D2SC: Dual-Stage Diffusion-Guided Semantic-Visual Coupling for Zero-Shot Learning

Qitong Fang
Jilin Jianzhu University
fangqitong@student.jlju.edu.cn

***Abstract*—Generative zero-shot learning remains challenging in practice. When conditioning on diffusion timesteps, semantics become step-dependent: the same concept drifts across noise levels, weakening conditional guidance for synthesis. Adversarial alignment alone further leaves a mismatch between synthesized and real features, both in global statistics and class-wise geometry; and in multi-branch setups, discriminator heads can produce inconsistent gradients that destabilize training. To address these issues, we design D2SC (Dual-Stage Diffusion-Guided Semantic-Visual Coupling) with a simple rationale: first stabilize the conditional prior, then align the visual space. Concretely, the Contrastive Prior Regularizer (CPR) enforces timestep consistency to suppress conditional drift, while the Visual Feature Synthesizer (VFS) performs multi-alignment via maximum mean discrepancy and center loss, complemented by lightweight discriminator distillation for coherent feedback. The two modules share semantic mappings and a unified schedule. With a single hyper-parameter setting across datasets, D2SC achieves robust zero-shot learning (ZSL) and generalized zero-shot learning (GZSL) gains. The code is available at https://github.com/btxq-ily/d2sc.**

***Index Terms*—Zero-Shot Learning, Diffusion Guidance, Dual-Stage Coupling, MMD, Center Loss**

I. Introduction

Zero-shot learning (ZSL) aims to recognize unseen classes using semantic side information such as attributes or word vectors [1]–[3]. In generative ZSL, visual features are synthesized to train a conventional classifier on pseudo-data [3]. Despite recent progress, two critical challenges persist. First, in diffusion-based generators, semantic conditioning often suffers from timestep-dependent drift, where a fixed semantic concept induces varying representations across different noise levels, thereby weakening the conditional guidance [4]. Second, standard adversarial alignment often fails to bridge the gap between synthesized and real distributions in terms of both global statistics and class-wise geometry, while multi-branch architectures may suffer from gradient inconsistency across different discriminator heads [5]. If these issues remain unresolved, the generator will be provided with unstable conditional anchors and imprecise alignment signals, leading to degraded synthesis quality and poor generalization on unseen classes.

Existing ZSL/GZSL solutions typically strengthen the semantics-to-visual bridge from different perspectives. However, a common challenge remains in preserving class-conditional geometry under noisy semantic guidance, especially in the generalized setting where a bias toward seen classes is prevalent. This difficulty is further amplified in diffusion-conditioned synthesis, where temporal drift in the latent space and the lack of class-wise compactness can lead to overlapping feature clusters. Consequently, achieving a favorable balance between seen and unseen recognition requires not only distributional alignment but also a stable conditional prior and explicit geometric constraints.

To address these challenges, we introduce D2SC (Dual-Stage Diffusion-Guided Semantic-Visual Coupling), a framework that stabilizes the conditional prior before aligning the visual space. The Contrastive Prior Regularizer (CPR) enforces timestep consistency of contrastive representations to suppress temporal drift and produce a dependable semantic anchor. Building on this stable condition, the Visual Feature Synthesizer (VFS) performs multi-alignment by combining Maximum Mean Discrepancy (MMD) with a center constraint to ensure both global distribution matching and class-wise compactness. Furthermore, a lightweight distillation objective synchronizes gradients across multiple discriminator branches, ensuring optimization stability.

The contributions are three-fold. (1) A dual-stage formulation that first consolidates the conditional prior and then focuses on visual alignment, improving stability and transfer across datasets. (2) A timestep-consistency regularization for contrastive representations that reduces variation across diffusion steps. (3) From an empirical perspective, D2SC achieves competitive performance on mainstream datasets. Compared with a representative diffusion-conditioned generator (ZeroDiff [6]), D2SC improves the GZSL harmonic mean (H) on AWA2 by 2.6 points while maintaining comparable model complexity (see Section IV-C); this gain is consistent with our design goals of suppressing step-dependent conditional drift and improving class-conditional alignment. Experiments further show that the timestep regularizer, MMD/center constraint, and lightweight distillation bring complementary gains, and a short diffusion chain improves stability and efficiency.

II. Related Work

We briefly review representative ZSL/GZSL paradigms and clarify where the difficulty lies for diffusion-conditioned generative learning. While existing baselines excel on specific aspects such as boosting seen accuracy or tightening cross-modal alignment, our objective is to enhance *unseen* generalization while maintaining overall optimization stability.

Generative ZSL and the GZSL bias. Feature synthesis trains a conventional classifier on pseudo features, but the

GZSL setting introduces a notorious bias toward seen classes: a method can obtain high seen accuracy (S) while producing low unseen accuracy (U), resulting in a modest harmonic mean (H). In practical zero-shot deployment, the core objective is to ensure the model correctly recognizes novel classes rather than being biased toward seen categories. Therefore, U serves as a direct proxy for zero-shot generalization, and H reflects the balance between maintaining seen recognition and expanding unseen coverage.

