Intervention-Guided Image-Free Classifier Expansion for Fine-Grained Recognition

Xiangyu Wang, Changxin Rong, Yanze Gao, Lyuzhou Chen, Xiren Zhou, Huanhuan Chen*
School of Computer Science and Technology
University of Science and Technology of China, Hefei, China

Abstract

Image-free zero-shot classifier expansion aims to extend recognition to unseen classes by establishing an association mapping between class semantics and visual classifier weights, without accessing original images. However, most existing methods rely on an association-based optimization mechanism, which struggles to meet the demands for fine-grained classification of visually and semantically similar classes. Since co-occurring attributes dominate class semantics, the model gravitates towards co-occurring attributes rather than discriminative attributes. This leads to geometrically entangled classifier weights, resulting in ambiguous decision boundaries. To address this problem, this paper proposes a novel approach that reframes weight generation through the lens of causal intervention. The method first employs comparative reasoning to disentangle discriminative and co-occurring attributes. Subsequently, it synthesizes intervened semantics to impose invariance and sensitivity constraints. This mechanism compels the model to suppress co-occurring attributes while enhancing sensitivity to discriminative attributes. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches, yielding clearer decision boundaries in fine-grained settings.

1 Introduction

The constant emergence of new classes in real-world applications necessitates deployed classifiers to continuously extend their recognition capabilities, a process known as classifier expansion [Zhou et al., 2024b]. However, conventional strategies such as full retraining are often inefficient due to the prohibitive costs associated with large-scale data collection and computational resources [Zhou et al., 2024a]. To alleviate this burden, Zero-Shot Learning (ZSL) was introduced to recognize unseen classes by transferring knowledge via semantic information, thereby eliminating the need to collect images for new classes [Lampert et al., 2013]. Despite its efficiency, the reliance on original images poses a severe bottleneck in data-sensitive or resource-constrained deployments where original images are inaccessible [Xian et al., 2018a, Tang et al., 2024]. Recently, Image-Free Zero-Shot Learning (I-ZSL) has emerged as a pragmatic paradigm that synthesizes classifier weights for unseen classes from attribute lists or textual descriptions, without requiring access to any original images [Christensen et al., 2023, Lyu et al., 2025]. Specifically, most existing methods optimize the model by establishing an association mapping between class semantics and visual classifier weights.

In classifier expansion, a key and common characteristic is that expanded classes often include fine-grained classes that are visually and semantically similar [Lee et al., 2023, Pan et al., 2024, Zhao et al., 2025]. However, existing association-based optimization methods face challenges in fine-grained settings. Specifically, for similar classes, their semantics are dominated by abundant co-occurring attributes, while the discriminative attributes that distinguish similar classes are few. Since association-based optimization lacks an explicit mechanism to prioritize discriminative attributes,

*Corresponding authors: Huanhuan Chen (hchen@ustc.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

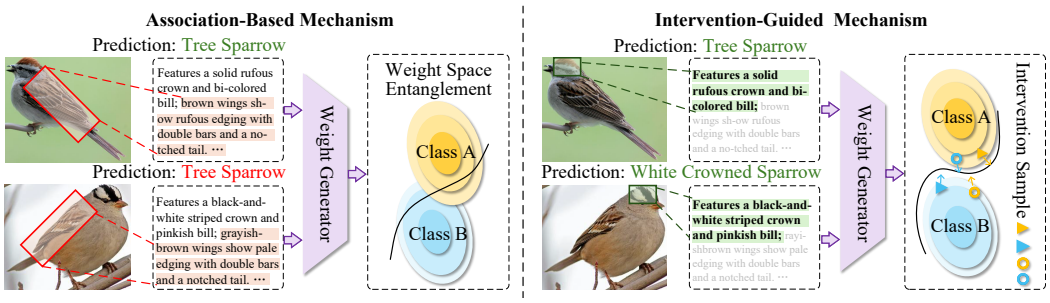


Figure 1: Comparison of association-based and intervention-guided mechanisms. The former relies on dominant co-occurring attributes and yields entangled weights, while the latter isolates discriminative attributes via causal intervention, producing disentangled weights and clearer boundaries.

it tends to indiscriminately associate visual weights with dominant co-occurring attributes, thereby optimizing the model toward co-occurring attributes rather than discriminative attributes. This leads to geometrically entangled classifier weights[2], resulting in ambiguous decision boundaries.

To overcome the limitations of association-based optimization mechanism, causal intervention can offer a novel perspective to reframe weight generation. Causal intervention identifies the genuine drivers of an outcome by actively manipulating variable factors to distinguish true causal effects from indiscriminate associations [Pearl, 2009]. Leveraging this capability enables the imposition of constraints on the optimization process, enabling the model to capture discriminative attributes and eliminate weight entanglement, as illustrated in the Figure 1.

However, due to the characteristics of fine-grained settings, the implementation of causal intervention faces challenges. First, when distinguishing similar classes, the discriminative attributes for a specific class pair may degenerate into noise when distinguishing other classes[3]. Therefore, constructing fixed attribute descriptions for different classes using a fixed rule or a unified pattern is ineffective, even if they are accurate. Second, due to the semantic similarity across classes, classes often exhibit abundant co-occurring attributes, while the attributes that can distinguish similar classes are few and subtle[4]. Thus, merely leveraging discriminative attributes struggles to counter the dominance of co-occurring attributes. Crucially, it is difficult to perform interventions in the semantic space based on discriminative and co-occurring attributes and to translate such abstract interventions into explicit geometric constraints in the classifier weight space that optimize the model to suppress co-occurring attributes while enhancing sensitivity to discriminative attributes.

To address these challenges, this paper proposes a novel framework termed Intervention-Guided Image-free Classifier Expansion (IGICE). Specifically, this paper first introduces a pairwise causal disentanglement module. It utilizes comparative reasoning to capture the relativity of attribute discriminability, isolating discriminative attributes from co-occurring attributes for specific class pairs. Subsequently, the joint causal intervention module actively synthesizes intervened semantics to simulate causal scenarios of contextual variations and identity shifts. These semantic variations are then fused into class embeddings through the shared-query semantic aggregation. Finally, the intervention-guided weight refinement module projects these synthesized semantics into the visual weight space. By treating weights derived from these interventions as positive and negative samples, it imposes invariance and sensitivity constraints to suppress the influence of co-occurring attributes while aggressively focusing on core discriminative attributes, forging robust fine-grained decision boundaries. The contributions of this paper are:

- This paper introduces causal intervention into fine-grained I-ZSL, reframing fine-grained classifier expansion to eliminate weight entanglement.
- This paper proposes an intervention strategy that simulates contextual variations and identity shifts, enabling interventions in the semantic space.

[2]We perform a theoretical analysis of this phenomenon, with more details in Appendix B.2.
[3]For instance, "wings" serve as a discriminative attribute for bird vs. dog, but degenerate into co-occurring for bird vs. bird.
[4]For instance, two bird species may share 90% of attributes like "wings" and "tail", differing only in a subtle "beak color".

- This paper introduces invariance and sensitivity constraints, enforcing robustness to co-occurring attributes, enhancing discriminative sensitivity.

2 Related Work

This section first discuss the limitations of existing Image-Free ZSL methods regarding correlation-based optimization, and subsequently examine the application of causal principles in vision tasks, highlighting the gaps that motivate our proposed approach.

2.1 Zero-Shot Learning

Existing Zero-Shot Learning approaches can be categorized into three paradigms including embedding-based, generative-based, and vision-language model-based methods [Xian et al., 2018a, Wang et al., 2019]. Embedding-based methods aim to learn a projection function that maps visual features and text descriptions into a common embedding space, where classification is performed via nearest-neighbor search [Frome et al., 2013, Akata et al., 2015, Lampert et al., 2013, Xian et al., 2016]. Generative-based methods leverage generative models, such as generative adversarial networks and variational autoencoders, to learn seen data distributions. By synthesizing pseudo-visual features for unseen classes, they convert ZSL into a supervised classification problem [Xian et al., 2018b, Narayan et al., 2020, Mishra et al., 2018, Schonfeld et al., 2019]. More recently, vision-language model-based methods treat ZSL as open-vocabulary recognition by encoding images and natural-language prompts into a shared space, and performing classification via prompt-conditioned similarity [Radford et al., 2021, Zhou et al., 2022, Liu et al., 2026]. Despite their success, all these approaches typically require original images of seen classes to align visual and text modalities, limiting their applicability in privacy-sensitive or data-restricted scenarios.

To mitigate this dependency, Image-Free ZSL learns an association mapping from attribute descriptions to classifier weights without accessing images [Christensen et al., 2023, Lyu et al., 2025]. However, most methods rely on association-based optimization. In fine-grained settings, this paradigm is limited as descriptions are dominated by co-occurring attributes. Consequently, the model gravitates towards co-occurring noise rather than discriminative attributes, leading to geometrically entangled weights and ambiguous decision boundaries.

2.2 Causal Inference for Zero-Shot Learning

To overcome the limitations of association-based optimization, causal intervention offers a principled framework for separating causal factors from spurious correlations. Formalized via do-calculus [Pearl, 2009, Peters et al., 2017], this mechanism simulates controlled interventions to sever non-causal dependencies between variables, thereby isolating factors that genuinely affect the target outcome. Due to its efficacy in eliminating biases [Yang et al., 2021], causal intervention has been widely adopted in various computer vision tasks, including long-tailed recognition [Tang et al., 2020], visual question answering [Niu et al., 2021, Li et al., 2025], and domain generalization [Shao et al., 2024].

In the context of Zero-Shot Learning, many studies have incorporated causal principles to disentangle semantic attributes or separate objects from environmental contexts [Yue et al., 2021, Chen et al., 2024]. However, these approaches are mainly designed for conventional ZSL settings where visual samples or image-level representations are available, and they typically operate on global feature selection or compositional interventions at the image level. They therefore do not directly address image-free classifier expansion, where original images are inaccessible and the key objective is to generate classifier weights from semantics alone.

Crucially, existing causal ZSL methods overlook the relativity of attribute importance, where an attribute can be discriminative for one class pair but co-occurring for another. Different from prior methods, our approach captures this pairwise relativity and translates semantic interventions into invariance and sensitivity constraints in the classifier weight space, thereby addressing co-occurrence-induced weight entanglement in image-free classifier expansion.

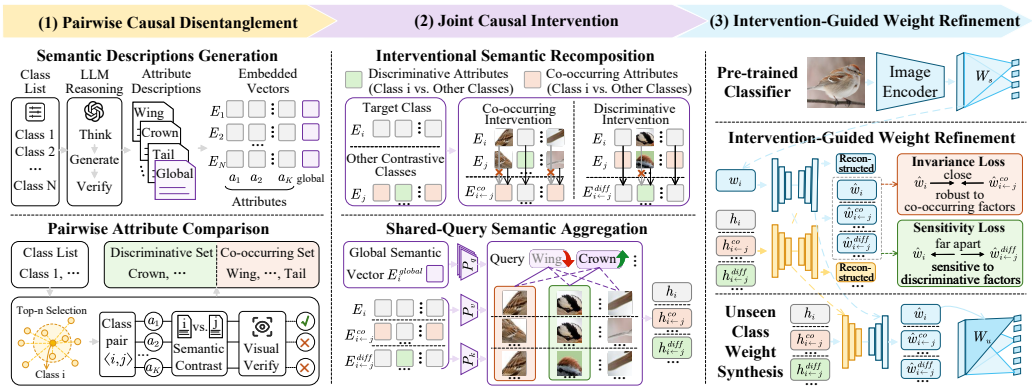


Figure 2: The overall architecture of the proposed framework. The figure illustrates the data flow across its three core modules: (a) Pairwise Causal Disentanglement (PCD), (b) Joint Causal Intervention (JCI), and (c) Intervention-Guided Weight Refinement (IWR).

3 Method

This section first formalizes the I-ZSL problem, then details the proposed IGICE framework. It comprises three key modules: pairwise causal disentanglement to isolate semantic attributes, joint causal intervention to simulate causal scenarios, and intervention-guided weight refinement to optimize decision boundaries. The overall architecture is depicted in Figure 2.

3.1 Problem Formulation

For class expansion in the I-ZSL setting, we start with a classifier pre-trained on seen classes $\mathcal{S}$, whose classification layer is parameterized by weight vectors $W_{\mathcal{S}} = \{w_s \in \mathbb{R}^d \mid s \in \mathcal{S}\}$, where d denotes the dimension of the visual features. The objective of I-ZSL is to extend recognition to disjoint unseen classes $\mathcal{U}$ strictly without accessing original images, by learning a mapping $\mathcal{G}$ that projects semantics into the weight space. $\mathcal{G}$ synthesizes weights $W_{\mathcal{U}}$ solely from text descriptions, which are then concatenated with the frozen $W_{\mathcal{S}}$ to enable classification on the joint space $\mathcal{S} \cup \mathcal{U}$.

To overcome the limitations of association-based optimization in fine-grained settings, our work introduces a causal intervention framework designed to explicitly guide the optimization of the model. Specifically, we first employ causal disentanglement to isolate discriminative and co-occurring attributes, constructing intervened semantics that simulate distinct causal scenarios. Subsequently, we project these inputs into the weight space to impose invariance and sensitivity constraints. This optimization regime enables the model to suppress dependency on co-occurring attributes while enhancing sensitivity to discriminative ones, thereby forging clear and robust decision boundaries.

3.2 Pairwise Causal Disentanglement

To effectively implement the causal intervention, identifying the intervention targets is an essential step. To this end, we introduce the pairwise causal disentanglement module to partition the attribute space of class pair into a co-occurring attribute set $\mathcal{C}$ and a discriminative attribute set $\mathcal{D}$. Specifically, $\mathcal{C}$ comprises dominant yet non-discriminative attributes serving as the primary noise source, while $\mathcal{D}$ encapsulates critical visual discrepancies representing essential causal signals. This structured disentanglement isolates discriminative and co-occurring attributes, providing operable targets for the subsequent causal intervention mechanism.

We first generate text descriptions for all classes. To ensure structural alignment of different classes in the semantic space, for each class, its text description is built on a set of shared attributes $\mathcal{A} = \{a_1, \ldots, a_K\}$. Under each attribute, there is a corresponding natural language text describing that class. To improve practicality and description accuracy, an LLM is used as the execution engine for description generation. It produces concise attribute-wise descriptions over $\mathcal{A}$ and generates a global class description T_i^{global}. Specific prompt designs and implementation details are provided in

Appendix F. The semantic description for a class i is represented as a set of attribute-specific texts:

$$\mathcal{T}_i = \{T_{i,1}, T_{i,2}, \ldots, T_{i,K}\} \tag{1}$$

where $T_{i,k}$ denotes the generated textual description of class i corresponding to the k-th attribute dimension $a_k \in \mathcal{A}$. Subsequently, we utilize a pre-trained text encoder $\mathcal{F}_{\text{text}}$ (e.g., CLIP) to independently encode these textual components. The attribute-specific texts are encoded and stacked to form the semantic matrix E_i:

$$E_i = [E_{i,1}, E_{i,2}, \ldots, E_{i,K}]^\top \in \mathbb{R}^{K \times d_T} \tag{2}$$

where $E_{i,k} = \mathcal{F}_{\text{text}}(T_{i,k})$ denotes the d_T-dimensional embedding vector for the k-th attribute. In parallel, the global description is encoded into a global semantic vector $E_i^{global} = \mathcal{F}_{\text{text}}(T_i^{global}) \in \mathbb{R}^{d_T}$, serving as a semantic anchor for subsequent aggregation. Consequently, the final semantic representation for class i preserves the fine-grained attribute structure within E_i while retaining the global description via E_i^{global}.

While obtaining structured attribute descriptions, utilizing these original semantics directly leads to weight entanglement. To address this, we formulate the weight generation process via a Structural Causal Model (SCM). Within this framework, $\mathcal{C}$ induces a spurious shortcut pathway that establishes a non-causal link between semantic descriptions and visual weights via co-occurring attributes. This path introduces indiscriminate associations, causing the optimization to mistakenly fit dominant noise rather than discriminative attributes. In contrast, $\mathcal{D}$ constitutes the genuine drivers governing the classification weight. Consequently, suppressing the shortcut influence of $\mathcal{C}$ while preserving the causal contribution of $\mathcal{D}$ represents the theoretical prerequisite for valid intervention. Please refer to Appendix B.1 for the detailed theoretical derivation.

