ZeroDiff++: Balancing Semantic Diffusion Dynamics for Robust Zero-Shot Learning

Supplementary Material

1. Experimental Design for Verifying SDI

We provide a detailed explanation of the experimental procedure used to visually and quantitatively verify the SDI phenomenon described in Fig. 1 of the main paper. To isolate the imbalance caused by a single fixed diffusion schedule, all results in this section are obtained using the trained ZeroDiff baseline.

Measuring Class Complexity and Selecting Two Classes. We use the 2048-dimensional features generated by the DFG of trained ZeroDiff to compute the intra-class scatter of each class on AWA2:

$$\text{Scatter}_c = \frac{1}{N}\sum_{n=1}^{N}\left\|\mathbf{v}^{(n)} - \boldsymbol{\mu}_c\right\|_2, \tag{1}$$

where $\boldsymbol{\mu}_c = \frac{1}{N}\sum_{n=1}^{N}\mathbf{v}^{(n)}$. We use intra-class scatter as a proxy for class complexity, and then select

$$c_{\text{simple}} = \arg\min_c \text{Scatter}_c, \tag{2}$$

$$c_{\text{complex}} = \arg\max_c \text{Scatter}_c. \tag{3}$$

Sampling Features Across Diffusion Steps. For each selected class, we first sample 200 random latent variables and use the DRG of trained ZeroDiff, conditioned on the class attribute embedding, to generate semantic representations. Then, conditioned on the generated representation, the class attribute embedding, the current diffusion timestep, and the noisy feature, the DFG performs a four-step reverse diffusion process to produce the final clean feature together with its intermediate states. All generated features of c_{simple} and c_{complex} at each diffusion step are saved for subsequent scatter computation and visualization.

PCA + t-SNE Visualization. For each diffusion step, we visualize the feature distributions of c_{simple} and c_{complex} using PCA followed by t-SNE. Specifically, the original 2048-dimensional features are first reduced to 30 dimensions using PCA while retaining more than 95% of the variance. The resulting features are then embedded into two dimensions using t-SNE with a perplexity of 25 and a learning rate of 150. In the visualization, features of c_{simple} are shown in blue, whereas those of c_{complex} are shown in red. For both classes, we additionally plot the 95% confidence ellipses to illustrate their dispersion patterns. These visualizations clearly reveal that the simple class tends to collapse rapidly into compact clusters, while the complex class remains widely dispersed under the same diffusion schedule, highlighting the structural imbalance induced by a fixed schedule.

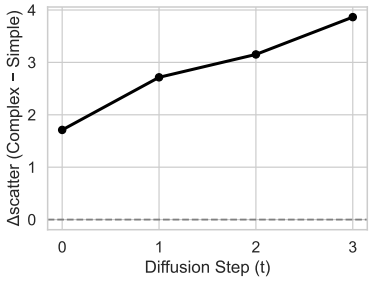

Figure 1. The curve of Δ_{scatter} during the four-step reverse diffusion.

We also compute the SDI gap curve:

$$\Delta_{\text{scatter}}(t) = \text{Scatter}_{c_{\text{complex}}}(t) - \text{Scatter}_{c_{\text{simple}}}(t). \tag{4}$$

Fig. 1 shows the curve of Δ_{scatter}. The gap in intra-class scatter between c_{simple} and c_{complex} increases continuously with diffusion step t, indicating that their denoising behaviors progressively diverge and thus exhibit markedly different denoising dynamics.

2. W-distillation Loss Calculation

Let $W_a = \mathbb{E}[D'_a(\mathbf{v}_0^s, \mathbf{a}^s)] - \mathbb{E}[D'_a(\tilde{\mathbf{v}}_0^s, \mathbf{a}^s)]$ and $W_d = \mathbb{E}\big[D'_d(\mathbf{v}_{t-1}^s, \mathbf{v}_t^s, \mathbf{r}_0^s, \mathbf{a}^s, t)\big] - \mathbb{E}\big[D'_d(\hat{\mathbf{v}}_{t-1}^s, \mathbf{v}_t^s, \mathbf{r}_0^s, \mathbf{a}^s, t)\big]$ denote the Wasserstein distances in the original ZeroDiff. Similarly, we define the branch-specific Wasserstein distances for the W- and M-Branches as

$$W_w = \mathbb{E}[D'_a(\mathbf{v}_0^s, \mathbf{a}^s)] - \mathbb{E}[D'_a(\tilde{\mathbf{v}}_0^{s,w}, \mathbf{a}^s)], \tag{5}$$

$$W_m = \mathbb{E}\big[D'_d(\mathbf{v}_{t-1}^{s,m}, \mathbf{v}_t^{s,m}, \mathbf{r}_0^s, \mathbf{a}^s, t)\big] - \mathbb{E}\big[D'_d(\hat{\mathbf{v}}_{t-1}^{s,m}, \mathbf{v}_t^{s,m}, \mathbf{r}_0^s, \mathbf{a}^s, t)\big]. \tag{6}$$

The corresponding W-distillation losses are defined as

$$\mathcal{L}_{\text{mu}}^w = (\kappa_{w,t})^\gamma \left(\|W_w - W_a\|_1 + \|W_w - W_d\|_1\right), \tag{7}$$

$$\mathcal{L}_{\text{mu}}^m = (\kappa_{m,t})^\gamma \left(\|W_m - W_d\|_1 + \|W_m - W_w\|_1\right), \tag{8}$$

where $\kappa_{w,t} = 1 - \sqrt{\prod_{i=1}^{t}(1-\beta_i^w)}$ and $\kappa_{m,t} = 1 - \sqrt{\prod_{i=1}^{t}(1-\beta_i^m)}$ denote the Noise-to-Data (N2D) ratios for the W- and M-Branches, respectively, and γ is a smoothing factor.