Diffusion-conditioned synthesis and step-dependent semantics. Recent diffusion-based generators (e.g., ZeroDiff [6]) improve synthesis quality by exploiting iterative denoising, yet conditioning can become timestep-dependent: the same semantic concept may correspond to different effective conditions at different noise levels. Such drift weakens conditional guidance and can manifest as unstable class-conditional geometry in the synthesized feature space, particularly when the semantic descriptors are ambiguous (a common issue on fine-grained datasets). Our CPR directly targets this failure mode by enforcing timestep consistency on the contrastive state used as condition.

Alignment beyond global matching. A strong line of work focuses on reducing cross-modal discrepancy and projection bias. For instance, SDVAE [7] emphasizes disentangling semantic factors to better cover unseen classes, and DFCAFlow [8] strengthens dual alignment and confusion alleviation, showing clear advantages on challenging datasets (e.g., SUN [9] in Table I). Meanwhile, ICIS [10] demonstrates that injecting classifiers can yield extremely strong seen-side performance on AWA2 [11], but it also highlights the trade-off that high S does not necessarily translate into high U. These methods motivate two observations that guide our design: (i) matching only marginal statistics is insufficient—class-wise geometry matters; (ii) improvements on U (and thus H) are particularly valuable when the seen side is already strong. D2SC follows these insights by combining distributional alignment (MMD) with class-conditional compactness (center constraint), while stabilizing the condition across diffusion steps.

III. Methodology

A. Problem definition

We first outline preliminaries and notation. Given seen classes with visual features $\mathbf{x} \in \mathbb{R}^{d_v}$ and attribute vectors $\mathbf{a} \in \mathbb{R}^{d_s}$, the goal is to recognize unseen classes. D2SC consists of two modules: the Contrastive Prior Regularizer (CPR) generates contrastive representations $\mathbf{c}$ under diffusion timestep consistency; the Visual Feature Synthesizer (VFS) generates visual features $\hat{\mathbf{x}}$ with multi-level alignment. Outputs are: (i) synthesized visual features $\hat{\mathbf{x}}$ (and optionally synthesized contrastive features) used to train downstream classifiers; and (ii) a final classifier $f(\cdot)$ for ZSL and GZSL, producing labels on unseen or seen+unseen test sets.

We use τ for the timestep regularizer weight. Module-wise adversarial losses are $\mathcal{L}_{\text{adv}}^{\text{CPR}}$ (CPR) and $\mathcal{L}_{\text{adv}}^{\text{VFS}}$ (VFS). Auxiliary objectives include the variational autoencoder loss $\mathcal{L}_V$, maximum mean discrepancy loss $\mathcal{L}_M$, center loss $\mathcal{L}_C$, attribute reconstruction loss $\mathcal{L}_R$, and the multi-branch distillation loss $\mathcal{L}_D$. Scalar weights are denoted by $\gamma_{x_0}, \gamma_{x_t}, \gamma_{\text{ADV}}, \gamma_{\text{VAE}}, \gamma_{\mathcal{M}}, \gamma_{\mathcal{C}}, \gamma_{\mathcal{R}}, \gamma_{\text{dist}}$, with concrete values specified in experiments; γ_{ADV} weights adversarial terms, γ_{VAE} weights the VAE objective, $\gamma_{\mathcal{M}}$ weights MMD, $\gamma_{\mathcal{C}}$ weights the center constraint, $\gamma_{\mathcal{R}}$ weights attribute reconstruction, γ_{dist} weights distillation, and $\gamma_{x_0}, \gamma_{x_t}$ weight the x_0 and x_t branch critic gaps. For long operators, we adopt short symbols: $\mathcal{R}(\cdot,\cdot)$ for reconstruction, $\mathcal{K}(\cdot \,\|\, \cdot)$ for Kullback–Leibler divergence (KLD), $\mathcal{M}(\cdot,\cdot)$ for maximum mean discrepancy (MMD), and $\mathcal{S}(\cdot)$ for the stop-gradient operator. Throughout, we denote a single diffusion step by $\mathcal{T}$; $t \in \{1,\dots,n_T\}$ indexes the current $\mathcal{T}$, where n_T is the number of timesteps. For any discriminator branch $p \in \{c_0, c_t, x_0, x_t, x_c\}$, we define the critic gap as $\Delta D_p = \mathbb{E}[D_p(\text{real})] - \mathbb{E}[D_p(\text{fake})]$.

B. Dual-stage diffusion-guided semantic-visual coupling

We first motivate the design, then detail two components. D2SC follows a simple but practical principle: **make the condition reliable before asking the generator to be precise**. Figure 1 illustrates the overall framework. Given attributes $\mathbf{a}$ and an optional contrastive state $\mathbf{c}_t$, CPR first produces a timestep-stable contrastive condition $\hat{\mathbf{c}}_t$ that serves as a semantic "anchor" across diffusion noise levels. Conditioned on $(\mathbf{a}, \hat{\mathbf{c}}_t)$, VFS then synthesizes visual features $\hat{\mathbf{x}}_0, \hat{\mathbf{x}}_t$ from a latent $\mathbf{z}$ for downstream classifier training. Discriminators $(D_{c_0}, D_{c_t}, D_{x_0}, D_{x_t}, D_{x_c})$ provide adversarial feedback, while complementary alignment objectives ensure that synthesized features match real features not only in global distribution but also in class-wise geometry.