Furthermore, the discriminability of attributes is not absolute, but relative. To accommodate this, we propose a comparative reasoning strategy. For any pair (i, j), we partition attributes into pair-specific sets $\mathcal{C}_{ij}$ and $\mathcal{D}_{ij}$, ensuring that the identification of discriminative attributes is aligned with the specific decision boundary of that pair. In practice, we restrict pairwise disentanglement to the top-n semantically nearest classes of each target class, which focuses comparisons on the most confusing pairs and substantially reduces the practical preprocessing cost, as detailed in Appendix E.9.

To operationalize this with stabity and accuracy, we leverage an LLM evaluator Ψ to assess attributes, guided by a Chain-of-Thought strategy (see Appendix F). Formally, for the k-th attribute, the LLM generates a binary indicator:

$$m_{i,j,k} = \Psi_{\text{LLM}}(T_{i,k}, T_{j,k} \mid \mathcal{I}) \tag{3}$$

where $\mathcal{I}$ represents the prompt that defines the criteria for visual discriminability. Specifically, to address the text-visual gap, $m_{i,j,k}$ is set to 1 if and only if the attribute implies an explicit visual discrepancy. Ultimately, we partition the attribute set $\mathcal{A}$ into two complementary subsets based on this indicator:

$$D_{ij} = \{a_k \in \mathcal{A} \mid m_{i,j,k} = 1\}, \qquad C_{ij} = \{a_k \in \mathcal{A} \mid m_{i,j,k} = 0\}. \tag{4}$$

where $\mathcal{D}_{ij}$ and $\mathcal{C}_{ij}$ denote the discriminative and co-occurring attribute sets, satisfying $\mathcal{C}_{ij} \cup \mathcal{D}_{ij} = \mathcal{A}$ and $\mathcal{C}_{ij} \cap \mathcal{D}_{ij} = \emptyset$.

3.3 Joint Causal Intervention

To explicitly construct intervened semantics, we design a joint causal intervention module. This module establishes a mechanism that jointly manipulates distinct attribute sets. Specifically, the first branch intervenes exclusively on co-occurring attributes to simulate contextual variations while maintaining class identity, whereas the second branch intervenes on discriminative attributes to simulate class identity variations. By jointly recomposing these intervened attributes at the semantic level, the module generates comprehensive intervened semantics for each class.

First, we construct the branch that intervenes exclusively on co-occurring attributes for each class pair (i, j). Formally, we define this process as applying a do-intervention on the co-occurring attribute set $\mathcal{C}_{ij}$ within the semantic space. The resulting intervened semantics $E_{i \leftarrow j}^{co}$ is defined as:

$$E_{i \leftarrow j}^{co} \triangleq E_i(do(E_{i,k} = E_{j,k} \mid a_k \in \mathcal{C}_{ij})) \tag{5}$$

Intuitively, this operation severs the link to i's original non-discriminative information and forcibly injects the background of j, thereby simulating contextual variation while maintaining class identity.

Complementarily, we construct the branch that intervenes exclusively on discriminative attributes. Corresponding to the causal operation, we apply a do-intervention on the discriminative attribute set $\mathcal{D}_{ij}$. The resulting intervened semantics $E^{diff}_{i\leftarrow j}$ is defined as:

$$E^{diff}_{i\leftarrow j} \triangleq E_i(do(E_{i,k} = E_{j,k} \mid a_k \in \mathcal{D}_{ij})) \tag{6}$$

This formulation captures a distinct intervention scenario where the intervened semantics retains the context of class i, but its class identity has been changed by those of j, compelling a robust model to change its recognition from i.

After obtaining the original and interventional semantic matrices, we need to aggregate them into unified class embeddings. Crucially, due to the discrepancy between original and intervened semantics, inconsistent aggregation standards induce feature extraction patterns that rely on specific semantics, hindering the focus on core attributes. Thus, it is crucial to establish a consistent aggregation mechanism that maintains focus on discriminative attributes across both original and intervened semantics. To this end, we propose a shared-query semantic aggregation strategy. Specifically, we apply the same query vectors Q to aggregate the semantic information from all three distinct branches: the original attributes E_i, the intervened co-occurring attributes $E^{co}_{i\leftarrow j}$, and the intervened discriminative attributes $E^{diff}_{i\leftarrow j}$.

Meanwhile, to capture diverse discriminative aspects (e.g., shape and color) beyond the capacity of a single and static vector, we employ a multi-head mechanism. Specifically, we employ a projection network $\mathcal{P}(\cdot)$ to map the global anchor E^{global}_i into H distinct query heads:

$$\{Q_{i,1}, \ldots, Q_{i,H}\} = \mathcal{P}(E^{global}_i), \tag{7}$$

where each head $Q_{i,h}$ is optimized to attend to a specific semantic subspace. Formally, we utilize these generated heads to perform shared-query aggregation across the three branches. The unified aggregation process is formulated as:

$$\begin{aligned} h_i &= \text{Concat}^H_{h=1}(\text{Attention}(Q_{i,h}, E_i, E_i)), \\ h^{co}_{i\leftarrow j} &= \text{Concat}^H_{h=1}(\text{Attention}(Q_{i,h}, E^{co}_{i\leftarrow j}, E^{co}_{i\leftarrow j})), \\ h^{diff}_{i\leftarrow j} &= \text{Concat}^H_{h=1}(\text{Attention}(Q_{i,h}, E^{diff}_{i\leftarrow j}, E^{diff}_{i\leftarrow j})). \end{aligned} \tag{8}$$

Through this mechanism, the model performs consistent aggregation, learning a robust policy to consistently focus on discriminative attributes, unaffected by semantic variations.

3.4 Intervention-Guided Weight Refinement

Following semantic intervention, our goal is to leverage the intervened semantic samples to guide the optimization process and refine decision boundaries within the visual weight space. To this end, we design the intervention-guided weight refinement module, which maps the aggregated original and intervened semantics into classifier weights. We treat the weights derived from co-occurring interventions as positive samples, and those from discriminative interventions as negative samples relative to the original anchor. On this basis, we employ complementary Invariance and Sensitivity losses to optimize the generator. This strategy effectively translates semantic causal interventions into direct geometric constraints, forging clearer fine-grained classification boundaries.

We employ a shared weight generator $\mathcal{G}(\cdot)$, instantiated as a MLP within an autoencoder framework [Christensen et al., 2023], to project the aggregated semantic embeddings into the visual weight space. Specifically, we generate the original weight $\hat{w}_i$ and its intervened weight by passing the respective semantic features through $\mathcal{G}$:

$$\hat{w}_i = \mathcal{G}(h_i), \hat{w}^{co}_{i\leftarrow j} = \mathcal{G}(h^{co}_{i\leftarrow j}), \hat{w}^{diff}_{i\leftarrow j} = \mathcal{G}(h^{diff}_{i\leftarrow j}) \tag{9}$$

Targeting the weights $\hat{w}^{co}_{i\leftarrow j}$ generated by intervening solely on co-occurring attributes, we impose an invariance constraint. Since this intervention preserves the core discriminative attributes, the class identity remains strictly unchanged. Consequently, we explicitly define $\hat{w}^{co}_{i\leftarrow j}$ as a positive sample relative to the original anchor $\hat{w}_i$. We formulate the invariance loss $\mathcal{L}_{inv}$ by minimizing the cosine distance between this positive pair:

$$\mathcal{L}_{inv} = \mathbb{E}_{i,j\in\mathcal{S}\cup\mathcal{U}}[d_{inv}(\hat{w}_i, \hat{w}^{co}_{i\leftarrow j})]. \tag{10}$$

This loss compels the model to suppress the co-occurring attributes.

Conversely, targeting the weights $\hat{w}_{i\leftarrow j}^{diff}$ generated by intervening in discriminative attributes, we impose a sensitivity constraint. Since the identity-defining discriminative attributes have been altered, we explicitly treat $\hat{w}_{i\leftarrow j}^{diff}$ as a negative sample relative to the original anchor $\hat{w}_i$. To enforce separability, we employ a sensitivity loss $\mathcal{L}_{sen}$ designed to maximize geometric divergence between this negative pair:

$$\mathcal{L}_{sen} = \mathbb{E}_{i,j\in\mathcal{S}\cup\mathcal{U}}[d_{sen}(\hat{w}_i, \hat{w}_{i\leftarrow j}^{diff})]. \tag{11}$$

By penalizing high similarity, we force the generated weights to shift in response to changes in discriminative attributes. This forces the weight space geometry to align with discriminative attributes, ensuring that fine-grained boundaries are governed by true causal distinctions.

The overall objective combines the baseline loss $\mathcal{L}_{base}$ (detailed in the Appendix C.3) with constraints:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda_{inv}\mathcal{L}_{inv} + \lambda_{sen}\mathcal{L}_{sen} \tag{12}$$

where λ denotes the balancing hyperparameters.

During inference, we generate weights for unseen classes $\mathcal{U}$ using the trained $\mathcal{G}$. These are concatenated with the frozen seen weights $W_{\mathcal{S}}$, enabling recognition on the joint space $\mathcal{S}\cup\mathcal{U}$ without accessing any unseen visual data.

4 Experiments

This section aims to comprehensively validate the effectiveness of the IGICE framework. We first demonstrate the performance of our method compared to existing state-of-the-art methods on fine-grained classification tasks. Then, through systematic ablation studies and robustness analyses, we analyze the core role of the causal intervention mechanism in addressing fine-grained challenges.

4.1 Experimental Setup

Datasets. We used three benchmark datasets for evaluation: CUB [Wah et al., 2011] containing 200 bird species, SUN [Patterson et al., 2014] containing 717 scene classes, and AWA2 [Xian et al., 2018a] containing 50 animal classes. The CUB and SUN datasets, with their high inter-class similarity, are fine-grained datasets and can serve as primary testbeds for validating our co-occurrence interference resolution scheme. Each dataset is divided into seen and unseen classes according to the segmentation scheme [Xian et al., 2018a]. Following image-free setting, we did not use any images from datasets.

Baselines. To validate the effectiveness of our method, we compare it against a comprehensive set of representative Image-Free ZSL methods, all operating under the strict image-free setting. The first class comprises classical methods, including ConSE [Norouzi et al., 2014] which uses convex combinations of semantic embeddings, and COSTA [Mensink et al., 2014] which leverages co-occurrence statistics. The second class consists of advanced optimization methods, including SubReg [Akyürek et al., 2021] which applies subspace regularization to constrain unseen weights, and VGSE [Xu et al., 2022] which aligns visual and semantic spaces via visual clustering. Finally, we compare with state-of-the-art methods, including ICIS [Christensen et al., 2023] utilizing dual autoencoders for weight synthesis, and the recent LLM-driven CEMIL [Lyu et al., 2025] method, which generates multi-view descriptions for classifier expansion.

Evaluation Metrics. We perform evaluations under Standard ZSL and Generalized ZSL (GZSL), where the evaluation scope are restricted to unseen classes and the union of seen and unseen classes, respectively. Accordingly, we report the average per-class Top-1 accuracy T for Standard ZSL. For GZSL, we measure the accuracies on seen classes S and unseen classes U, reporting their Harmonic Mean H, defined as $H = (2 \times S \times U)/(S + U)$, to assess the overall performance balance.

Implementation Details. For visual features, we utilize a ResNet101 pre-trained on ImageNet to extract 2048-dimensional vectors without fine-tuning. For semantic features, we use GPT-4o for text generation and the pre-trained CLIP text encoder with a ViT-B/32 backbone to extract attribute-level embeddings. The weight generator is implemented as a Multi-Layer Perceptron and optimized using the Adam optimizer. Detailed experimental settings are provided in the Appendix E.1.

Table 1: Comparison with baselines on three benchmarks under ZSL (Top-1 Acc **T**) and GZSL (Harmonic Mean **H**, unseen **u**, seen **s**) settings. "ΔImp" denotes the relative improvement $(p_2 - p_1)/p_1$ of our method (p_2) over the SOTA baseline (p_1) in terms of **T** and **H**.

Semantic Source	Method	Zero-Shot Accuracy			Generalized Zero-Shot Accuracy								
		CUB	SUN	AWA2	CUB			SUN			AWA2		
		T	T	T	H	u	s	H	u	s	H	u	s
Expert-constructed Class Attributes	ConSE	41.9	44.4	44.0	0.9	0.5	88.0	0.1	0.1	47.9	5.7	3.0	96.1
	COSTA	31.9	19.9	40.9	0.0	0.0	87.6	0.0	0.0	50.1	0.0	0.0	96.1
	SubReg	37.6	48.3	37.5	0.0	0.0	87.6	0.0	0.0	50.1	0.0	0.0	96.1
	wDAE	38.2	49.9	37.0	0.0	0.0	87.3	0.0	0.0	49.3	0.3	0.1	96.0
	VGSE	45.1	42.7	55.4	44.8	39.2	52.3	3.1	42.5	1.6	47.3	31.8	92.4
	ICIS	60.6	51.8	64.6	56.5	45.8	73.7	32.7	45.2	25.6	51.6	35.6	93.3
LLM-generated Class Attributes	ConSE	45.8	44.4	51.4	0.9	0.5	88.3	0.3	0.1	50.2	4.5	2.3	96.3
	COSTA	35.8	29.4	57.3	0.0	0.0	87.6	0.0	0.0	52.1	0.0	0.0	96.2
	SubReg	57.3	55.3	65.0	3.2	1.6	87.9	1.9	1.0	52.3	0.1	0.1	96.2
	wDAE	59.3	50.6	64.1	3.4	1.7	87.4	0.1	0.1	51.1	8.1	4.2	95.8
	VGSE	49.8	44.4	58.3	48.7	42.3	57.2	17.5	42.4	11.0	53.8	38.9	87.1
	ICIS	61.4	55.8	65.9	56.9	46.2	74.0	37.8	44.8	32.7	57.4	42.0	90.8
	CEMIL	62.2	56.5	68.9	57.9	50.2	68.3	38.5	41.0	36.3	61.2	46.8	88.6
	Ours	**71.5**	**62.7**	**74.9**	**63.2**	64.8	61.6	**43.9**	45.2	42.6	**66.7**	60.3	74.6
	ΔImp	14.9%	11.0%	8.7%	9.2%	-	-	14.0%	-	-	9.0%	-	-

4.2 Comparison with State-of-the-Arts

To rigorously evaluate the effectiveness of IGICE, we conduct a comprehensive comparative study against a diverse range of methods under both standard and generalized ZSL settings. To ensure a fair assessment regarding semantic quality, we evaluate baselines using two distinct semantic sources: 1) traditional expert attributes, and 2) our LLM-generated semantics. For the latter, to accommodate baselines requiring holistic inputs, we aggregate the attribute-level embeddings via mean pooling to form unified semantic vectors.

As reported in Table 1, IGICE achieves state-of-the-art performance across all configurations, with larger gains on fine-grained CUB and SUN than on coarse-grained AWA2. Since CUB and SUN contain stronger co-occurring noise, this result supports our premise that causal intervention mitigates co-occurrence-dominated weight entanglement. Moreover, the gains over ICIS and CEMIL show that explicit geometric constraints, including invariance and sensitivity, are more effective for precise weight learning than reconstruction objectives or unstructured semantic enrichment.

4.3 Ablation Studies

To rigorously verify the individual contribution of each component within our framework, we performed a systematic ablation study involving five variants compared against the full model. (1) Regarding the model architecture, we investigated the necessity of shared-query aggregation by replacing the shared-query attention with standard mean pooling, respectively. (2) To assess the causal strategy, we substituted our fine-grained pairwise causal disentanglement with a coarse-grained global disentanglement approach based on overall dataset statistics. (3) Furthermore, we examined the geometric constraints by independently excluding the sensitivity loss, the invariance loss, and finally removing both causal losses simultaneously to isolate their specific roles in weight refinement.

Table 2: Ablation study on the core components of the proposed framework. We report Zero-Shot Accuracy (**T**) and Generalized Zero-Shot metrics (**H**, **u**, **s**) on CUB, SUN, and AWA2 datasets.

