, s=8. In Fig.6, the bar plot shows that all CLIP-IRM process-reward settings outperform vanilla GRPO, with OOD accuracy clustered in the 57.3–59.3% band; the best configuration is w=96, s=8 at 59.25%, closely followed by w=64, s=4 (59.15%) and w=96, s=4 (59.15%), indicating that moderate-to-large windows combined with moderate stride yield the most robust improvements. Very small windows (w=32) underperform larger windows across strides (58.35% at s=4, 57.80% at s=8, 57.45% at s=16), suggesting insufficient context hampers stable invariant alignment. Conversely, overly sparse evaluation (s=16) consistently degrades performance relative to s=4 or s=8 for the same w, implying that denser reward feedback helps credit assignment. Overall, the results support using $w \in [64, 96]$ with $s \in [4, 8]$; our default (w=64, s=8) attains 58.80% and balances accuracy with compute, while the slight gains at w=96, s=8 reflect better phrase-level grounding from longer token windows without incurring the latency of s=4.