CPR: Timestep-consistent contrastive generation. Contrastive representations serve as intermediate conditional embeddings that bridge semantic attributes and visual features; their role is to provide a stable reference point for the generator across diffusion noise levels. Under the diffusion framework, given a real contrastive representation $\mathbf{c}_0$ and a timestep t, we construct $\mathbf{c}_t$ and $\mathbf{c}_{t+1}$. The generator G_{con} takes noise $\mathbf{z}$, semantics $\mathbf{a}$, and $\mathbf{c}_t$ to predict a clean contrastive state $\hat{\mathbf{c}}_0$, from which we obtain $\hat{\mathbf{c}}_t$ via posterior sampling. Two discriminators, D_{c0} and D_{ct}, supervise the realism of $\hat{\mathbf{c}}_0$ and the conditional transition at timestep t (see Figure 1):

$$\begin{aligned} \mathcal{L}_{\text{adv}}^{\text{CPR}} &= \gamma_{x_0}\,\Delta D_{c_0} + \gamma_{x_t}\,\Delta D_{c_t}, \\ \Delta D_{c_0} &= \mathbb{E}[D_{c_0}(\mathbf{c}_0, \mathbf{a})] - \mathbb{E}[D_{c_0}(\hat{\mathbf{c}}_0, \mathbf{a})], \\ \Delta D_{c_t} &= \mathbb{E}[D_{c_t}(\mathbf{c}_t, \mathbf{c}_{t+1}, \mathbf{a}, t)] - \mathbb{E}[D_{c_t}(\hat{\mathbf{c}}_t, \mathbf{c}_{t+1}, \mathbf{a}, t)]. \end{aligned} \tag{1}$$

with Wasserstein GAN with gradient penalty (WGAN-GP). To suppress timestep drift, we add a timestep-consistency regularizer:

$$\mathcal{L}_{\text{step}} = \|\hat{\mathbf{c}}_t - \mathbf{c}_t\|_2^2, \quad \mathcal{L}_G^{\text{CPR}} = \mathcal{L}_{\text{adv}}^{\text{CPR}} + \tau\,\mathcal{L}_{\text{step}}. \tag{2}$$

This regularizer explicitly constrains the **condition** to be stable across diffusion noise levels, which is essential because the downstream visual synthesis depends on $\hat{\mathbf{c}}_t$ as a semantic anchor; without such stability, the same attribute vector can