Method Ablation	Zero-Shot Accuracy			Generalized Zero-Shot Accuracy								
	CUB	SUN	AWA2	CUB			SUN			AWA2		
	T	T	T	H	u	s	H	u	s	H	u	s
Full Model (Ours)	**71.5**	**62.7**	**74.9**	**63.2**	64.8	61.6	**43.9**	45.2	42.6	**66.7**	60.3	74.6
w/o Shared-query Attention	70.7	60.8	73.1	61.4	58.7	64.4	42.7	41.9	43.5	64.8	55.2	78.5
w/o Sensitivity Loss	68.6	58.3	71.0	59.3	56.2	62.8	40.3	41.6	39.1	63.9	53.1	80.2
w/o Invariance Loss	67.3	58.6	70.2	59.8	57.5	62.3	39.0	43.3	35.5	62.5	50.3	82.4
w/o Pairwise Disentanglement	66.9	57.6	69.1	58.8	55.8	62.1	38.8	40.8	36.9	61.8	48.8	84.1
w/o Causal Losses	63.0	56.8	68.1	58.4	54.0	63.5	37.9	37.6	38.1	59.6	45.5	86.5

As shown in Table 2, all variants suffer consistent performance drops, confirming the necessity of our overall design. The degradation caused by global disentanglement shows that attribute causality is context-dependent, while the inferior result of mean pooling indicates that simple averaging cannot effectively filter dominant co-occurring noise. The loss ablations further reveal the complementarity of our constraints: removing invariance or sensitivity weakens noise suppression and discriminative sensitivity, respectively, and removing both causes the largest drop. These results confirm that translating semantic causal interventions into geometric constraints is essential for robust fine-grained decision boundaries.

4.4 Further Empirical Analysis

To comprehensively validate the effectiveness and robustness of our framework, we conducted a series of extensive analyses. Due to space constraints, we briefly summarize two key studies here.

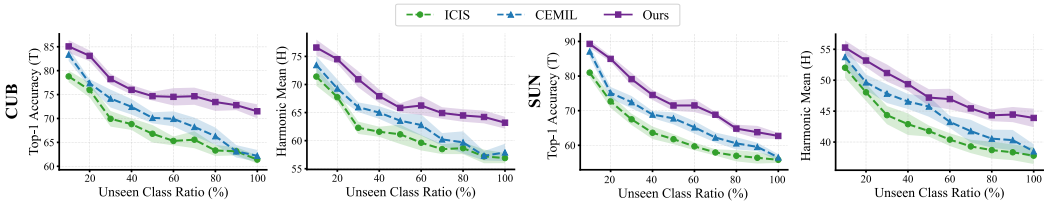


Figure 3: Comparative Harmonic Mean and Top-1 accuracy of the baseline (ICIS, CEMIL) and our method across proportions of expanded classes.

Fine-Grained Robustness. To verify anti-interference capabilities in high-density semantic environments where co-occurring attributes increase rapidly, we designed a progressive expansion protocol that increases the ratio of unseen classes from 20% to 100% on CUB and SUN dataset. As shown in Fig. 3, while baseline methods suffer severe performance collapse due to geometric entanglement, our method demonstrates exceptional stability. Notably, the performance gap maximizes under the extremely crowded 100% condition, confirming that our model maintains good performance even when discriminative features are heavily overshadowed by co-occurrence noise. Detailed settings and results are provided in Appendix E.3.

Table 3: Comparative analysis of our method with various LLMs and embedding models on CUB.

LLM	ZSL (Top-1 Accuracy)				GZSL (Harmonic Mean)			
	CLIP	SBERT	LLaMA	Qwen	CLIP	SBERT	LLaMA	Qwen
GPT-4o	**71.5±1.1**	**66.5±1.8**	**65.8±2.5**	**66.8±2.1**	**63.2±0.9**	58.4±1.3	58.8±1.6	59.2±1.6
Gemini-2.5	71.2±1.3	66.1±2.0	65.4±1.5	66.5±1.8	62.8±1.1	**59.1±1.5**	**58.8±1.2**	**59.5±1.4**
Claude-4	70.9±1.5	65.8±1.9	64.9±2.1	66.1±1.2	62.5±1.4	58.2±1.1	57.6±1.4	59.0±1.0
LLaMA-3.3	67.8±1.9	63.5±1.5	63.8±2.1	64.4±1.8	60.5±1.4	57.4±0.9	57.4±0.9	57.8±1.2
DeepSeek-V3	67.5±1.6	63.8±2.0	64.2±1.8	63.9±1.5	59.8±1.3	57.5±1.1	58.0±1.0	57.8±1.4
Qwen-2.5	66.9±1.1	62.7±2.5	63.6±2.5	63.2±1.4	58.9±0.8	56.0±0.9	57.4±1.1	57.4±0.8

Stability Analysis. To assess whether IGICE depends on a specific semantic generator or embedding space, we conduct broad substitution experiments across different LLMs and text encoders. Specifically, we evaluate both proprietary LLMs, including GPT-4o, Gemini-2.5, and Claude-4, and open-source models, including LLaMA-3.3, DeepSeek-V3, and Qwen-2.5. We further consider different embedding models, including the vision-language-aligned CLIP encoder, the text-only SBERT encoder, and LLM-based embeddings from LLaMA and Qwen. As shown in Table 3, IGICE maintains stable performance across these configurations, indicating that the proposed intervention-guided refinement is not tied to a particular LLM or semantic embedding architecture. See Appendix E.4 for full comparisons.

Case-Level Probability Analysis. To further examine fine-grained class separability, we analyze eight challenging samples from CUB and compare the prediction probabilities of the baseline and our method in Fig. 4. The baseline often assigns relatively low confidence to the ground-truth class while producing high responses for visually similar alternatives. In contrast, our method consistently increases the confidence of the correct class and suppresses confusing predictions, indicating that

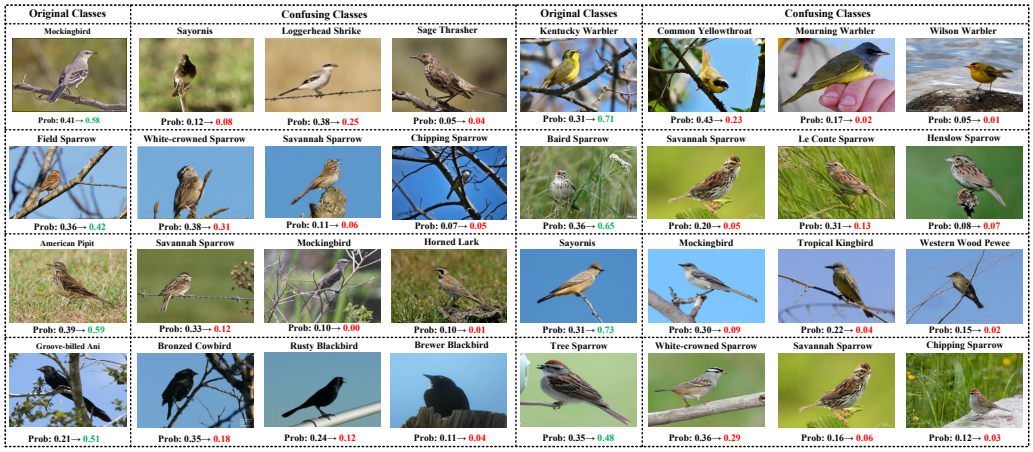


Figure 4: Case-level probability analysis on eight challenging CUB samples. We compare the confidence assigned to the ground-truth class and confusing alternatives before and after intervention-guided refinement.

intervention-guided refinement improves prediction margins by emphasizing discriminative cues and reducing interference from co-occurring attributes.

Further experiments are detailed in Appendix E. Specifically, we provide confusion and case analyses to examine how IGICE reduces ambiguity among visually similar classes, and causal disentanglement analyses to study the effects of neighbor selection, reasoning strategies, and pairwise attribute partitioning. We further evaluate robustness to simulated LLM reasoning errors and conduct sensitivity analyses on the main intervention-related hyperparameters. Extended experiments on additional datasets are included to assess generalization across different domains and class granularities. Finally, we analyze computational efficiency, LLM preprocessing cost, and scalability with respect to the number of classes to evaluate the practical deployment overhead of the proposed framework.

5 Conclusion

Extending recognition to fine-grained classes is important yet challenging due to high class similarity. Existing association-based optimization struggles in this setting, often indiscriminately associating visual weights with dominant co-occurring noise and resulting in entangled classifier weights. To address this, we proposed IGICE, a framework that reframes weight generation via causal intervention. Building upon the intervened semantics constructed by pairwise causal disentanglement and joint causal intervention, we successfully translated abstract intervened semantics into geometric invariance and sensitivity constraints. This mechanism compels the model to suppress co-occurring attributes while enhancing sensitivity to discriminative attributes. Extensive experiments confirm that IGICE outperforms state-of-the-art methods, yielding clearer decision boundaries and offering a robust and scalable solution for precise fine-grained classifier expansion.

Acknowledgements

This research is supported in part by the National Key R&D Program of China (No. 2026YFE0104800), in part by the National Natural Science Foundation of China (No. 62406302, 62137002, 62576327), in part by the Natural Science Foundation of Anhui Province (No. 2408085QF195), in part by the Fundamental Research Funds for the Central Universities under Grant WK2150110035, in part by the USTC Research Funds of the Double First-Class Initiative (Grant No. YD9110002085 and YD2155202608), and in part by the Research Funds of Centre for Leading Medicine and Advanced Technologies of IHM (Grant No. 2025IHM01030).

References

Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

Afra Feyza Akyürek, Ekin Akyürek, Derry Tanti Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. *arXiv preprint arXiv:2110.07059*, 2021.

Shuhuang Chen, Dingjie Fu, Shiming Chen, Shuo Ye, Wenjin Hou, and Xinge You. Causal visual-semantic correlation for zero-shot learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4246–4255, 2024.

Anders Christensen, Massimiliano Mancini, A Koepke, Ole Winther, and Zeynep Akata. Image-free classifier injection for zero-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19072–19081, 2023.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2013.

Byung Hyun Lee, Okchul Jung, Jonghyun Choi, and Se Young Chun. Online continual learning on hierarchical label expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11761–11770, 2023.

Wei Li, Zhixin Li, Fuyun Deng, Kun Zeng, and Canlong Zhang. Counterfactual causal inference for robust visual question answering. *Neural Networks*, page 108115, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Honglin Liu, Chao Sun, Peng Hu, Yunfan Li, and Xi Peng. Conditional representation learning for customized tasks. *Advances in Neural Information Processing Systems*, 38:31706–31737, 2026.

Derui Lyu, Xiangyu Wang, Taiyu Ban, Lyuzhou Chen, Xiren Zhou, and Huanhuan Chen. Expanding the category of classifiers with llm supervision. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5905–5913, 2025.

Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. COSTA: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014.

Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 2188–2196, 2018.

Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, pages 479–495, 2020.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710, 2021.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2014.

Zicheng Pan, Weichuan Zhang, Xiaohan Yu, Miaohua Zhang, and Yongsheng Gao. Pseudo-set frequency refinement architecture for fine-grained few-shot class-incremental learning. *Pattern Recognition*, 155:110686, 2024.

Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1–2): 59–81, 2014.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.

Youjia Shao, Shaohui Wang, and Wencang Zhao. A causality-aware perspective on domain generalization via domain intervention. *Electronics*, 13(10):1891, 2024.

Bowen Tang, Jing Zhang, Long Yan, Qian Yu, Lu Sheng, and Dong Xu. Data-free generalized zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5108–5117, 2024.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Computation & Neural Systems Technical Report*, 2011.

Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10 (2):1–37, 2019.

Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2018a.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5542–5551, 2018b.

Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. VGSE: Visually-grounded semantic embeddings for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9316–9325, 2022.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021.

Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021.

Li-Jun Zhao, Zhen-Duo Chen, Yongxin Wang, Xin Luo, and Xin-Shun Xu. Evolving and regularizing meta-environment learner for fine-grained few-shot class-incremental learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: A survey. *arXiv preprint arXiv:2401.16386*, 2024a.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024b.

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.

A Appendix Overview

This appendix supplements the main paper with theoretical analysis, implementation details, discussion, additional experiments, and the LLM-based semantic generation pipeline. The main paper presents IGICE as an intervention-guided framework that translates semantic causal interventions into geometric constraints in the classifier weight space. The appendix further explains the motivation, implementation, robustness, and limitations of this design.

Appendix B provides the theoretical analysis. It uses a Structural Causal Model to explain how dominant co-occurring attributes introduce a spurious shortcut path and lead to entangled classifier weights. It further provides a local geometric interpretation of the proposed invariance and sensitivity constraints, showing how they suppress weight displacement along observed co-occurring intervention directions and enlarge displacement along observed discriminative intervention directions under a first-order approximation.

Appendix C details the implementation of IGICE. Following the main text, it describes Pairwise Causal Disentanglement, Joint Causal Intervention, and Intervention-Guided Weight Refinement, including semantic construction, pairwise attribute partitioning, intervened semantic recomposition, weight generation, training objectives, and image-free inference.

Appendix D discusses key design choices and limitations. It clarifies that the improvement comes not merely from richer LLM-generated semantics, but from weight-space causal regularization. It also discusses pair-dependent attribute roles, incomplete disentanglement under attribute overlap, LLM reliability, computational overhead, and future directions.

Appendix E reports additional experiments, including fine-grained robustness analysis, LLM and embedding model compatibility, hyperparameter sensitivity, confusion and case analysis, causal disentanglement analysis, extended dataset comparisons, computational efficiency, LLM cost estimation, and robustness under simulated LLM reasoning errors.

Appendix F presents the LLM-based generation and causal disentanglement pipeline. It describes how class names are converted into visually grounded attribute descriptions, how pairwise discriminative and co-occurring attributes are identified, and how structured prompts are designed for reliable semantic generation and verification.

Together, these appendices support the central claim that fine-grained image-free classifier expansion benefits not only from richer semantic descriptions, but more importantly from intervention-guided disentanglement in the classifier weight space.

B Theoretical Analysis

This section provides a theoretical analysis of why fine-grained image-free classifier expansion suffers from weight entanglement and how the proposed intervention-guided objectives alleviate this problem. We first formulate the semantic-to-weight generation process from a causal perspective and explain how dominant co-occurring attributes introduce a spurious shortcut path. This path biases the generated classifier weights toward non-discriminative semantic factors, which causes geometrically entangled weights for visually similar classes. We then analyze the proposed intervention mechanism in the classifier weight space from a local geometric perspective. Specifically, under a first-order approximation of the semantic-to-weight generator, the invariance loss suppresses the generator response along observed co-occurring intervention directions, while the sensitivity loss enlarges the generator response along observed discriminative intervention directions. Finally, we connect this local geometric effect to fine-grained classification margins under an explicit alignment assumption between discriminative semantic directions and visual class evidence.

B.1 Structural Causal Model and Weight Entanglement

We begin by analyzing the source of weight entanglement in fine-grained image-free zero-shot learning. In the image-free setting, the model cannot access original visual samples. Instead, it learns a generator G that maps a textual semantic representation T to a visual classifier weight W. For a class i, the generated weight is written as

$$\hat{w}_i = G(T_i).$$

The key difficulty in fine-grained recognition is that the class semantics contain two types of attributes. The first type consists of discriminative attributes D, which determine the class identity in a specific comparison context. The second type consists of co-occurring attributes C, which frequently appear in the class description but do not distinguish visually similar classes.

We construct a Structural Causal Model consisting of four variables, namely the class identity Y, the discriminative attributes D, the co-occurring attributes C, and the generated classifier weight W. The textual description T is the observed semantic manifestation of the latent attribute set $\{D, C\}$. The class identity Y gives rise to both discriminative attributes and co-occurring attributes:

$$Y \to D, \qquad Y \to C.$$

Ideally, the semantic-to-weight generator should synthesize classifier weights according to the causal mechanism induced by discriminative attributes, namely $D \to W$. However, because co-occurring attributes are frequent, stable, and highly represented in textual descriptions, the generator may also exploit the shortcut $C \to W$. As a result, the association between discriminative attributes and generated weights can be affected by the following spurious shortcut path:

$$D \leftarrow Y \to C \to W. \tag{13}$$

This path creates a non-causal association between D and W through the shared class identity and the dominant co-occurring attributes. Therefore, the observed association between semantic descriptions and classifier weights does not necessarily reflect the causal attributes that define the fine-grained decision boundary.

This issue is particularly severe in fine-grained settings. Since similar classes share a large portion of attributes, co-occurring attributes usually dominate the textual representation in both frequency and magnitude. When the generator is trained with association-based objectives, such as reconstruction or semantic alignment, it is encouraged to fit the most stable and most frequent semantic signals. Consequently, the generator tends to assign excessive weight to C, even though C is not sufficient for distinguishing similar classes.

