Figure 1: **Visualization of the object-centric training set in the AbC benchmark.** The image on the left shows a real exemplar, while the 20 images on the right illustrate a subset of its paired synthetic counterparts. In the AbC benchmark, the quality of synthetic data varies significantly. Some images, such as those outlined in black at the bottom, provide strong supervisory signals. In contrast, many others—such as those highlighted in red at the top—are noisy or ambiguous, potentially hindering effective training. These 20 examples represent only a small portion of the hundreds of synthetic images generated per exemplar and do not fully capture the overall variability and noise present in the dataset.

Figure 2: **Visualization of attribution results.** Given a synthetic image generated by CustomDiffusion (based on Stable Diffusion), we retrieve its exemplars from a pool of 1 million LAION images. Alignment denotes attribution using the pretrained DINO model. Supervised refers to DINO finetuned with paired training data from the AbC benchmark. A&D is our proposed method, which performs unsupervised finetuning on DINO. Red bounding boxes indicate the ground-truth exemplars.

Source		ImageNet-Seen				BAM-FG				ImageNet-Unseen				Artchive				
Prompts		GPT		Media		GPT		Object		GPT		Media		GPT		Object		**Average**
F_{base}	Method	R@5	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5	mAP
ViT-B	Alignment	0.355	0.310	0.242	0.210	0.168	0.193	0.224	0.259	0.785	0.726	0.474	0.410	0.201	0.210	0.156	0.161	0.326	0.310
	A&D	0.440	0.395	0.295	0.259	0.203	0.229	0.251	0.288	0.845	0.805	0.561	0.500	0.229	0.240	0.176	0.182	**0.375**	**0.362**
ViT-L	Alignment	0.336	0.288	0.218	0.184	0.116	0.126	0.161	0.178	0.707	0.639	0.415	0.354	0.143	0.142	0.117	0.112	0.277	0.253
	A&D	0.349	0.299	0.229	0.195	0.117	0.127	0.160	0.178	0.716	0.647	0.432	0.368	0.142	0.139	0.114	0.111	**0.282**	**0.258**

Table 1: **Attribution results for ViT with ViT-B and ViT-L** ViT-B and ViT-L refer to vit_base_patch16_384 and vit_large_patch32_384, respectively. We report retrieval performance of two methods: Alignment (zero-shot pretrained ViT) and A&D (Alignment plus Disentanglement) across all the test sets of AbC benchmark. Despite the larger capacity of ViT-L, it does not consistently outperform ViT-B in the alignment-only setting, highlighting that scaling the model size alone is insufficient for robust attribution. However, our proposed A&D method consistently improves performance for both backbones, demonstrating its effectiveness in enhancing attribution quality without requiring supervision.

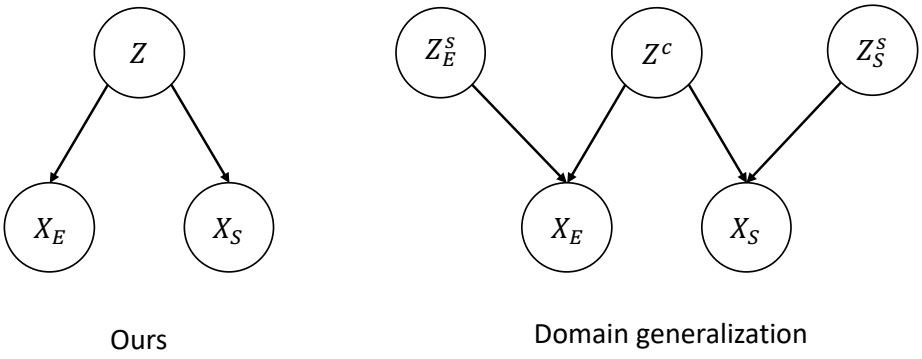

Figure 3: **Causal graph of data generation processes in our CCA setting versus the domain generalization setting.**