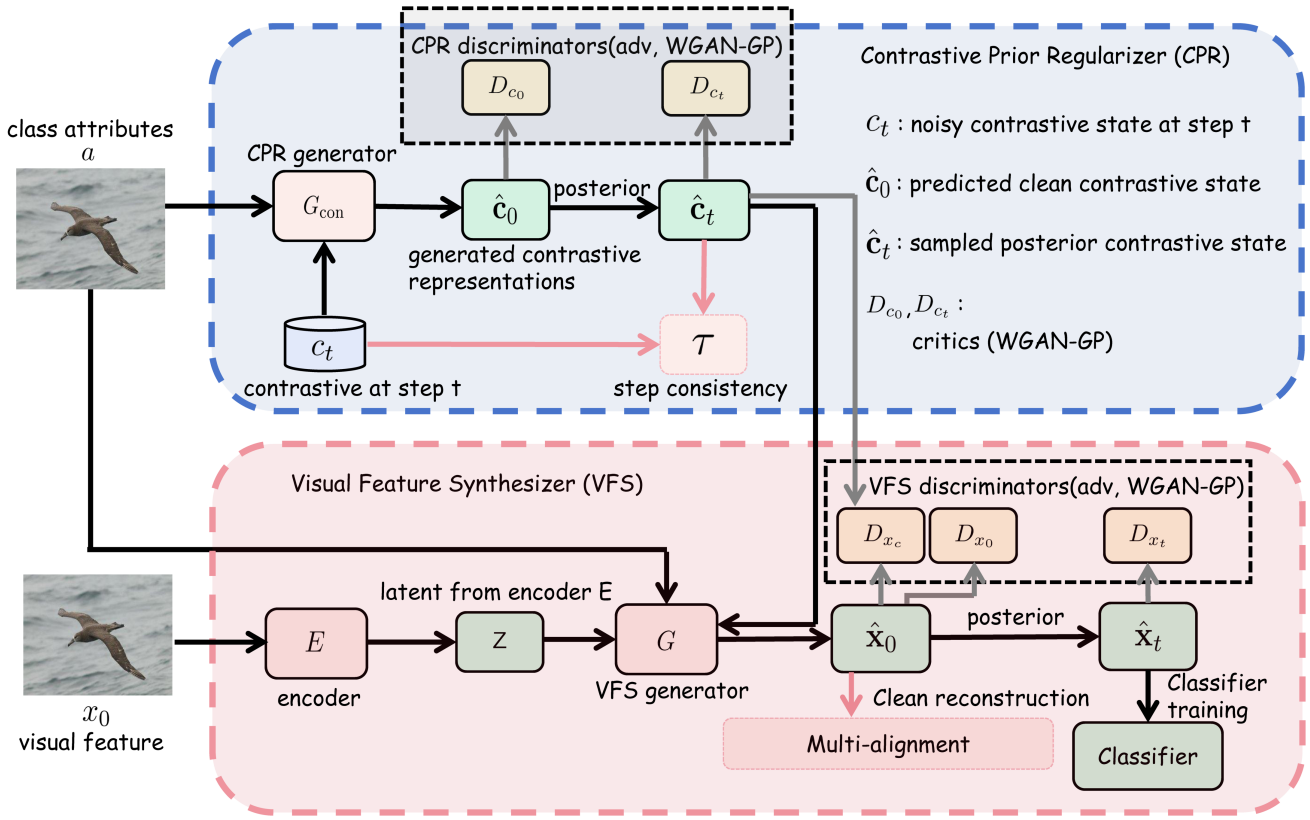

Fig. 1. D2SC training pipeline. The blue dashed region denotes CPR; the red dashed region denotes VFS. Solid arrows indicate processing flow; the thick grey arrow conveys the stabilized condition $\hat{\mathbf{c}}_t$ from CPR to VFS; dashed arrows represent adversarial supervision from discriminators. CPR receives class attributes $\mathbf{a}$ and a noisy contrastive state $\mathbf{c}_t$, and enforces timestep consistency (via $\mathcal{T}$) to suppress step-dependent conditional drift. VFS synthesizes visual features conditioned on $(\mathbf{a}, \hat{\mathbf{c}}_t, \mathbf{z})$ and aligns synthesized and real features via complementary objectives (distributional alignment and class-wise geometry), with lightweight discriminator distillation for coherent feedback.

induce different class-wise geometry at different t, amplifying the seen/unseen imbalance in GZSL.

VFS: Multi-alignment visual generation. In VFS, the encoder E extracts latent $\mathbf{z}$ and the visual generator G synthesizes a clean feature $\hat{\mathbf{x}}_0$ under $(\mathbf{z}, \mathbf{a}, \hat{\mathbf{c}}_t, \mathbf{x}_{t+1}, t)$, followed by posterior sampling to obtain $\hat{\mathbf{x}}_t$. Three discriminators D_{x_0}, D_{x_t}, and D_{x_c} supervise $\hat{\mathbf{x}}_0$, $\hat{\mathbf{x}}_t$, and the semantic-visual coupling $(\hat{\mathbf{x}}_0, \hat{\mathbf{c}}_t)$, respectively (Figure 1). The overall adversarial term is

$$\mathcal{L}_{\text{adv}}^{\text{VFS}} = \gamma_{x_0}\,\Delta D_{x_0} + \gamma_{x_t}\,\Delta D_{x_t} + \Delta D_{x_c}, \tag{3}$$

with WGAN-GP penalties. To enhance fidelity, we use a variational autoencoder (VAE) reconstruction objective

$$\mathcal{L}_V = \mathcal{R}(\hat{\mathbf{x}}_0, \mathbf{x}_0) + \mathcal{K}(\mathcal{N}(\mu, \sigma^2)\,\|\,\mathcal{N}(0,1)). \tag{4}$$

When multiple discriminator branches evaluate the same generated sample, their critic gaps can diverge, producing conflicting gradient directions for the generator. We introduce a distillation term to encourage agreement among these signals:

$$\mathcal{L}_D = \sum_{p \neq q} \left\| \Delta D_p - \mathcal{S}(\Delta D_q) \right\|_2^2, \quad p, q \in \{x_0, x_t, x_c\}, \tag{5}$$

weighted by the ratio induced by the diffusion schedule. By encouraging agreement among critic gaps while preventing collapse via stop-gradient, this term reduces gradient conflict across branches and yields more coherent supervision for the generator. To reduce distribution gap, we add a maximum mean discrepancy (MMD) term

$$\mathcal{L}_M = \mathcal{M}\big(\hat{\mathbf{x}}_0, \mathbf{x}_0\big), \tag{6}$$

and a center constraint via cross-modal semantic centers $\mathbf{m}_y = \Phi(\mathbf{a}_y)$ (where Φ denotes the semantic-to-visual mapping)

$$\mathcal{L}_C = \frac{1}{N} \sum_i \left\| \hat{\mathbf{x}}_0^{(i)} - \mathbf{m}_{y_i} \right\|_2^2. \tag{7}$$