To illustrate this effect, consider a local decomposition of the generated classifier weight:

$$\hat{w}_i = G(T_i) = \alpha\phi_D(D_i) + \beta\phi_C(C_i) + \epsilon_i, \tag{14}$$

where $\phi_D(\cdot)$ and $\phi_C(\cdot)$ denote the semantic factors induced by discriminative and co-occurring attributes, respectively. The coefficients α and β measure the response strength of the generator to the two semantic factors, and ϵ_i denotes the residual term of this local decomposition. In fine-grained image-free classifier expansion, association-based optimization often leads to $\beta \gg \alpha$, because the co-occurring attributes provide a stronger and more stable shortcut for reconstruction.

Now consider two fine-grained classes i and j. They share similar co-occurring attributes but differ in subtle discriminative attributes:

$$C_i \approx C_j \approx C_{\text{shared}}, \qquad D_i \neq D_j.$$

Their generated weights can be approximated as

$$\hat{w}_i \approx \alpha\phi_D(D_i) + \beta\phi_C(C_{\text{shared}}),$$

and

$$\hat{w}_j \approx \alpha\phi_D(D_j) + \beta\phi_C(C_{\text{shared}}).$$

When $\beta \gg \alpha$, the shared co-occurring component dominates the geometry of both weights. Thus, the angular similarity between $\hat{w}_i$ and $\hat{w}_j$ becomes mainly determined by the shared term:

$$\cos(\hat{w}_i, \hat{w}_j) \approx \frac{\|\beta\phi_C(C_{\text{shared}})\|_2^2}{\|\hat{w}_i\|_2\|\hat{w}_j\|_2} \to 1. \tag{15}$$

This high cosine similarity indicates that the two classifier weights are geometrically entangled. Since the decision boundary between two linear classifiers is determined by the direction and distance between their weight vectors, such entanglement directly leads to an ambiguous fine-grained boundary. Therefore, the central theoretical goal is to reduce the contribution of co-occurring attributes while preserving the contribution of discriminative attributes in the generated weight space.

B.2 Intervention-Guided Local Weight-Space Disentanglement

We now analyze how the proposed causal intervention mechanism translates semantic interventions into geometric constraints on classifier weights. For a class pair (i, j), we denote the pair-specific discriminative attribute set as D_{ij} and the pair-specific co-occurring attribute set as C_{ij}. The pair-specific nature is important because an attribute can be discriminative for one class pair but co-occurring for another. Thus, causal disentanglement should not be defined globally over the whole dataset, but locally with respect to the current fine-grained decision boundary.

For clarity, we first describe the desired semantic effect of the two interventions. The co-occurring intervention replaces the co-occurring attributes of class i with those from class j while preserving the discriminative attributes of class i:

$$T^{co}_{i\leftarrow j} = (D_i, C_j).$$

Since the identity-discriminative attributes of class i are preserved, this intervention is expected to maintain the pairwise identity of class i. It therefore corresponds to a class-preserving contextual variation.

The discriminative intervention replaces the discriminative attributes of class i with those from class j while preserving the co-occurring context of class i:

$$T^{diff}_{i\leftarrow j} = (D_j, C_i).$$

Since the identity-defining attributes are changed, this intervention is expected to alter the pairwise identity of class i relative to class j. It therefore corresponds to an identity-shifting variation.

These two interventions define the desired property of weight-space causal disentanglement:

$$G(T_i) \approx G(T^{co}_{i\leftarrow j}), \qquad \text{dist}\left(G(T_i), G(T^{diff}_{i\leftarrow j})\right) \geq \rho, \tag{16}$$

where $\rho > 0$ denotes a desired separation margin. The first condition requires the generated weight to be invariant to non-discriminative co-occurring changes, while the second condition requires it to be sensitive to identity-changing discriminative variations.

We further analyze this effect through the local geometry of the generator. To avoid differentiating discrete textual descriptions, we conduct the analysis on the continuous semantic embeddings after shared-query semantic aggregation. Let h_i denote the aggregated semantic embedding of class i, and let $\hat{w}_i = G(h_i)$ be the generated classifier weight. For a class pair (i, j), the co-occurring intervention and the discriminative intervention induce two semantic displacement directions:

$$\delta^{co}_{i\leftarrow j} = h^{co}_{i\leftarrow j} - h_i, \qquad \delta^{diff}_{i\leftarrow j} = h^{diff}_{i\leftarrow j} - h_i.$$

Here, $\delta^{co}_{i\leftarrow j}$ corresponds to a class-preserving contextual variation, while $\delta^{diff}_{i\leftarrow j}$ corresponds to an identity-changing discriminative variation.

Assume that G is differentiable in a neighborhood of h_i. Under a first-order Taylor approximation, for a small semantic displacement δ, we have

$$G(h_i + \delta) - G(h_i) = J_i\delta + o(\|\delta\|_2), \tag{17}$$

where J_i is the local Jacobian of G with respect to the continuous semantic embedding h_i. Therefore, the co-occurring intervention induces the local weight displacement

$$\hat{w}^{co}_{i\leftarrow j} - \hat{w}_i = J_i\delta^{co}_{i\leftarrow j} + o(\|\delta^{co}_{i\leftarrow j}\|_2),$$

and the discriminative intervention induces

$$\hat{w}^{diff}_{i\leftarrow j} - \hat{w}_i = J_i\delta^{diff}_{i\leftarrow j} + o(\|\delta^{diff}_{i\leftarrow j}\|_2).$$

The invariance loss minimizes the distance between $\hat{w}^{co}_{i\leftarrow j}$ and $\hat{w}_i$. Under the above local approximation, it encourages

$$\left\|J_i\delta^{co}_{i\leftarrow j}\right\|_2 \tag{18}$$

to be small. This means that the generator becomes locally insensitive to the observed co-occurring intervention directions. Importantly, this does not imply that the full Frobenius norm of the Jacobian

is globally reduced. Rather, the constraint acts on the sampled class-preserving directions constructed by our interventions.

Conversely, the sensitivity loss encourages separation between $\hat{w}_{i\leftarrow j}^{diff}$ and $\hat{w}_i$. Under the same local approximation, it encourages

$$\left\| J_i \delta_{i\leftarrow j}^{diff} \right\|_2 \tag{19}$$

to be large. This means that the generator preserves or enlarges its local response to observed identity-changing discriminative directions. This should not be interpreted as a global increase of the minimum singular value of a discriminative subspace. Instead, the objective prevents discriminative semantic variations from being collapsed into nearly identical classifier weights.

In implementation, the geometric distances are computed by cosine-based losses. When generated classifier weights are ℓ_2-normalized, cosine distance is equivalent to squared Euclidean distance on the unit hypersphere:

$$1 - \cos(a, b) = \frac{1}{2} \left\| \frac{a}{\|a\|_2} - \frac{b}{\|b\|_2} \right\|_2^2.$$

Therefore, the above local Euclidean interpretation is consistent with the cosine-based implementation in the normalized weight space.

Overall, the proposed losses reshape the generator in a directional manner. The invariance loss compresses weight displacement along class-preserving co-occurring directions, while the sensitivity loss expands weight displacement along identity-changing discriminative directions. This explains how semantic interventions are translated into weight-space disentanglement without requiring a global guarantee over all possible semantic directions.

B.3 Connection to Fine-Grained Classification Margins

We finally connect the above local geometric analysis to fine-grained classification performance. For a visual feature z_i from class i, the pairwise classification margin between class i and class j is

$$\gamma_{ij} = \langle z_i, \hat{w}_i \rangle - \langle z_i, \hat{w}_j \rangle.$$

A positive and large γ_{ij} means that the classifier assigns a higher logit to the correct class i than to the confusing class j. Therefore, improving the pairwise margin is important for fine-grained recognition.

Using the local decomposition in Eq. (14), the difference between two generated classifier weights can be written as

$$\hat{w}_i - \hat{w}_j = \alpha \left[\phi_D(D_i) - \phi_D(D_j) \right] + \beta \left[\phi_C(C_i) - \phi_C(C_j) \right] + (\epsilon_i - \epsilon_j).$$

Substituting this into the pairwise margin gives

$$\gamma_{ij} = \alpha \left\langle z_i, \phi_D(D_i) - \phi_D(D_j) \right\rangle + \beta \left\langle z_i, \phi_C(C_i) - \phi_C(C_j) \right\rangle + \left\langle z_i, \epsilon_i - \epsilon_j \right\rangle.$$

The first term is the desired discriminative margin, because it is induced by identity-defining attributes. The second term is the co-occurring leakage, because it is induced by non-discriminative attributes. In fine-grained settings, this leakage term is unstable. Since C_i and C_j are highly similar and not aligned with the true decision boundary, the co-occurring component can either contribute little useful separation or distort the boundary direction.

By the Cauchy-Schwarz inequality, the pairwise margin can be lower bounded as

$$\gamma_{ij} \geq \alpha \left\langle z_i, \phi_D(D_i) - \phi_D(D_j) \right\rangle - |\beta| \|z_i\|_2 \left\| \phi_C(C_i) - \phi_C(C_j) \right\|_2 - \|z_i\|_2 \|\epsilon_i - \epsilon_j\|_2. \tag{20}$$

This bound shows that the effective fine-grained margin is determined by two competing factors. The discriminative term increases the margin when it is positively aligned with the visual evidence of class i, while the co-occurring leakage and approximation residual reduce the guaranteed margin.

The connection between the proposed losses and the pairwise margin relies on a local alignment assumption. Specifically, the discriminative semantic displacement induced by $\delta_{i\leftarrow j}^{diff}$ should be positively aligned with the visual evidence of the corresponding class, while the co-occurring displacement induced by $\delta_{i\leftarrow j}^{co}$ should not provide stable class-specific evidence. This assumption is

consistent with the purpose of the baseline semantic-to-visual alignment objective, which trains the generator to map semantic differences into visual classifier differences.

Under this alignment assumption, suppressing the generator response along co-occurring directions reduces unstable boundary perturbations caused by non-discriminative attributes. Meanwhile, enlarging the response along discriminative directions increases the component of the weight difference that is useful for distinguishing visually similar classes. Therefore, the proposed losses do not provide an unconditional guarantee that the margin increases for every sample. Rather, they provide a local geometric explanation for why intervention-guided weight refinement tends to improve fine-grained decision boundaries when the generated discriminative semantic directions are aligned with visual class evidence.

In summary, the proposed method can be interpreted as performing local causal disentanglement directly in the classifier weight space. The co-occurring intervention identifies directions that should not affect class identity and compresses them through invariance regularization. The discriminative intervention identifies directions that should affect class identity and expands them through sensitivity regularization. This intervention-guided geometric reshaping reduces the influence of co-occurring shortcut directions and encourages the generated weight space to rely more on discriminative semantic directions. The analysis should be understood as a local geometric interpretation of the proposed objectives rather than a global identifiability guarantee.

C Method Details

This section provides supplementary implementation details of the proposed framework. Following the structure of the main text, we organize the method into three parts: Pairwise Causal Disentanglement, Joint Causal Intervention, and Intervention-Guided Weight Refinement. The first module constructs pair-specific discriminative and co-occurring attribute sets. The second module recomposes intervened semantic representations by manipulating these attribute sets. The third module maps factual and intervened semantics into the classifier weight space and optimizes the generator with both baseline reconstruction losses and intervention-guided geometric constraints.

C.1 Pairwise Causal Disentanglement

To make causal intervention operable in fine-grained image-free classifier expansion, we first construct structured class semantics at the attribute level. Let $\mathcal{S}$ and $\mathcal{U}$ denote the seen and unseen class sets. We define a shared attribute schema

$$\mathcal{A} = \{a_1, a_2, \ldots, a_K\},$$

where each a_k corresponds to a visually meaningful attribute dimension. For each class $i \in \mathcal{S} \cup \mathcal{U}$, an LLM generates an attribute-wise semantic description

$$T_i = \{T_{i,1}, T_{i,2}, \ldots, T_{i,K}\},$$

where $T_{i,k}$ denotes the textual description of class i under the attribute a_k. In addition, a global description T_i^{global} is generated to provide a holistic class-level semantic anchor.

A frozen text encoder $\mathcal{F}_{text}$ is used to encode each attribute description independently:

$$E_{i,k} = \mathcal{F}_{text}(T_{i,k}), \qquad k = 1, \ldots, K.$$

The encoded attributes are stacked into an attribute-level semantic matrix:

$$E_i = [E_{i,1}, E_{i,2}, \ldots, E_{i,K}]^\top \in \mathbb{R}^{K \times d_T},$$

where d_T denotes the text embedding dimension. The global description is encoded as

$$E_i^{global} = \mathcal{F}_{text}(T_i^{global}) \in \mathbb{R}^{d_T}.$$

Thus, each class is represented by both the structured attribute matrix E_i and the global semantic vector E_i^{global}.

The key motivation of pairwise causal disentanglement is that attribute discriminability is not absolute, but depends on the comparison context. An attribute may be discriminative for one class pair but

co-occurring for another. Therefore, for each class pair (i, j), we use an LLM evaluator Ψ_{LLM} to determine whether the k-th attribute provides an explicit visual distinction between the two classes:

$$m_{i,j,k} = \Psi_{\text{LLM}}(T_{i,k}, T_{j,k}, a_k \mid \mathcal{I}),$$

where $\mathcal{I}$ denotes the instruction that specifies the criterion for visual discriminability. If $m_{i,j,k} = 1$, the attribute is considered discriminative for the pair (i, j). Otherwise, it is considered co-occurring or non-discriminative for this pairwise boundary.

Based on this indicator, we partition the shared attribute schema into a pair-specific discriminative set and a co-occurring set:

$$\mathcal{D}_{ij} = \{a_k \in \mathcal{A} \mid m_{i,j,k} = 1\}, \qquad \mathcal{C}_{ij} = \{a_k \in \mathcal{A} \mid m_{i,j,k} = 0\}. \tag{21}$$

These two sets satisfy

$$\mathcal{D}_{ij} \cup \mathcal{C}_{ij} = \mathcal{A}, \qquad \mathcal{D}_{ij} \cap \mathcal{C}_{ij} = \varnothing.$$

In practice, pairwise disentanglement can be restricted to the most semantically similar classes, since these pairs are most likely to suffer from co-occurring attribute interference. For each class i, we define its neighbor set as

$$\mathcal{N}_i = \text{Top-}n \left(\left\{ \cos(E_i^{global}, E_j^{global}) \mid j \in \mathcal{S} \cup \mathcal{U}, j \neq i \right\} \right).$$

The pair set used for intervention is then

$$\mathcal{P} = \{(i, j) \mid i \in \mathcal{S} \cup \mathcal{U}, j \in \mathcal{N}_i\}.$$

This strategy focuses causal disentanglement on the fine-grained boundaries where semantic overlap is most severe.

C.2 Joint Causal Intervention

After obtaining pair-specific attribute partitions, we construct intervened semantics by explicitly manipulating the attribute-level semantic matrices. For each pair $(i, j) \in \mathcal{P}$, we build two complementary interventions: a co-occurring intervention and a discriminative intervention.

The co-occurring intervention replaces the co-occurring attributes of class i with those of class j, while keeping the discriminative attributes of class i unchanged. The resulting semantic matrix $E^{co}_{i \leftarrow j}$ is defined as

$$E^{co}_{i \leftarrow j,k} = \begin{cases} E_{j,k}, & a_k \in \mathcal{C}_{ij}, \\ E_{i,k}, & a_k \in \mathcal{D}_{ij}. \end{cases} \tag{22}$$

Equivalently, this operation can be viewed as a do-intervention,

$$E^{co}_{i \leftarrow j} = E_i\big(do(E_{i,k} = E_{j,k} \mid a_k \in \mathcal{C}_{ij})\big).$$

Since the discriminative attributes are preserved, this intervention keeps the class identity of i unchanged. Therefore, it simulates an identity-preserving contextual variation.

The discriminative intervention replaces the discriminative attributes of class i with those of class j, while keeping the co-occurring context of class i unchanged:

$$E^{diff}_{i \leftarrow j,k} = \begin{cases} E_{j,k}, & a_k \in \mathcal{D}_{ij}, \\ E_{i,k}, & a_k \in \mathcal{C}_{ij}. \end{cases} \tag{23}$$

Equivalently,

$$E^{diff}_{i \leftarrow j} = E_i\big(do(E_{i,k} = E_{j,k} \mid a_k \in \mathcal{D}_{ij})\big).$$

Since the identity-defining attributes are changed, this intervention simulates an identity-shifting semantic variation. It should induce a clear displacement in the generated classifier weight.

The factual and intervened semantic matrices are then aggregated into class-level semantic embeddings through a shared-query aggregation strategy. For class i, the global semantic vector E_i^{global} is projected into H query heads:

$$\{Q_{i,1}, Q_{i,2}, \ldots, Q_{i,H}\} = \mathcal{P}_Q(E_i^{global}),$$

where $\mathcal{P}_Q$ is a learnable projection network. Given a semantic matrix E, each query head aggregates attribute information as

$$\mathrm{Attn}(Q_{i,h}, E) = \mathrm{softmax}\left(\frac{Q_{i,h}W_Q(EW_K)^\top}{\sqrt{d_a}}\right)EW_V,$$

where W_Q, W_K, and W_V are learnable projection matrices, and d_a is the attention dimension.

Using the same query heads, we aggregate the factual semantics and the two intervened semantics:

$$\begin{aligned} h_i &= \mathrm{Concat}_{h=1}^{H}\,\mathrm{Attn}(Q_{i,h}, E_i), \\ h_{i\leftarrow j}^{co} &= \mathrm{Concat}_{h=1}^{H}\,\mathrm{Attn}(Q_{i,h}, E_{i\leftarrow j}^{co}), \\ h_{i\leftarrow j}^{diff} &= \mathrm{Concat}_{h=1}^{H}\,\mathrm{Attn}(Q_{i,h}, E_{i\leftarrow j}^{diff}). \end{aligned} \tag{24}$$

The shared-query design ensures that factual and intervened semantics are read under the same aggregation policy. This makes the subsequent weight-space comparison meaningful, because the generated factual and intervened weights are produced from semantically aligned representations.

C.3 Intervention-Guided Weight Refinement

The final module maps factual and intervened semantic embeddings into the classifier weight space and optimizes the generator with both baseline generation losses and intervention-guided geometric constraints.

Following existing image-free classifier expansion methods, we use a dual-autoencoder architecture as the backbone. The semantic stream contains a semantic encoder $\mathcal{E}_T$ and a semantic decoder $\mathcal{D}_T$, while the visual stream contains a visual encoder $\mathcal{E}_V$ and a visual decoder $\mathcal{D}_V$. The semantic and visual streams are connected through a shared latent space. The weight generator $\mathcal{G}$ is instantiated as the cross-modal pathway from the semantic encoder to the visual decoder:

$$\mathcal{G}(\cdot) = \mathcal{D}_V(\mathcal{E}_T(\cdot)). \tag{25}$$

Thus, the factual and intervened weights are generated as

$$\hat{w}_i = \mathcal{D}_V(\mathcal{E}_T(h_i)), \quad \hat{w}_{i\leftarrow j}^{co} = \mathcal{D}_V(\mathcal{E}_T(h_{i\leftarrow j}^{co})), \quad \hat{w}_{i\leftarrow j}^{diff} = \mathcal{D}_V(\mathcal{E}_T(h_{i\leftarrow j}^{diff})).$$

The baseline objective provides the fundamental semantic-to-weight generation ability. For each class i, we define the augmented semantic set

$$\mathbb{H}_i = \{h_i\} \cup \{h_{i\leftarrow j}^{co} \mid j \in \mathcal{N}_i\} \cup \{h_{i\leftarrow j}^{diff} \mid j \in \mathcal{N}_i\}.$$

The semantic reconstruction loss is defined as

$$\mathcal{L}_{rec}^{S\to S} = \mathbb{E}_{i\in\mathcal{S}\cup\mathcal{U}}\left[\sum_{h\in\mathbb{H}_i} d_S(\mathcal{D}_T(\mathcal{E}_T(h)), h)\right].$$

The visual reconstruction loss is defined on seen classes:

$$\mathcal{L}_{rec}^{V\to V} = \mathbb{E}_{i\in\mathcal{S}}\left[d_V(\mathcal{D}_V(\mathcal{E}_V(w_i)), w_i)\right].$$

To align the semantic and visual modalities, we further use the forward semantic-to-visual alignment loss

$$\mathcal{L}_{align}^{S\to V} = \mathbb{E}_{i\in\mathcal{S}}\left[d_V(\mathcal{D}_V(\mathcal{E}_T(h_i)), w_i)\right],$$

and the backward visual-to-semantic consistency loss

$$\mathcal{L}_{align}^{V\to S} = \mathbb{E}_{i\in\mathcal{S}}\left[d_S(\mathcal{D}_T(\mathcal{E}_V(w_i)), h_i)\right].$$

The complete baseline loss is

$$\mathcal{L}_{base} = \mathcal{L}_{rec}^{S\to S} + \mathcal{L}_{rec}^{V\to V} + \mathcal{L}_{align}^{S\to V} + \mathcal{L}_{align}^{V\to S}. \tag{26}$$

Beyond the baseline objective, intervention-guided losses are imposed directly in the classifier weight space. For the co-occurring intervention, the class identity should remain unchanged. Therefore, $\hat{w}_{i\leftarrow j}^{co}$ is treated as a positive sample relative to $\hat{w}_i$. The invariance loss is

$$\mathcal{L}_{inv} = \mathbb{E}_{(i,j)\in\mathcal{P}}\left[d_{inv}(\hat{w}_i, \hat{w}_{i\leftarrow j}^{co})\right]. \tag{27}$$

Here, d_{inv} is implemented as a cosine-distance loss. Minimizing this loss suppresses the weight displacement caused by non-discriminative co-occurring variations.

For the discriminative intervention, the class identity should change. Therefore, $\hat{w}_{i\leftarrow j}^{diff}$ is treated as a negative sample relative to $\hat{w}_i$. The sensitivity loss is

$$\mathcal{L}_{sen} = \mathbb{E}_{(i,j)\in\mathcal{P}}\left[d_{sen}(\hat{w}_i, \hat{w}_{i\leftarrow j}^{diff})\right]. \tag{28}$$

Here, d_{sen} is implemented as a cosine-based loss that penalizes high similarity between the factual weight and the discriminative-intervened weight. Minimizing $\mathcal{L}_{sen}$ therefore decreases their cosine similarity and increases their angular separation.

The final training objective is

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda_{inv}\mathcal{L}_{inv} + \lambda_{sen}\mathcal{L}_{sen}, \tag{29}$$

where λ_{inv} and λ_{sen} control the strength of the two intervention-guided constraints.

During inference, no intervened semantics or visual samples are required. For each unseen class $u \in \mathcal{U}$, we construct its factual semantic embedding h_u and synthesize the corresponding classifier weight:

$$\hat{w}_u = \mathcal{D}_V(\mathcal{E}_T(h_u)).$$

The generated unseen weights are concatenated with the frozen seen classifier weights:

$$W_{total} = [W_{\mathcal{S}}; \{\hat{w}_u\}_{u\in\mathcal{U}}].$$

The final classifier performs prediction over the joint label space $\mathcal{S} \cup \mathcal{U}$, while strictly following the image-free setting.

D Discussion

Weight-space causal intervention and source of improvement. The proposed framework should not be interpreted as merely using an LLM to generate richer semantic descriptions. Enhanced semantics can improve the input quality, but they do not by themselves prevent dominant co-occurring attributes from being mapped into classifier weights. In fine-grained image-free classifier expansion, visually similar classes often share many semantic attributes, and association-based optimization tends to preserve these shared factors in the generated weights, leading to ambiguous decision boundaries. Our main contribution is to convert semantic interventions into geometric constraints in the classifier weight space. Co-occurring interventions are expected to preserve class identity and thus induce invariant weights, while discriminative interventions are expected to change class identity and thus induce separable weights. This mechanism differs from existing causal ZSL methods that mainly focus on semantic debiasing, visual feature adjustment, or image-level counterfactual reasoning. In our strict image-free setting, the central object is the generated classifier weight itself, and the improvement comes from reshaping its geometry rather than from semantic enrichment alone.

Pairwise attribute roles and incomplete disentanglement. Our method does not assume that co-occurring attributes are universally noisy or should be globally removed. The role of an attribute is defined relative to a specific class pair. For example, wing-related attributes may help distinguish a bird from a mammal, but may provide little discriminative information when comparing two similar bird species. Therefore, globally assigning each attribute as either discriminative or co-occurring can be misleading. Pairwise Causal Disentanglement addresses this issue by identifying discriminative and co-occurring attributes under each local comparison context. This design is also important because attribute overlap is inherent in fine-grained recognition. Many classes share most visible properties and differ only in a few subtle cues, making perfect global disentanglement unrealistic. Instead, our pairwise formulation focuses on the attributes that are most informative for the current confusing boundary, which provides more precise intervention targets for refining the relative geometry between class weights.

Reliability of LLM-based disentanglement. The framework does not require the LLM to provide perfectly correct descriptions or perfectly accurate attribute partitions. The LLM is used as a semantic reasoning engine to construct candidate descriptions and pairwise intervention targets, while the final classifier weights are learned through optimization under intervention-guided constraints. Structured

prompting, verification, refinement, and comparative reasoning are used to reduce noisy semantic partitions. Moreover, the method relies on the relative distinction between discriminative and co-occurring attributes rather than on absolute semantic annotation. Incorrect causal attribution can still weaken the intervention signal, but the framework is designed to tolerate moderate partition errors, since the learned generator aggregates intervention evidence across many class pairs rather than depending on a single comparison. This makes the method more robust than directly trusting unverified LLM-generated semantics.

Computational overhead and deployment. The additional cost of the proposed method mainly appears during preprocessing and training. Pairwise semantic comparison and intervention construction require extra computation, and the invariance and sensitivity losses introduce additional training-time optimization terms. However, these operations are not required during inference. Once the generator is trained, unseen classifier weights are synthesized through the same semantic-to-weight pathway as existing image-free methods. Therefore, the causal reasoning process is effectively distilled into the learned generator and the resulting classifier weights. This makes the framework practical for deployment, since it improves fine-grained separability without introducing additional inference-time LLM calls or requiring original visual data.

Limitations and future directions. Several limitations remain. First, severe LLM hallucinations or systematic visual misconceptions may still produce inaccurate intervention targets, especially for classes whose differences require expert-level visual knowledge. Second, attribute-level language descriptions may be insufficient for distinctions that depend on highly localized or subtle visual patterns. Third, the current framework uses deterministic generated weights and does not explicitly model uncertainty in semantic partitioning or weight synthesis. Future work could incorporate confidence-aware causal constraints, probabilistic weight generation, human-verifiable attribute correction, or multimodal prior knowledge to further improve reliability. More broadly, this work suggests that fine-grained image-free classifier expansion should not only ask how to obtain better class semantics, but also how generated classifier weights should respond to controlled semantic changes. This weight-space causal perspective may benefit other image-free or data-restricted learning settings where semantic information is available but visual data cannot be accessed.

Broader impacts. This work may have positive impacts in privacy-sensitive and data-restricted scenarios, since it enables classifier expansion without requiring access to original visual data. By reducing the dependence on image collection and storage, it can lower deployment costs and help build recognition systems under limited data availability. However, the same capability may also be misused to extend recognition systems in sensitive or inappropriate applications, such as surveillance or fine-grained classification of protected attributes. In addition, LLM-generated semantic descriptions may contain biases or visual misconceptions, which could be transferred into classifier weights if used without verification. Therefore, practical deployment should be restricted to appropriate domains, accompanied by human review when needed, and evaluated carefully for fairness, privacy, and misuse risks.

E Additional Experiments

E.1 Experimental setup

Datasets. To comprehensively evaluate our method, we conduct experiments on five widely used ZSL benchmarks. For fine-grained classification, we employ five datasets: CUB with 200 bird species, Oxford Flowers (FLO) with 102 flower classes and SUN covering 717 scene classes. These datasets are all characterized by high inter-class visual similarity. For coarse-grained classification, we utilize AWA2 with 50 animal classes and aPY with 32 general object classes. Following an image-free setting, we did not use any images from these datasets.

Baselines. To validate the effectiveness of our method, we compare it against a comprehensive set of representative Image-Free ZSL methods, all operating under the strict image-free setting. The first class comprises classical methods: ConSE [Norouzi et al., 2014], which utilizes convex combinations of semantic embeddings from seen classes to approximate unseen representations; and COSTA [Mensink et al., 2014], which leverages co-occurrence statistics to estimate classifier parameters. The second class consists of advanced optimization methods: SubReg [Akyürek et al., 2021] applies subspace regularization to constrain the generated weights of unseen classes, ensuring they lie within the manifold defined by seen classes; and VGSE [Xu et al., 2022] aligns visual and

semantic spaces via visual clustering to produce visually-grounded semantic embeddings. Finally, we compare with state-of-the-art methods: ICIS [Christensen et al., 2023] employs a dual-autoencoder architecture to synthesize visual classifier weights from semantic descriptions, effectively bridging the modality gap; and the recent LLM-driven CEMIL [Lyu et al., 2025] method generates multi-view descriptions to enrich the semantic space for classifier expansion.

LLMs. Our framework utilizes Large Language Models to generate semantic descriptions and causal disentanglement. To comprehensively evaluate the method's applicability, we tested two classes of LLMs. For closed-source models, we selected GPT-4o [Hurst et al., 2024], Gemini-2.5 [Comanici et al., 2025] and Claude-4 which represent the high-performance frontier in reasoning and instruction following. For open-source models, we employed LLaMA-3.3 [Dubey et al., 2024] , DeepSeek-V3 [Liu et al., 2024] and Qwen-2.5 [Yang et al., 2024], offering cost-effective and deployable alternatives.

Embedding Models. To transform the generated textual descriptions into semantic vectors, we evaluate three distinct types of embedding models. First, for vision-language alignment, we primarily utilize CLIP (ViT-B/32), which aligns image and text features via contrastive learning [Radford et al., 2021]. Second, as a pure text encoder, we select SBERT, which is optimized for sentence-level semantic similarity [Reimers and Gurevych, 2019]. Finally, we explore LLM-based embeddings by extracting the mean-pooled features from the last hidden layer of LLaMA and Qwen. This setup allows us to verify whether our method adapts to feature spaces with varying semantic structures and dimensionalities.

Implementation Details. The base classifier weights for the seen classes are sourced from the final classification layer of a pre-trained ResNet101 model. For semantic inputs, we leverage LLMs to generate semantic attributes for each class, controlling the ratio of pairwise causal interventions within 10% to balance feature diversity and computational efficiency. Regarding network configuration, the weight generator is constructed as a symmetric dual-stream autoencoder, with linear layer weights initialized using a Gaussian distribution. The shared-query semantic aggregation module uses $H = 128$ query heads with attention dimension $d_a = 1024$. During training, we employ the Adam optimizer with a learning rate of 10^{-4} and exponential decay rates $(\beta_1, \beta_2) = (0.9, 0.999)$. The balancing coefficients for the causal constraints are set to $\lambda_{inv} = 0.1$ and $\lambda_{sen} = 0.1$. The maximum training epochs are set to 500, coupled with an early stopping strategy to prevent overfitting. All experiments are executed on a single NVIDIA RTX 4090 GPU with 32GB of memory.

E.2 Confusion & Case Analysis

This section analyzes how the proposed method alleviates fine-grained confusion at the class level. We provide two complementary visual analyses: confusion matrices for examining the overall prediction distribution, and case-level probability comparisons for showing how the method refines prediction confidence on challenging samples.

Matrix Analysis. To analyze how the model handles fine-grained confusion, we compute confusion matrices for CEMIL and our method on CUB, SUN, and AWA2. The evaluation settings are selected to examine robustness under different scenarios. For CUB and SUN, we restrict the prediction scope to unseen classes under the ZSL setting, which directly evaluates discrimination among unseen fine-grained classes. For AWA2, we visualize confusion among all classes under the GZSL setting, which reflects the global decision boundary between seen and unseen classes. In each matrix, rows correspond to ground-truth labels and columns correspond to predicted labels. The values are normalized for each ground-truth class to show how samples from the same true class are distributed across predicted classes.