Additionally, a visual-to-semantic mapping Ψ reconstructs attributes, $\mathcal{L}_R = \|\Psi(\hat{\mathbf{x}}_0) - \mathbf{a}\|_2^2$. The final generator objective is

$$\begin{aligned} \mathcal{L}_G^{\text{VFS}} = {} & \gamma_{\text{ADV}} \mathcal{L}_{\text{adv}}^{\text{VFS}} + \gamma_{\text{VAE}} \mathcal{L}_V \\ & + \gamma_{\mathcal{M}} \mathcal{L}_M + \gamma_{\mathcal{C}} \mathcal{L}_C + \gamma_{\mathcal{R}} \mathcal{L}_R. \end{aligned} \tag{8}$$

The two stages are coupled via a shared attribute mapping and unified timestep scheduling, and synthesized features are used to train ZSL/GZSL classifiers. Together with the timestep-consistency regularizer τ, these choices reduce the discrepancy between predicted and real contrastive states across $\mathcal{T}$, tighten the conditional prior used by VFS, and mitigate mode drift. In the visual space, the MMD term $\mathcal{L}_M$ aligns global statistics while the center constraint $\mathcal{L}_C$ enforces class-conditional compactness; lightweight distillation $\mathcal{L}_D$ promotes agreement

among the x_0, x_t, and x_c discriminators, reducing gradient variance. A short diffusion chain (n_T=4) and shared mappings constitute a unified recipe that transfers across datasets with minimal tuning.

C. Diffusion Schedules and Coefficients

We use a continuous diffusion schedule with parameters (β_1, β_2) and a short chain of n_T timesteps to balance stability and efficiency: the schedule yields smooth posterior coefficients for sampling, while a small n_T reduces optimization variance and computational overhead without sacrificing the ability to model step-wise denoising. Let $\{\alpha_t\}$ denote the per-step noise multipliers and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. Following the implementation, we define

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \tag{9}$$

$$\mathbf{x}_{t-1} = \mu_t(\mathbf{x}_t, \mathbf{x}_0) + \sigma_t\,\boldsymbol{\epsilon}, \tag{10}$$

where μ_t and σ_t are posterior coefficients computed from the schedule. Analogous definitions hold for contrastive representations ($\mathbf{c}_t$). These coefficients are pre-computed for efficiency and used in both q-sampling and posterior sampling functions in training and synthesis.

1) Optimization and stability: Each discriminator is trained with WGAN-GP. Let $\mathcal{P}(\cdot)$ denote the gradient penalty term enforcing $\|\nabla\|_2 \approx 1$. The overall discriminator loss aggregates

$$\mathcal{L}_{\text{disc}} = \gamma_{\text{ADV}}\Big(\gamma_{x_0}\,\Delta D_{x_0} + \gamma_{x_t}\,\Delta D_{x_t} + \Delta D_{x_c}\Big) + \gamma_{\text{dist}}\,\mathcal{L}_D + \lambda_1\,\mathcal{P}, \tag{11}$$

with an adaptive λ_1 that increases if the average penalty exceeds a threshold and decreases otherwise, maintaining stable Lipschitz control. We further apply gradient clipping (max-norm 1.0) to E, G, and all discriminators to suppress exploding updates.

2) Semantic-visual mappings and reconstruction: We adopt two linear mappings, denoted by $\Psi : \mathbb{R}^{d_v} \to \mathbb{R}^{d_s}$ (visual→semantic) and $\Phi : \mathbb{R}^{d_s} \to \mathbb{R}^{d_v}$ (semantic→visual). Given class attribute vectors $\{\mathbf{a}_y\}$, the cross-modal visual centers are defined as $\mathbf{m}_y = \Phi(\mathbf{a}_y)$. The attribute reconstruction term $\|\Psi(\hat{\mathbf{x}}_0) - \mathbf{a}\|_2^2$ encourages consistency between synthesized visual features and their semantic descriptions, stabilizing cross-modal alignment and complementing adversarial objectives.

3) Classifier training: For ZSL, we train a softmax classifier on synthesized features of unseen classes and evaluate on unseen test features. For GZSL, we train on the union of real seen features and synthesized unseen features, and evaluate U/S/H. We additionally support variants that concatenate synthesized contrastive features or decode semantic features for joint training, consistent with the code's classifier interfaces.

IV. Experiments

A. Implementation details

Datasets and protocol. We evaluate on AWA2 [11], CUB [12] and SUN [9], as they are standard and complementary: AWA2 [11] is coarse-grained with relatively discriminative attribute descriptions; CUB [12] is fine-grained where class semantics can be highly overlapping; SUN [9] covers diverse scenes with non-trivial attribute correlations. This spectrum stresses both synthesis fidelity and semantic robustness, which is crucial for diffusion-conditioned generators. We follow the widely used ZSL/GZSL splits and protocols [11]. For ZSL, we train a classifier on synthesized features of unseen classes and test on unseen data; for GZSL, we train on real seen + synthesized unseen and test on the joint label space. We report unseen accuracy U, seen accuracy S, and their harmonic mean H, because U directly measures zero-shot generalization while H reflects the balance that GZSL inherently demands.