As shown in Figure 5, the CEMIL heatmaps exhibit evident local block-like confusion patterns, indicating that the baseline tends to confuse semantically similar classes. In contrast, our method produces sharper diagonal structures and lower off-diagonal responses. This reduction in off-diagonal confusion suggests that the proposed intervention-guided refinement improves class separability and suppresses interference from visually similar candidates.

Case-Level Probability Analysis. To further examine the behavior of our method on challenging samples, we conduct a fine-grained probability analysis on the CUB dataset. We select eight hard cases and compare the prediction probability distributions of the baseline and our method in Figure 6. Instead of only reporting whether the final prediction is correct, this analysis focuses on the confidence assigned to the ground-truth class and the probabilities assigned to confusing classes.

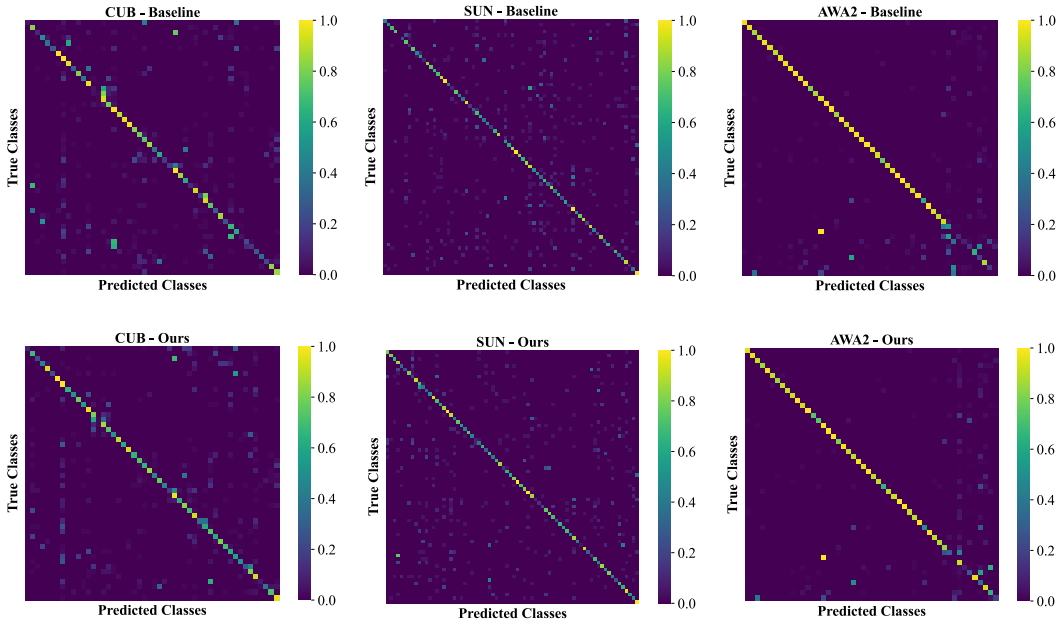


Figure 5: Visualization of confusion matrices. The top row shows the CEMIL results, and the bottom row shows our method across CUB, SUN, and AWA2.

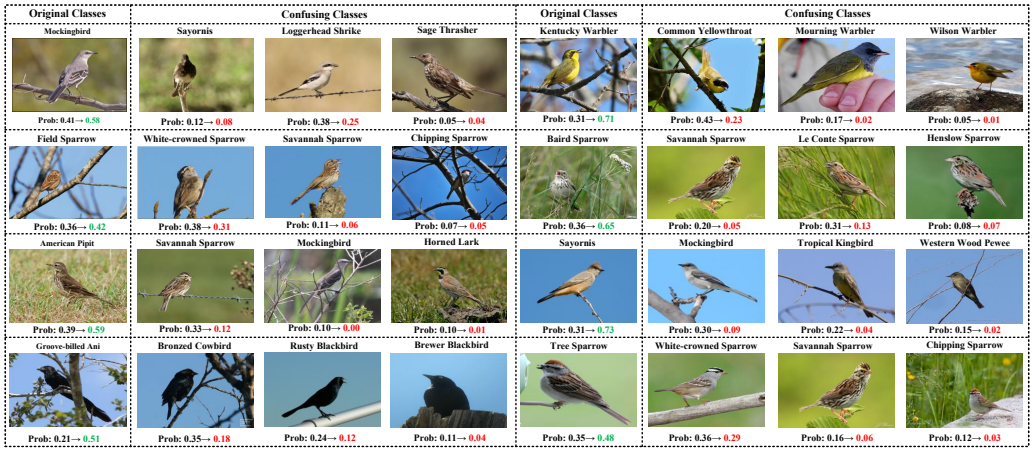


Figure 6: Qualitative analysis of probability refinement on hard cases. We visualize eight challenging samples to compare prediction confidence between the baseline and our method. For each case, the left panel highlights the probability increase for the ground-truth class, while the right panel shows the suppression of top confusing classes. The results indicate that our method improves confidence margins by strengthening discriminative cues and reducing interference from visually similar or co-occurring classes.

The comparisons in Figure 6 show that the baseline often assigns relatively low confidence to the ground-truth class and high probabilities to visually similar confusing classes. In contrast, our method increases the probability of the ground-truth class while suppressing the probabilities of confusing alternatives. This suggests that the proposed intervention-guided refinement mitigates co-occurrence interference by strengthening discriminative cues, thereby improving prediction confidence and enlarging the margin between the ground-truth class and confusing classes.

E.3 Fine-Grained Robustness Analysis

This experiment aims to verify the anti-interference capability of our method in high-density semantic environments. In fine-grained datasets, as the number of semantically similar classes increases, the interference from co-occurring attributes increase rapidly, leading to a performance collapse for models unable to disentangle attributes. To simulate this scenario, we designed a progressive expansion protocol on the CUB and SUN datasets. Specifically, we vary the number of candidate unseen classes from 20% to 100% of the total set. At each expansion step, we randomly sample the specified proportion of classes and evaluate both the Top-1 Accuracy (T) and Harmonic Mean (H). To eliminate bias from randomness, we repeat the sampling and testing process 5 times at each ratio and report the average performance.

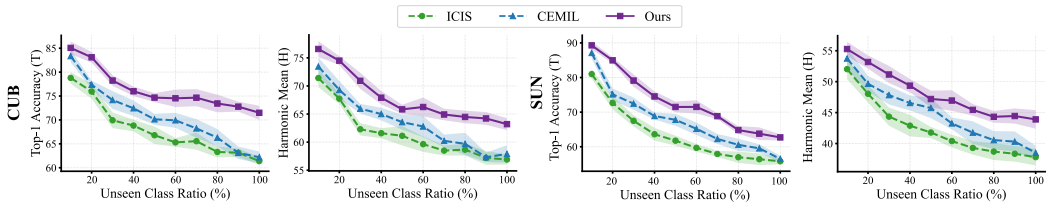


Figure 7: Comparative harmonic mean and top-1 accuracy of the baseline (ICIS, CEMIL) and our method across proportions of expanded classes.

The experimental results are illustrated in Figure 7. We observe a significant performance divergence: as the number of unseen classes increases, baseline methods exhibit a sharp decline in accuracy, indicating that their generated weights suffer from severe geometric entanglement dominated by co-occurring features. In contrast, our method demonstrates exceptional stability with a gentle decline curve. Notably, under the extremely crowded condition where all unseen classes are included (100%), the performance gap between our method and the baselines reaches its maximum. This result proves that, through the causal intervention mechanism, our model successfully mitigates co-occurrence noise, maintaining clear decision boundaries even in highly congested fine-grained semantic spaces.

E.4 Stability Analysis

To verify the versatility and robustness of our framework, we conducted a broad series of substitution and sensitivity experiments.

Compatibility with LLMs and Encoders. To examine whether our framework depends on a specific semantic generator or text encoder, we conduct extensive substitution experiments across the CUB, SUN, and AWA2 datasets, reporting both Top-1 Accuracy and Harmonic Mean. Specifically, we vary the semantic source by using closed-source LLMs, including GPT-4o, Gemini-2.5, and Claude-4, as well as open-source models, including LLaMA-3.3, DeepSeek-V3, and Qwen-2.5. In addition, we evaluate four distinct embedding architectures, ranging from the vision-aligned CLIP text encoder and the pure-text SBERT encoder to LLM-based embeddings derived from LLaMA and Qwen hidden states. This experiment is designed to assess whether the proposed intervention-guided refinement remains effective when the semantic descriptions and their embedding representations are replaced by different model families.

The results are summarized in Table 4. The default configuration, GPT-4o with CLIP, achieves the strongest overall performance on most datasets and metrics, indicating that high-quality visual semantics and vision-aligned text embeddings are beneficial for image-free classifier expansion. Nevertheless, the alternative LLM and encoder combinations still maintain competitive performance across datasets. For example, closed-source LLMs such as Gemini-2.5 and Claude-4 achieve results close to GPT-4o, while open-source models such as LLaMA-3.3, DeepSeek-V3, and Qwen-2.5 also preserve reasonable performance despite using weaker or more domain-general semantic priors. Similarly, although CLIP generally provides the strongest embedding space due to its vision-language alignment, SBERT and LLM-based embeddings still yield stable results.

These observations suggest that the proposed framework is not tied to a particular semantic generator or text encoder. The absolute performance varies with the semantic quality of the LLM and the visual alignment of the embedding model, but the intervention-guided refinement remains effective under

Table 4: Top-1 Accuracy and Harmonic Mean of our method under different LLM and embedding model configurations across multiple datasets.

Dataset	LLM	ZSL (Top-1 Accuracy)				GZSL (Harmonic Mean)			
		CLIP	SBERT	LLaMA	Qwen	CLIP	SBERT	LLaMA	Qwen
CUB	GPT-4o	**71.5±1.1**	**66.5±1.8**	**65.8±2.5**	**66.8±2.1**	**63.2±0.9**	58.4±1.3	58.8±1.6	59.2±1.6
	Gemini-2.5	71.2±1.3	66.1±2.0	65.4±1.5	66.5±1.8	62.8±1.1	**59.1±1.5**	**58.8±1.2**	**59.5±1.4**
	Claude-4	70.9±1.5	65.8±1.9	64.9±2.1	66.1±1.2	62.5±1.4	58.2±1.1	57.6±1.4	59.0±1.0
	LLaMA-3.3	67.8±1.9	63.5±1.5	63.8±2.1	64.4±1.8	60.5±1.4	57.4±0.9	57.4±0.9	57.8±1.2
	DeepSeek-V3	67.5±1.6	63.8±2.0	64.2±1.8	63.9±1.5	59.8±1.3	57.5±1.1	58.0±1.0	57.8±1.4
	Qwen-2.5	66.9±1.1	62.7±2.5	63.6±2.5	63.2±1.4	58.9±0.8	56.0±0.9	57.4±1.1	57.4±0.8
SUN	GPT-4o	**62.7±0.8**	58.8±1.4	57.9±1.3	57.5±0.5	**43.9±0.7**	39.9±1.3	38.5±1.7	**39.5±1.1**
	Gemini-2.5	62.4±0.5	**59.2±1.1**	**58.5±1.0**	**58.1±0.8**	43.5±1.4	**40.2±1.9**	**39.6±2.0**	38.8±1.8
	Claude-4	62.1±0.9	58.6±0.8	57.8±1.2	57.9±0.9	43.1±2.1	39.5±1.7	38.2±1.5	38.5±1.5
	LLaMA-3.3	59.5±0.8	56.8±0.4	56.5±1.6	56.2±0.5	41.2±2.3	37.4±2.0	37.3±1.4	36.9±0.4
	DeepSeek-V3	59.2±0.7	56.5±0.9	56.2±1.1	56.0±0.6	40.8±1.9	37.1±1.6	36.8±1.3	37.0±1.7
	Qwen-2.5	58.8±0.9	56.1±0.8	55.6±1.3	55.8±0.9	40.1±2.5	36.5±2.3	36.3±1.6	36.2±2.5
AWA2	GPT-4o	**74.9±1.5**	**70.8±1.6**	**70.5±3.0**	**70.9±1.9**	**66.7±2.1**	**62.9±2.6**	62.5±2.3	**62.8±6.1**
	Gemini-2.5	74.5±2.8	70.2±2.3	70.1±1.8	70.5±3.1	66.2±5.0	62.5±2.8	**62.8±2.8**	62.2±4.1
	Claude-4	74.1±2.9	69.8±2.9	69.5±2.1	70.1±2.7	65.9±4.9	62.0±2.6	61.4±2.6	61.8±7.2
	LLaMA-3.3	71.5±2.3	69.2±1.6	69.0±2.1	69.4±1.8	62.8±5.1	61.5±2.7	61.6±3.3	61.5±5.2
	DeepSeek-V3	71.2±2.0	69.5±1.9	68.9±2.2	69.1±1.6	62.5±4.5	61.1±2.5	60.9±3.0	61.2±5.5
	Qwen-2.5	70.8±3.4	69.0±2.9	68.7±2.1	69.6±2.4	62.1±3.5	60.6±3.3	60.4±2.7	61.9±5.7

diverse semantic configurations. Therefore, this experiment should be understood as an evaluation of compatibility and robustness under semantic-source substitution, rather than as a baseline comparison under every possible LLM and embedding configuration.

Hyperparameter Sensitivity Analysis. We investigated the impact of two key hyperparameters, namely the invariance loss weight λ_{inv} and the sensitivity loss weight λ_{sen}, across the CUB, SUN, and AWA2 datasets. Unless otherwise specified, we use $\lambda_{inv} = 0.1$ and $\lambda_{sen} = 0.1$ for all main experiments. To analyze robustness to these causal loss weights, we varied both parameters from 10^{-3} to 10^{2} and visualized the results using matrix heatmaps, where the horizontal axis represents λ_{inv} and the vertical axis denotes λ_{sen}. Each grid cell corresponds to a specific parameter combination, and the value represents the achieved Top-1 Accuracy or Harmonic Mean.

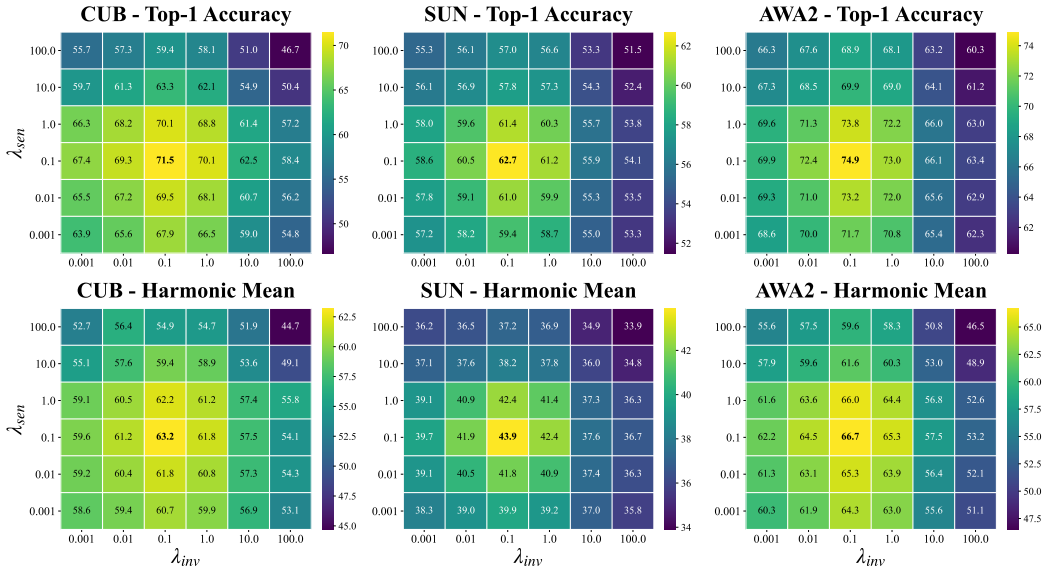


Figure 8: Hyperparameter sensitivity analysis of the proposed method with different invariance and sensitivity loss weights. The default setting used in the main experiments is $\lambda_{inv} = 0.1$ and $\lambda_{sen} = 0.1$.

As illustrated in Figure 8, model performance remains stable across a broad region of the parameter space. Specifically, we observe a high-performance plateau where λ_{inv} falls within the interval from

0.1 to 1 and λ_{sen} lies between 0.01 and 1. Since the default setting $(\lambda_{inv}, \lambda_{sen}) = (0.1, 0.1)$ is located within this plateau, the results indicate that our method is not overly sensitive to fine-grained parameter tuning.

E.5 Causal Disentanglement Analysis

This section investigates the impact of the LLM-based causal disentanglement pipeline on model performance. We systematically verify the approach across three dimensions: the selection of neighbor quantity, the accuracy of the disentanglement strategy, and the semantic quality of the disentangled content.

Impact of Neighbor Quantity. The number of pairwise comparisons serves as a pivotal hyperparameter governing the trade-off between disentanglement precision and computational overhead. To identify the optimal configuration, we formulated the neighbor quantity as a ratio of the total class space, evaluating the set 5% to 100% on the CUB dataset. To distinguish whether the optimal parameter depends on specific data distributions, we conducted a rigorous validation across varying data scales. By adjusting the seen class rates from 40% to 100%, we systematically assessed the stability of the parameter saturation point under different semantic density conditions, ensuring the generalization of our selection.

As illustrated in Figure 9, we observe a consistent performance trajectory across all five seen class settings: metrics exhibit a rapid ascent in the low-ratio regime followed by a distinct plateau once the neighbor quantity exceeds 10%. This uniform saturation phenomenon indicates that the effectiveness of local causal context is scale-invariant; a small fraction of nearest neighbors is sufficient to capture discriminative features without necessitating exhaustive global matching. Consequently, we fix the neighbor ratio at 10% as the default setting. This choice effectively minimizes computational costs while guaranteeing robust performance stability regardless of the available training data scale.

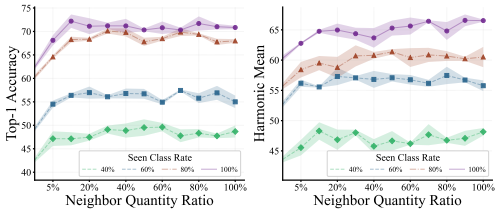


Figure 9: Impact of neighbor quantity on performance across varying seen class proportions (40%–100%).

Disentanglement Strategy Analysis. To verify the necessity and robustness of our Chain-of-Thought discrimination strategy, we conducted a comprehensive evaluation across the two dimensions of disentanglement mechanisms and Large Language Model architectures. Mechanistically, we designed four comparative settings on the CUB dataset, comprising a Baseline without disentanglement, a Random setting that arbitrarily partitions attributes, a Direct Prompt setting using simple instructions, and our proposed CoT strategy. Architecturally, to assess whether our strategy generalizes across different model capabilities, we deployed these settings on both a representative open-source model, LLaMA-3, and a closed-source state-of-the-art model, GPT-4o. To intuitively visualize the execution quality, we further selected six representative class pairs to compare the specific attribute partitions generated by these methods.

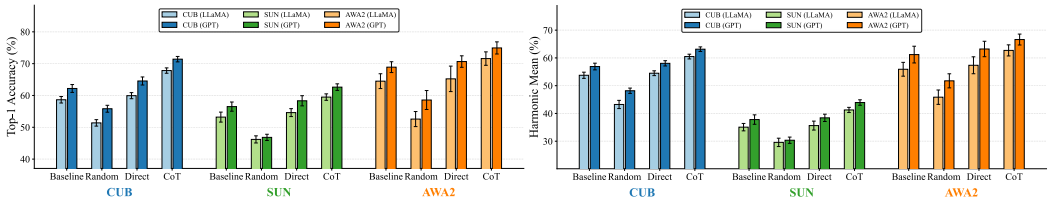


Figure 10: Impact of Disentanglement Strategies and LLM Architectures. We compare the performance of three disentanglement settings (Random, Direct Prompt, and our CoT strategy) evaluated across open-source and closed-source backbones across CUB, SUN, and AWA2 datasets.

The quantitative results in Figure 10 reveal two critical insights. First, regarding strategy effectiveness, the performance hierarchy where the Random setting underperforms the Baseline while the CoT approach surpasses the Direct Prompt remains consistent across both LLaMA and GPT architectures.

Table 5: Qualitative Analysis of Pair-specific Partitioning Errors on CUB. We show 6 confusing fine-grained pairs where directly querying an LLM about whether an attribute is common or discriminative can lead to: (1) False Co-occurrence (discriminative → common), and (2) False Discriminative (common → discriminative). Our CoT strategy corrects these cases by explicitly conditioning on the pairwise comparison context and checking consistency.

Confusing Pair	Attribute	Direct Prompt	CoT Output	Reasoning
1. Laysan Albatross vs. **Black-footed Albatross**	Bill Length	✗ Discriminative	✓ Common	Both are large albatrosses with long, robust bills; bill length is not a reliable separator under typical views and varies with pose.
2. Chipping Sparrow vs. **Clay-colored Sparrow**	Breast Pattern	✗ Discriminative	✓ Common	Both typically appear with a largely plain breast in standard field views; breast pattern is weak and easily confounded by lighting fluff.
3. White-crowned Sparrow vs. **White-throated Sparrow**	Crown Pattern	✗ Discriminative	✓ Common	Both present prominent head striping in many views; relying on "striped crown" alone is non-specific without resolving exact stripe configuration and head angle.
	Eye Region	✗ Common	✓ Discriminative	The eye region provides a stable discriminator: White-throated typically has a distinct yellow lores patch, whereas White-crowned lacks this yellow facial patch.
4. Glaucous-winged Gull vs. **Western Gull**	Leg Color	✗ Discriminative	✓ Common	Both are commonly described as pink-legged large gulls; leg color overlaps strongly and is sensitive to mud, so it should be treated as co-occurring.
	Wingtip Color	✗ Common	✓ Discriminative	Wingtip coloration is discriminative: Western typically shows very dark wingtips with strong contrast, while Glaucous-winged more often has gray wingtips that blend with the mantle.
5. Forster's Tern vs. **Common Tern**	Bill Color	✗ Discriminative	✓ Common	Both can show orange-red bills with a black tip depending on season; bill color alone is ambiguous without additional constraints, so it is not reliably discriminative here.
6. Blue-headed Vireo vs. **Yellow-throated Vireo**	Wing Bars	✗ Discriminative	✓ Common	Both typically show two pale wing bars; bar count or presence overlaps and is not a stable separator, especially under blur or partial wing visibility.
	Belly Color	✗ Common	✓ Discriminative	Underparts coloration is discriminative: Yellow-throated typically shows a strongly yellow throat underparts that extends toward the belly, while Blue-headed is usually whiter below with more limited yellow on the flanks.

This confirms that incorrect causal attribution is detrimental and validates the superiority of our CoT approach. Second, regarding model generalization, while GPT-4o exhibits a performance advantage due to its extensive knowledge base, our CoT strategy enables the open-source LLaMA to achieve competitive results by effectively narrowing the gap with closed-source models. This quantitative success is corroborated by the qualitative case study in Table 5. We observe that the Direct Prompt method consistently struggles with two specific types of logical errors: incorrectly grouping discriminative attributes as co-occurring and incorrectly grouping co-occurring attributes as discriminative. In contrast, our CoT strategy accurately rectifies these errors, demonstrating its ability to provide high-quality causal reasoning.

E.6 Robustness to LLM Reasoning Errors

Although the proposed framework relies on LLMs to construct pairwise discriminative and co-occurring attribute partitions, it does not require these reasoning results to be perfectly accurate. To explicitly examine how LLM reasoning errors affect the proposed causal intervention mechanism, we conduct a controlled error injection experiment. Starting from the original pairwise partition results, we randomly corrupt a fixed proportion of LLM decisions. Specifically, for each selected class pair (i, j) and each attribute a_k, the binary partition indicator $m_{i,j,k}$ is flipped with probability r, where r denotes the simulated LLM reasoning error rate. After corruption, an attribute originally assigned to the discriminative set is moved to the co-occurring set, and vice versa. We vary r from 0% to 50%, where 0% corresponds to the original LLM partition and 50% approximates a random binary partition. All other components, including semantic descriptions, neighbor selection, model architecture, and training hyperparameters, are kept unchanged. Each setting is repeated with multiple

random seeds, and we report the average ZSL Top-1 accuracy and GZSL harmonic mean on CUB, SUN, and AWA2. For comparison, we include CEMIL as an external image-free ZSL baseline and w/o Pairwise Causal Disentanglement as an internal mechanism-level control. The former reflects method-level comparison with a representative baseline, while the latter isolates whether corrupted pairwise causal partitions still provide useful intervention signals when the remaining framework is kept unchanged.

Table 6: Robustness analysis under simulated LLM reasoning errors. We randomly flip a proportion of pairwise attribute partition decisions and report ZSL Top-1 accuracy (**T**) and GZSL harmonic mean (**H**). Here, r denotes the injected LLM reasoning error rate.

Setting	**Error Rate** r	**CUB**		**SUN**		**AWA2**	
		T	**H**	**T**	**H**	**T**	**H**
CEMIL	–	62.2±1.1	57.9±1.4	56.5±0.9	38.5±1.2	68.9±2.4	61.2±3.1
w/o Pairwise Causal Disentanglement	–	66.9±1.6	58.8±1.6	57.6±1.4	38.8±1.3	69.1±1.6	61.8±2.4
Ours with corrupted LLM partitions	0%	71.5±1.1	63.2±0.9	62.7±0.8	43.9±0.7	74.9±1.5	66.7±2.1
	10%	69.4±1.4	61.1±1.2	60.6±1.3	41.8±0.9	72.5±1.8	64.6±2.4
	20%	68.2±1.8	60.4±1.5	58.5±1.5	39.7±1.2	70.4±2.7	62.9±3.6
	30%	59.4±2.2	52.3±1.9	47.6±2.1	31.5±1.8	61.2±3.5	53.7±4.4
	40%	51.3±2.5	44.8±2.3	38.2±2.4	25.1±2.2	53.5±4.2	46.2±5.3
	50%	47.6±2.9	40.2±2.7	31.6±3.2	20.4±2.9	48.4±5.1	39.2±6.4

The results in Table 6 show that performance decreases monotonically as the injected error rate increases. When $r = 0\%$, the model recovers the full version of our framework and achieves the best performance across all datasets. Under a mild error rate of 10%–20%, it still outperforms both CEMIL and the internal variant without pairwise causal disentanglement on all datasets and metrics, indicating that the proposed framework can tolerate limited LLM reasoning noise. This robustness comes from the fact that intervention-guided training aggregates causal constraints across many class pairs and attributes, rather than depending on a single partition decision. However, when the error rate increases to 30% and above, the degradation becomes more pronounced. In this case, corrupted partitions may invert the intended intervention semantics, causing the invariance loss to suppress useful discriminative variations or the sensitivity loss to enlarge non-discriminative variations. As the error rate approaches 50%, the partition becomes close to random and the intervention signal is substantially damaged. These results show that moderate LLM reasoning errors are tolerable, while accurate pairwise causal partitioning remains important for fully exploiting intervention-guided weight refinement.

E.7 Comprehensive Dataset Analysis

Table 7: Extended comparison with baselines on aPY and FLO datasets under ZSL (Top-1 Accuracy **T**) and GZSL (Harmonic Mean **H**, unseen **u**, seen **s**) settings. "ΔImp" denotes the relative improvement of our method over the SOTA baseline.

Semantic Source	**Method**	**Zero-Shot Accuracy**		**Generalized Zero-Shot Accuracy**					
		aPY	**FLO**	**aPY**			**FLO**		
		T	**T**	**H**	**u**	**s**	**H**	**u**	**s**
Expert-constructed Class Attributes	ConSE	25.1±0.3	31.8±0.6	4.6±0.1	2.4±0.1	92.4±0.2	0.8±0.1	0.4±0.1	98.4±0.1
	COSTA	37.2±0.4	35.9±0.8	0.0±0.0	0.0±0.0	92.5±0.1	0.0±0.0	0.0±0.0	98.6±0.1
	SubReg	36.9±2.5	36.5±4.8	0.0±0.0	0.0±0.0	92.5±0.1	0.0±0.0	0.0±0.0	98.6±0.1
	wDAE	36.5±1.6	39.2±3.2	0.2±0.1	0.1±0.0	92.1±0.4	0.0±0.0	0.0±0.0	98.3±0.2
	VGSE	38.1±0.3	38.2±1.9	22.9±1.1	13.1±0.7	89.6±0.4	36.9±3.0	23.1±2.3	93.0±1.2
	ICIS	34.5±1.8	50.4±1.5	17.0±1.4	9.4±0.9	90.5±0.2	47.6±2.8	31.9±2.5	94.1±0.5
LLM-generated Class Attributes	ConSE	41.2±0.3	35.8±0.5	4.9±0.1	2.5±0.1	92.5±0.1	0.7±0.1	0.4±0.1	98.6±0.1
	COSTA	48.6±0.4	47.9±0.6	0.0±0.0	0.0±0.0	92.5±0.2	0.0±0.0	0.0±0.0	98.6±0.1
	SubReg	46.7±2.1	44.6±4.5	0.2±0.1	0.1±0.0	92.5±0.1	0.9±0.4	0.5±0.2	98.6±0.1
	wDAE	46.1±2.8	47.1±2.5	0.2±0.1	0.1±0.0	92.0±0.3	0.0±0.0	0.0±0.0	98.5±0.2
	VGSE	45.2±0.5	41.2±0.9	27.1±1.3	15.9±1.1	89.9±0.4	38.5±2.0	24.3±1.5	92.9±0.9
	ICIS	44.8±2.3	53.2±2.4	19.1±1.8	10.7±1.1	90.6±0.2	45.6±2.7	30.1±2.4	94.5±0.4
	CEMIL	46.5±1.9	55.4±2.1	22.8±1.5	13.2±1.0	84.5±0.6	48.2±2.3	32.4±2.1	93.8±0.7
	Ours	**53.1±1.1**	**63.8±1.0**	**34.4±1.1**	21.3±0.9	89.4±0.2	**55.5±1.5**	39.5±1.4	93.4±0.9
	ΔImp	9%	15%	27%	-	-	15%	-	-

To comprehensively evaluate the generalization capability of our method, we extended the experimental scope from three standard benchmarks to five datasets covering diverse domains and granularities. Table 7 presents the complete comparison results against state-of-the-art methods on the extended datasets including FLO and aPY. Our method achieves SOTA or highly competitive performance across all datasets. Particularly on the challenging Oxford Flowers datasets, we achieved significant improvements relative to the best baseline.

By comparing the improvement margins across different datasets, we observe a critical trend where the performance gains on fine-grained datasets such as CUB, SUN and FLO are notably higher than those on coarse-grained datasets like AWA2 and aPY. This phenomenon is highly consistent with our theoretical expectations. The primary bottleneck in fine-grained classification is feature entanglement caused by high co-occurrence which is precisely the core issue our causal intervention mechanism aims to address. In contrast, discriminative features in coarse-grained classes are inherently more salient rendering the marginal benefit of disentanglement relatively smaller.

Table 8: Ablation study on the core components of the proposed framework. We report Zero-Shot Accuracy (**T**) and Generalized Zero-Shot metrics (**H**, **u**, **s**) with standard deviations on aPY and FLO datasets.

Method Ablation	Zero-Shot Accuracy		Generalized Zero-Shot Accuracy					
	aPY	FLO	aPY			FLO		
	T	T	H	u	s	H	u	s
Full Model (Ours)	**53.1±1.1**	**63.8±1.0**	**34.4±1.1**	21.3±0.9	89.4±0.2	**55.5±1.5**	39.5±1.4	93.4±0.9
w/o Shared-query Attention	52.4±1.2	62.9±0.9	33.1±1.3	20.4±0.9	89.6±0.2	54.2±1.6	38.2±1.5	93.6±0.8
w/o Sensitivity Loss	51.5±1.4	61.2±1.3	31.5±1.5	19.2±1.1	89.5±0.2	52.1±1.9	36.1±1.7	93.5±0.9
w/o Invariance Loss	50.6±1.5	60.5±1.4	30.2±1.6	18.2±1.2	89.8±0.2	51.3±1.8	35.4±1.6	93.3±0.9
w/o Pairwise Disentanglement	50.1±1.6	59.8±1.4	29.4±1.6	17.6±1.3	90.0±0.3	50.4±2.0	34.5±1.8	93.9±1.0
w/o Causal Losses	46.5±2.1	56.4±2.2	25.1±2.0	14.6±1.5	90.5±0.4	45.9±2.5	30.2±2.3	94.8±1.1

Furthermore, we repeated the systematic ablation experiments on these extended datasets. We evaluated the individual contributions of the semantic architecture, causal disentanglement strategy, and geometric constraints. Table 8 shows the complete ablation results across all datasets. Replacing fine-grained disentanglement with global statistics consistently leads to a general performance decrease across all domains, confirming that the context-dependent nature of causal relationships is a universal attribute in fine-grained recognition. Similarly, removing geometric constraints consistently reduces the separability of difficult classes on specialized datasets such as Oxford Flowers. This cross-dataset validation strongly demonstrates that our overall framework is a general solution applicable to various fine-grained tasks.