Training recipe. Both CPR and VFS adopt WGAN-GP with gradient clipping and Adam. We use a short diffusion chain (n_T=4) and shared semantic mappings. CPR applies the timestep-consistency regularizer (τ=0.02) and trains with batch size 64 and learning rate 1×10^{-3}; adversarial and VAE weights are γ_{ADV}=10 and γ_{VAE}=1.0 with branch weights γ_{x_0}=γ_{x_t}=1.0. VFS follows the same schedule, uses VAE reconstruction, and enables MMD and center losses together with lightweight discriminator distillation; default weights are γ_{ADV}=10, γ_{VAE}=1.0, $\gamma_{\mathcal{M}}$=0.25, $\gamma_{\mathcal{C}}$=0.25, $\gamma_{\mathcal{R}}$=1.0, and γ_{dist}=5.0. All weights were selected via grid search on the AWA2 validation split and directly transferred to CUB and SUN without further tuning.

B. Results and Analysis

1) The Trade-off between Seen and Unseen Accuracy: Beyond absolute accuracy, the GZSL trade-off between seen and unseen classes is a central challenge. A method may obtain a very high S by favoring seen classes, yet still under-recognize unseen classes (low U), yielding a modest H. We report U, S, and their harmonic mean H, as they provide a comprehensive characterization of unseen generalization and the balance between seen and unseen recognition. As shown in Table I, different methods often exhibit distinct trade-offs. D2SC is designed to stabilize semantic conditioning across diffusion timesteps and to align synthesized and real feature geometry, which aims to improve unseen recognition while maintaining seen-side performance, leading to a more balanced U/S trade-off.

2) Overall comparison with the state-of-the-art: We benchmark on AWA2 [11], CUB [12] and SUN [9]. Under a single set of hyper-parameters, D2SC achieves strong ZSL and balanced GZSL (Table I). On AWA2 [11], the harmonic mean reaches 84.0 with high unseen accuracy; on CUB [12], performance remains competitive; on SUN [9], the main gain concentrates on the unseen side (U = 65.0), indicating improved transfer in the more ambiguous semantic regime.

3) Why gains vary across datasets: Performance improvements in diffusion-conditioned generative ZSL are strongly affected by the discriminability of the semantic space. On datasets with highly correlated class attributes, different classes can be nearly indistinguishable in semantics, which limits how much any conditional generator can improve unseen recognition. Concretely, the maximum cosine similarity between

TABLE I

COMPARISONS WITH THE STATE-OF-THE-ART ON AWA2 [11], CUB [12] AND SUN [9]. ZSL REPORTS TOP-1 ACCURACY (T1, %), AND GZSL REPORTS U (UNSEEN), S (SEEN), AND THEIR HARMONIC MEAN H (%). THE BEST PER COLUMN IS IN **BOLD**. †: RESULTS USING OUR FINE-TUNED FEATURES; ‡: USING OTHER FINE-TUNED FEATURES. RES101 DENOTES RESNET-101; VIT DENOTES VISION TRANSFORMER.

Method	Venue	Backbone	ZSL (T1)			GZSL (AWA2 [11])			GZSL (CUB [12])			GZSL (SUN [9])		
			AWA2 [11]	CUB [12]	SUN [9]	U	S	H	U	S	H	U	S	H
CLIP [13]*	ICML21	ViT	-	-	-	-	-	-	55.2	54.8	55.0	-	-	-
CoOp [14]*	IJCV22	ViT	-	-	-	-	-	-	49.2	63.8	55.6	-	-	-
ICIS [10]	ICCV23	Res101	64.6	60.6	51.8	35.6	**93.3**	51.6	45.8	73.7	56.5	45.2	25.6	32.7
ReZSL [15]	TIP23	Res101	70.9	80.9	63.0	63.8	85.6	73.1	72.8	74.8	73.8	47.4	34.8	40.1
PSVMA [16]	CVPR23	ViT	-	-	-	73.6	77.3	75.4	70.1	77.8	73.8	61.7	45.3	52.3
ZSLViT [17]	CVPR24	ViT	70.7	78.9	68.3	66.1	84.6	74.2	69.4	78.2	73.6	45.9	48.4	47.3
f-CLSWGAN [18] †	CVPR18	Res101	75.9	84.5	75.5	65.1	68.9	66.9	76.4	83.3	79.7	63.8	55.7	59.5
f-VAEGAN [19] †	CVPR19	Res101	75.8	85.1	75.4	67.3	65.6	66.4	77.4	**83.5**	80.3	63.6	54.1	58.4
TFVAEGAN [20] †	ECCV20	Res101	72.4	85.8	74.1	54.2	89.6	67.5	79.0	83.3	81.1	51.8	53.8	52.8
SDVAE [7] †	ICCV21	Res101	69.3	85.1	77.0	57.0	72.3	63.8	**81.6**	74.2	77.7	62.3	56.9	59.5
CEGAN [21] †	CVPR21	Res101	72.8	84.6	74.1	57.1	89.0	69.6	78.9	80.9	79.9	58.1	57.4	57.8
DFCAFlow [8]†	TCSVT23	Res101	74.4	83.9	77.2	67.6	81.0	73.7	77.3	82.9	80.0	63.0	**59.6**	**61.2**
CoOp+SHIP [22]*‡	ICCV23	ViT	-	-	-	-	-	-	55.3	58.9	57.1	-	-	-
DSP [23]‡	ICML23	Res101	-	-	-	60.0	86.0	70.7	51.4	63.8	56.9	48.3	43.0	45.5
DML [24]	ACCV24	Res101	-	-	-	62.2	82.3	70.9	57.1	81.6	67.2	39.6	52.7	45.9
VADS [25]‡	CVPR24	ViT	82.5	86.8	76.3	75.4	83.6	79.3	74.1	74.6	74.3	64.6	49.0	55.7
ZeroDiff [6]†	ICLR25	Res101	87.3	**87.5**	**77.3**	74.7	89.3	81.4	80.0	83.2	**81.6**	63.0	56.9	59.8
D2SC†	Ours	Res101	**87.8**	86.4	76.7	**79.6**	89.0	**84.0**	79.8	78.8	79.3	**65.0**	54.6	59.4

class attribute vectors is 0.9556 on AWA2 [11], but rises to 0.9960 on CUB [12] and 0.9823 on SUN [9]; moreover, CUB [12] and SUN [9] contain "near-duplicate" class pairs with similarity larger than 0.99. In such cases, strengthening timestep consistency helps stabilize conditioning, yet the upper bound is constrained by semantic ambiguity, especially for fine-grained categories.

This observation is consistent with our results: D2SC yields larger gains on AWA2 [11] (where semantics are more discriminative) while maintaining competitive performance on CUB [12]/SUN [9] (e.g., when multiple bird species share almost identical attribute profiles, a stabilized condition reduces "semantic swapping" in synthesized features and improves U without inflating S). Moreover, the fine-grained and data-scarce regimes (e.g., CUB [12] and SUN [9]) make class-conditional alignment more sensitive to semantic ambiguity; in fine-grained regimes, D2SC prioritizes correcting the generative bias toward seen classes, yielding a more balanced GZSL trade-off rather than overfitting to seen categories.

For ablations, removing MMD and center losses primarily degrades unseen coverage while leaving seen accuracy nearly unchanged (Table II). These results indicate that the two terms align global statistics and compact class-wise geometry, lifting U and H without sacrificing S.

C. Efficiency and complexity analysis

In addition to accuracy, we evaluate computational complexity and empirical efficiency under the same backbone and batch size. As summarized in Table III, D2SC has comparable model complexity to ZeroDiff [6] (only +0.64M parameters and +0.041 GFLOPs at batch size 64); the added regularization and alignment objectives introduce negligible parameter/FLOP overhead. In practice, D2SC reduces GPU memory usage by 28.5% and shortens training time by 8.6% per iteration, which is consistent with using a short diffusion chain (n_T=4) and shared mappings, while keeping the optimization signals coherent via lightweight branch distillation. In our measurements, D2SC also exhibits slightly lower time variability across epochs (coefficient of variation 1.16% vs. 1.31%), suggesting a steadier training profile.

TABLE II

ABLATION STUDY ON MMD AND CENTER LOSS. "FULL D2SC" DENOTES THE COMPLETE MODEL REPORTED IN TABLE I; "W/O MMD & CL" REMOVES BOTH OBJECTIVES.

Setting	ZSL (T1)	U	S	H
SUN [9] (Full D2SC)	76.7	65.0	54.6	59.4
SUN [9] (w/o MMD & CL)	76.0	63.8	53.3	58.1
AWA2 [11] (Full D2SC)	87.8	79.6	89.0	84.0
AWA2 [11] (w/o MMD & CL)	83.8	71.7	89.0	79.4

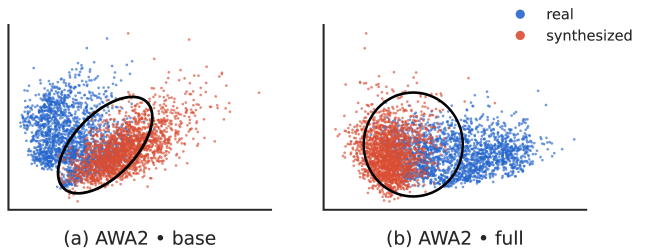

Fig. 2. Qualitative ablation of MMD/center. Left: base (without MMD/center). Right: full D2SC. Full produces tighter class clusters and better mixing between synthesized and real features.

D. Qualitative analysis

To further illustrate the effect of MMD and center loss, Figure 2 visualizes synthesized (red) and real (blue) features via 2D projection. The left panel shows the base model without

TABLE III
MODEL COMPLEXITY AND EMPIRICAL EFFICIENCY (BATCH SIZE 64, SINGLE GPU). TIME: SECONDS PER ITERATION; MEMORY: AVERAGE ALLOCATED GPU MEMORY.