Table 9: Comparison of computational efficiency across CUB, SUN, and AWA2 datasets. We report Training Time and Inference Time in seconds.

Method	Training Time (s)			Inference Time (s)		
	CUB	SUN	AWA2	CUB	SUN	AWA2
ICIS	43±2.2	151±3.5	26±0.8	1.1±0.1	2.0±0.1	2.2±0.1
CEMIL	51±4.5	162±7.1	32±4.1	2.2±0.2	2.3±0.1	2.4±0.2
Ours	75±6.8	223±5.4	47±7.5	2.2±0.1	2.3±0.1	2.2±0.1

E.8 Computational Efficiency Analysis

To assess the practical deployability of our framework, we compare its computational cost with the baseline methods, ICIS and CEMIL. We report two metrics, including Training Time and Inference Time. Training Time is defined as the wall-clock time required for model optimization until convergence. Inference Time is defined as the wall-clock time of the evaluation stage after preprocessing, including classifier-weight synthesis, logit computation over the pre-extracted test features, and metric computation, while excluding LLM calls and visual feature extraction.

The results are summarized in Table 9. During training, our method incurs a moderate increase in time consumption compared with the baselines. This overhead mainly comes from constructing intervened semantic variants and optimizing the additional invariance and sensitivity losses required

for intervention-guided refinement. Since these operations are performed during training, they do not introduce additional LLM calls or causal reasoning during deployment.

At inference time, the trained generator directly synthesizes classifier weights, and the subsequent evaluation follows the same classifier-based prediction pipeline as the baselines. As shown in Table 9, our method maintains inference time comparable to ICIS and CEMIL across all datasets. This indicates that the intervention-guided mechanism is largely absorbed into the learned weight generator, leading to improved classification performance without introducing substantial deployment-time overhead. Therefore, the proposed method remains practical for image-free classifier expansion scenarios where inference efficiency is important.

E.9 LLM Cost Estimation

To clarify the scalability of our LLM-assisted pipeline, this section estimates the LLM overhead as a function of the total number of classes $N = |\mathcal{S} \cup \mathcal{U}|$. The dominant LLM cost comes from pairwise causal disentanglement, where the model compares semantically similar class pairs to identify discriminative and co-occurring attributes. A naive implementation would enumerate all class pairs, resulting in $N(N-1)/2$ LLM comparisons, which is $O(N^2)$ and becomes costly for datasets with many classes.

In practice, we avoid exhaustive all-pair comparison by applying nearest-neighbor pruning before LLM querying. For each target class, we retrieve its top-n semantically nearest classes according to the global semantic embeddings and only perform pairwise disentanglement within this local hard-neighbor set. This design is consistent with the fine-grained motivation of our method, since weight entanglement and ambiguous decision boundaries mainly arise among visually and semantically similar classes rather than among distant class pairs. In our experiments, we set $n = \rho N$ with $\rho = 10\%$, which substantially reduces the number of evaluated class pairs and focuses the LLM budget on the most confusing comparisons.

Moreover, the pairwise relation is treated as undirected for LLM inference. That is, once a pair (i, j) has been evaluated, the result is cached and reused for the reverse direction when needed. Since nearest-neighbor relations are often approximately symmetric in semantic space, the actual number of unique LLM calls is typically smaller than the crude directed upper bound Nn. More precisely, the number of evaluated unique pairs is upper bounded by Nn, and is close to $Nn/2$ when most neighbor relations are reciprocal. Therefore, when n is fixed as a small constant, the number of LLM comparisons scales as $O(Nn)$, which is linear in N in practice. When n is set as a fixed ratio of N, as in our experiments, the worst-case asymptotic order remains $O(N^2)$, but the constant factor is greatly reduced compared with exhaustive all-pair comparison. For $\rho = 10\%$, the evaluated pair set is only a small fraction of all possible class pairs.

The per-call token usage is also controlled by a fixed prompt template and structured JSON-style outputs. As a result, the total token budget is approximately proportional to the number of evaluated unique pairs. Importantly, all LLM-based semantic generation and pairwise disentanglement are performed only during preprocessing and training. During inference, the trained weight generator directly synthesizes unseen classifier weights without additional LLM calls.

F LLM-Based Generation and Causal Disentanglement

This section details the LLM-assisted pipeline used in our framework. The purpose of this pipeline is not to treat the LLM as a direct classifier or an unrestricted semantic generator. Instead, we use the LLM as a structured knowledge elicitation engine that converts class names into attribute-level descriptions and further separates pair-specific discriminative attributes from co-occurring attributes. This design is important for fine-grained image-free classifier expansion, because the model cannot access original images and must rely on semantic descriptions to synthesize classifier weights. If the semantic descriptions are unstructured, inconsistent, or hallucinated, the generated weights may inherit these errors and become geometrically entangled. Therefore, we impose three principles on the LLM pipeline. First, the generated semantics should be visually verifiable. Second, all classes should be described under a shared attribute schema to ensure structural alignment. Third, discriminative and co-occurring attributes should be determined through attribute-wise judgments in a pairwise comparison context rather than through a global attribute assignment.

F.1 Semantic Description Generation

Directly asking an LLM to generate complete descriptions for all classes often leads to unstable outputs. The model may mix visual and non-visual knowledge, generate literary expressions that are difficult to encode, or hallucinate traits that are not reliably observable in standard images. These problems are especially harmful in the image-free setting, because there are no visual samples available to correct the semantic noise during training. To alleviate this issue, we decompose semantic generation into two structured stages, namely global attribute construction and class-specific description generation.

Global attribute construction. We first ask the LLM to construct a global attribute list shared by all classes in the dataset. This step defines the semantic coordinate system used by subsequent modules. Instead of allowing each class to be described with arbitrary attributes, we require all classes to be characterized under the same attribute schema. For example, in a bird dataset, the schema may include bill shape, crown pattern, wing color, tail shape, belly color, and body size. In a flower dataset, the schema may include petal shape, petal color, center pattern, leaf form, and overall structure. This shared schema ensures that semantic descriptions are comparable across classes and that later interventions can be performed attribute by attribute.

To improve the quality of the attribute schema, we use a closed-loop reasoning process consisting of Think, Generate, Verify, and Refine stages. In the Think stage, the LLM analyzes the dataset taxonomy and identifies visual dimensions that are likely to be useful for distinguishing classes. In the Generate stage, it proposes a candidate list of attributes. In the Verify stage, it removes abstract, non-visual, overly broad, or ambiguous attributes. In the Refine stage, it standardizes the naming format and merges redundant attributes. This procedure reduces schema inconsistency and provides a visually grounded basis for downstream causal intervention.

Class-specific semantic description. Given the global attribute list, we then generate class-specific descriptions for each class. For a class i, the LLM fills every attribute slot in the schema with a concise visual description. This produces an attribute-level semantic representation rather than a single holistic sentence. The structured format is important because our intervention mechanism manipulates semantic factors at the attribute level. If the descriptions were generated as unstructured paragraphs, it would be difficult to determine which semantic component should be intervened on.

During this step, we again use a verification and refinement mechanism. The LLM is instructed to check whether each generated description corresponds to a physically observable visual trait. Descriptions that rely on habitat, behavior, taxonomy, metaphor, or uncertain knowledge are removed or rewritten. This self-correction step is designed to reduce hallucinated semantics and enforce visual verifiability. The final output is a JSON object whose keys exactly match the global attribute list. This ensures that every class has a structurally aligned semantic representation and that each attribute can be independently encoded and manipulated.

The global class description T_i^{global} used as the semantic anchor is not generated by an additional LLM query. After obtaining the structured attribute-level descriptions, we deterministically concatenate the class name and the verified attribute-wise descriptions with a fixed template:

$$T_i^{global} = \text{Template}\left(y_i, \{(a_k, T_{i,k})\}_{k=1}^{K}\right).$$

This construction ensures that the global semantic anchor remains consistent with the attribute-level representation used for intervention.

F.2 Pairwise Causal Disentanglement

Although structured descriptions provide aligned semantic inputs, they do not by themselves solve the fine-grained ambiguity problem. In fine-grained recognition, many attributes are shared by visually similar classes. These shared attributes may dominate the semantic description but contribute little to the local decision boundary. Therefore, the key question is not only what attributes a class has, but which attributes are truly useful for distinguishing one class from another.

A central property of our framework is that the role of an attribute is pair-dependent. An attribute should not be globally labeled as discriminative or co-occurring. For example, wing shape can be discriminative when comparing a bird with a mammal, but it may become co-occurring when comparing two similar bird species. Conversely, a subtle crown stripe or eye-ring pattern may be

irrelevant for coarse-grained recognition but decisive for a specific pair of bird species. Therefore, we perform causal disentanglement at the class-pair level.

Efficient hard negative sampling. Exhaustively comparing all class pairs would introduce unnecessary LLM overhead, especially for datasets with hundreds of classes. Moreover, many class pairs are visually distant and do not require fine-grained causal separation. We therefore restrict pairwise comparison to hard negative neighbors. For each target class, we first retrieve its top-n semantically nearest classes according to the global semantic embeddings. These nearest neighbors are more likely to share co-occurring attributes and to produce ambiguous decision boundaries. By focusing on these local comparison sets, the LLM budget is spent on the class pairs where causal disentanglement is most needed.

Pairwise binary discrimination. For each selected class pair, we ask the LLM to evaluate every attribute one by one under a binary decision criterion. The LLM must make an attribute-wise judgment for each attribute a_k in the shared schema. Specifically, it determines whether the attribute provides a necessary and visually observable difference for separating the two classes, or whether it is shared, weakly distinguishable, ambiguous, unstable, or irrelevant in the current comparison. Attributes judged to provide necessary visual differences are assigned to the discriminative set. Attributes judged to be shared or non-decisive are assigned to the co-occurring set. In implementation, all attributes of a class pair can be provided in a single prompt, but the LLM is explicitly instructed to return one binary decision for every input attribute.

This design differs from simple keyword matching. The LLM is required to reason about the local decision boundary. For instance, if both classes have striped crowns, the attribute crown pattern should not automatically be treated as discriminative unless the specific stripe configuration provides a reliable distinction. Similarly, if both classes have similar wing bars, the presence of wing bars should be treated as co-occurring even though it is visually salient. This comparison-based reasoning helps avoid two common errors. The first is false discriminative assignment, where a shared attribute is incorrectly treated as class-defining. The second is false co-occurring assignment, where a subtle but decisive difference is ignored.

Role in causal intervention. The output of pairwise causal disentanglement provides the intervention targets for the model. Co-occurring attributes define the semantic factors that should be intervened on without changing class identity. Discriminative attributes define the semantic factors whose intervention should change class identity. Therefore, the LLM output does not directly determine the classifier prediction. Instead, it provides structured supervision for constructing positive and negative interventional semantic variants. The final effect is learned through the weight-space invariance and sensitivity losses.

F.3 Prompt Design Details

To make LLM outputs reliable and reproducible, we use role-specific prompts, explicit reasoning protocols, and structured JSON output schemas. The prompts are designed according to three principles. First, the LLM is cast as a fine-grained visual classification expert, which encourages it to focus on observable visual cues rather than background knowledge. Second, the output schema is fixed, which prevents format drift and ensures that downstream parsing is stable. Third, the prompts include verification instructions, so the LLM must check whether each generated or selected attribute is visually grounded.

Below we detail the prompts for global attribute construction, class-specific description generation, and pairwise causal disentanglement.

Prompt 1: Global Attribute Construction

System Role: You are an expert taxonomist specializing in fine-grained visual classification. Your task is to identify visual attributes that can be observed from standard images. Do not use habitat, behavior, taxonomy, geographic distribution, or abstract concepts as attributes.

Task: Construct a comprehensive and standardized Global Attribute List based on the provided dataset taxonomy. The attribute list will be used to describe every class in a structurally aligned way.

Chain-of-Thought Protocol:
1. **Think:** Analyze the class names and infer the visual dimensions that are likely to vary across this dataset.
2. **Generate:** Propose a candidate list of fine-grained visual attributes.
3. **Verify:** Remove attributes that are non-visual, ambiguous, overly broad, redundant, or difficult to verify from images.
4. **Refine:** Standardize the remaining attributes using concise canonical names. Merge duplicates and keep the schema compact but complete.

Input: [List of all class names from the dataset]

Output Format: JSON list: `["bill_shape", "wing_pattern", "belly_color", ...]`

Figure 11: Prompt for constructing a visually grounded global attribute schema.

Prompt 2: Class-Specific Description Generation

System Role: You are an expert taxonomist specializing in fine-grained visual classification. You must describe only visually observable traits that may appear in standard images.

Task: Generate a structured semantic description for `<Target Class>` using the provided `<Attribute List>`. Each attribute must be filled with a concise visual description.

Chain-of-Thought Protocol:
1. **Retrieve and Fill:** For each attribute, recall the typical visual appearance of the target class and generate a concise description.
2. **Self-Correction:** Check each description according to the following criteria:
   - *Visual Verifiability:* Keep only traits that can be observed from images.
   - *Metaphor Removal:* Remove literary, metaphorical, or subjective descriptions.
   - *Hallucination Check:* If a trait is uncertain, invented, or not reliably present, remove or rewrite it.
   - *Attribute Alignment:* Ensure that every key exactly matches the given attribute list.
3. **Finalize:** Return the result as a JSON object with no extra commentary.

Input: Target class name and global attribute list.

Output Format: JSON object: `{"bill_shape": "...", "wing_pattern": "...", ...}`

Figure 12: Prompt for generating visually grounded and structurally aligned class descriptions.

Prompt 3: Pairwise Causal Disentanglement

System Role: You are a fine-grained visual discrimination expert. Your goal is to determine which attributes are necessary for distinguishing two visually similar classes.

Task: Compare `<Class A>` and `<Class B>` using their structured attribute descriptions. Separate attributes into discriminative attributes and co-occurring attributes. You must make an attribute-wise binary judgment for every attribute in the input list.

Decision Criterion: An attribute is discriminative only if its difference provides an explicit, visually observable, and necessary cue for defining the boundary between the two classes. An attribute is co-occurring if it is shared, visually similar, weakly distinguishable, ambiguous, unstable, or not necessary for this pairwise distinction.

Chain-of-Thought Protocol:
1. **Compare:** Examine the two descriptions under each attribute one by one.
2. **Causal Reasoning:** Ask whether the difference in the current attribute is necessary and visually reliable for distinguishing Class A from Class B.
3. **Decision:**

- Assign $m_{i,j,k} = 1$ and place the attribute into the **Discriminative Set** only if it provides a necessary and visually reliable difference.
- Assign $m_{i,j,k} = 0$ and place the attribute into the **Co-occurring Set** if it is shared, similar, unstable, ambiguous, or insufficient for distinction.

4. **Verification:** Check whether any shared attribute has been incorrectly marked as discriminative, or whether any subtle but decisive attribute has been incorrectly marked as co-occurring.
5. **Completeness Check:** Ensure that every input attribute appears exactly once in either the discriminative set or the co-occurring set.

Input: Class A name, Class B name, and their aligned JSON attribute descriptions under the same global attribute list.

Output Format: JSON object: `{"discriminative": ["attr_1", ...], "co_occurring": ["attr_2", ...]}`

Figure 13: Prompt for context-aware pairwise causal disentanglement.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made match theoretical and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly discussed in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Full set of assumptions and a complete and correct proof are provided for all theoretical results, statements.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All necessary settings and codes are available for the experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Data and codes are available in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and test details necessary to understand results are specified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Statsitical significance are considered and reported in the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sufficient information on the computer resources is provided in the experimental setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both potential positive societal impacts and potentially negative societal impacts are discussed in Appendix D

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects, crowdsourcing, user studies, or personal data collection, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are used for semantic description generation and pairwise attribute disentanglement, as described in Appendix F.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.