Method	Params (M)	GFLOPs	Time (s/it)	Mem (MB)
D2SC	55.16	3.53	28.15	2484.4
ZeroDiff [6]	54.52	3.49	30.81	3473.2

MMD/center, and the right panel shows the full D2SC. The full model forms tighter class clusters and exhibits higher synthesized–real mixing, indicating improved class compactness and distributional alignment.

Failure cases mainly occur on fine-grained categories with highly overlapping semantics (e.g., multiple bird species in CUB [12] can share nearly identical attribute profiles, making their synthesized clusters overlap even when the marginal distribution is well matched). In such regimes, sharpening class-conditional anchors or adopting class-adaptive τ is helpful (e.g., assigning larger τ to semantically ambiguous classes stabilizes the condition across timesteps and reduces "semantic swapping"), and studying semantic-ambiguity-aware conditioning and alignment is therefore a meaningful direction on its own.

V. CONCLUSION

We introduced D2SC, a diffusion-guided dual-stage semantic-visual coupling framework. CPR stabilizes contrastive representations via timestep consistency, while VFS achieves distributional and geometric alignment in the visual space through distillation, MMD, and center constraints. Using unified hyper-parameters, D2SC delivers consistent gains on unseen classes and in GZSL. Moreover, D2SC maintains comparable model complexity to representative diffusion-based baselines while reducing training memory and time, making it more practical for deployment. Future work includes finer timestep scheduling and class-adaptive center modeling to further enhance generalization.

REFERENCES

[1] P. Bhagat, P. Choudhary, and K. M. Singh, "A study on zero-shot learning from semantic viewpoint," *The Visual Computer*, vol. 39, no. 5, pp. 2149–2163, 2023.

[2] M. G. Atigh, S. Nargang, M. Keller-Ressel, and P. Mettes, "Simzsl: Zero-shot learning beyond a pre-defined semantic embedding space," *International Journal of Computer Vision*, pp. 1–17, 2025.

[3] S. Chen, Z. Hong, X. You, and L. Shao, "Semantics-conditioned generative zero-shot learning via feature refinement," *International Journal of Computer Vision*, pp. 1–18, 2025.

[4] H. Huang, X. Qiao, Z. Chen, H. Chen, B. Li, Z. Sun, M. Chen, and X. Li, "Crest: Cross-modal resonance through evidential deep learning for enhanced zero-shot learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 5181–5190.

[5] L. Meegahapola, H. Hassoune, and D. Gatica-Perez, "M3bat: Unsupervised domain adaptation for multimodal mobile sensing with multi-branch adversarial training," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–30, 2024.

[6] Z. Ye, S. N. Gowda, X. Huang, H. Xu, Y. Jin, K. Huang, and X. Jin, "Zerodiff: Solidified visual-semantic correlation in zero-shot learning," *arXiv preprint arXiv:2406.02929*, 2024.

[7] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8712–8720.

[8] H. Su, J. Li, K. Lu, L. Zhu, and H. T. Shen, "Dual-aligned feature confusion alleviation for generalized zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[9] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2751–2758.

[10] A. Christensen, M. Mancini, A. S. Koepke, O. Winther, and Z. Akata, "Image-free classifier injection for zero-shot classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 19 072–19 081.

[11] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.

[12] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-ucsd birds 200," 2010.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[15] Z. Ye, G. Yang, X. Jin, Y. Liu, and K. Huang, "Rebalanced zero-shot learning," *IEEE Transactions on Image Processing*, 2023.

[16] M. Liu, F. Li, C. Zhang, Y. Wei, H. Bai, and Y. Zhao, "Progressive semantic-visual mutual adaption for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 337–15 346.

[17] S. Chen, W. Hou, S. Khan, and F. S. Khan, "Progressive semantic-guided vision transformer for zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 964–23 974.

[18] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551.

[19] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-vaegan-d2: A feature generating framework for any-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 275–10 284.

[20] S. Narayan, A. Gupta, F. S. Khan, C. G. Snoek, and L. Shao, "Latent embedding feedback and discriminative features for zero-shot classification," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 2020, pp. 479–495.

[21] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2371–2381.

[22] Z. Wang, J. Liang, R. He, N. Xu, Z. Wang, and T. Tan, "Improving zero-shot generalization for clip with synthesized prompts," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3032–3042.

[23] S. Chen, W. Hou, Z. Hong, X. Ding, Y. Song, X. You, T. Liu, and K. Zhang, "Evolving semantic prototype improves generative zero-shot learning," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 4611–4622.

[24] C. Zhang, M. Jin, Q. Yu, H. Xue, S. N. Gowda, and X. Jin, "Bridging the projection gap: Overcoming projection bias through parameterized distance learning," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 3327–3343.

[25] W. Hou, S. Chen, S. Chen, Z. Hong, Y. Wang, X. Feng, S. Khan, F. S. Khan, and X. You, "Visual-augmented dynamic semantic prototype for generative zero-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 23 627–23 637.