CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts

APPENDIX

Yichao Cai, Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi

Australian Institute for Machine Learning, University of Adelaide, SA 5000, Australia
{yichao.cai,yuhang.liu01,zhen.zhang02,javen.shi}@adelaide.edu.au

Overview of the Appendix:

- More details on experiments using the CLIP pre-trained ViT-B/16 model are provided in Appendix A, including implementation details in Appendix A.1, investigations into prompt augmentation combinations in Appendix A.2, analysis of different training prompt sources in Appendix A.3, and detailed experiment results for each dataset in Appendix A.4.
- The processes of data synthesis with large models used in our approach are outlined in Appendix B: The image synthesis procedure for Im.Aug is detailed in Appendix B.1, and the approach for generating "LLM" prompts, used in analyzing prompt sources, is described in Appendix B.2.
- In Appendix C, we detail our repeated zero-shot experiments conducted with the CLIP pre-trained ViT-L/14 (Appendix C.1) and ResNet50x16 (Appendix C.2) models.
- In section Appendix D, we present discussions covering the underlying rationale for basing CLAP on the CLIP pre-trained models in Appendix D.1, and the impact of image augmentation and text augmentation in Appendix D.2.

A More on Experiments with ViT-B/16

A.1 Implementation Details

In this section, we detail the implementation of our experiments utilizing the CLIP pre-trained ViT-B/16 model:

Network. The network's output dimension is aligned with the 512-dimensional CLIP features, thereby obviating the need for input feature downsampling. The latent dimensions are tailored to each dataset: 256 for PACS, 448 for OfficeHome, and 512 for VLCS and DomainNet, to accommodate the variety of categories and complexity of datasets. The weight parameter α is adjusted to 0.208 for PACS, 0.056 for VLCS, 0.14 for OfficeHome, and 0.2 for DomainNet, while it is consistently maintained at 1 throughout the training phase.

Training CLAP. Training parameters are consistent across datasets, employing the Adam optimizer with a learning rate of 0.0001, limiting training to 8,000 steps with checking the average loss every 480 steps, and instituting early stopping after five checkpoints without a loss decrease of at least 0.01. Batch sizes are adjusted to 8 for PACS and VLCS, 96 for OfficeHome, and 384 for DomainNet, with the temperature parameter τ set at 0.5 for PACS and VLCS, and 0.3 for OfficeHome and DomainNet. The loss coefficient λ is set to 1 for PACS and VLCS, and 0.0001 for OfficeHome and DomainNet, due to the first two datasets have less classes. Prompt augmentations, OSD+OCD+SPO, are applied across datasets all with a 0.5 probability. For the PACS and VLCS datasets, Gaussian noise with a zero mean and a standard deviation of 0.02 is randomly inserted at the beginning, middle, or end of the augmented-view prompts to enrich the training samples. In the linear probe evaluations for few-shot analysis, L2 normalization and cross-entropy loss are utilized for training over 1,000 epochs with a batch size of 32, incorporating early stopping with a patience threshold of 10 epochs and a loss decrease criterion of 0.001.

Training Im.Aug. We train a disentangled network using image augmentation, applying the InfoNCE loss with a temperature parameter τ set to 0.5. This include image augmentation techniques, image cropping ($scale \in [0.64, 1.0]$) and color distortion ($brightness = 0.5, hue = 0.3$), each with a probability of 0.5. Other training and inference configurations for Im.Aug are consistent with those used for CLAP across all datasets.

A.2 Prompt Augmentation Combinations

In Tab. 1, we explore different combinations of our tailored prompt augmentation techniques and EDA (Easy Data Augmentation) [42] techniques on the VLCS dataset. Each combination demonstrates CLAP's effectiveness in enhancing CLIP's performance and reducing performance disparities. The combination of OSD+OCS+SPO+IGN achieves the highest average accuracy and the least variance, outperforming the EDA techniques. Notably, even without incorporating random noise in the augmentations, CLAP significantly surpasses CLIP in handling perturbations on prompts, as evidenced by the largely reduced $\Delta_{(NC)}$.

A.3 Prompt Sources

In Tab. 2, we examine the effects of various training prompt formats, sourced from different synthetic origins, on the VLCS dataset performance, utilizing

EDA techniques. The prompt formats are defined as follows: "Template" refers to the template-based prompts fundamental to our primary approach; "LLM" designates prompts created by ChatGPT-3.5 [3], with the generation process elaborated in Appendix B.2; "Random" describes prompts formatted as "a [random] style of [class]," with "[random]" being replaced by terms from a random word generator; and "Prm.Stl." indicates vectorized prompts generated through PromptStyler [9].

Table 1: We evaluate prompt augmentation combinations on the VLCS dataset: OSD (①), OCD (②), ITD (③), ASD (④), SPO (⑤), and IGN (⑥). ZS(Avg.) shows average zero-shot accuracy acoss four distinct inference prompts. CLAP boosts CLIP's accuracy and reduces variances, with ①②⑤⑥ as the optimal combination.

Metrics	CLIP (base)	Avg. top-1 acc. (%) of different augmentations							
		EDA	①②③④⑤⑥	①②③④⑤	①②③④⑥	①②③④	③④⑤⑥	①②⑤⑥
ZS(Avg.) (↑)	77.3	81.6	82.0	80.1	82.0	79.6	82.1	**82.6**
R (↓)	6.1	1.9	1.2	2.5	0.9	3.2	1.6	**0.8**
δ (↓)	2.8	0.9	0.6	1.2	**0.4**	1.5	0.7	**0.4**
$\Delta_{(NC)}$ (↓)	8.1	2.3	1.7	3.0	1.8	3.4	2.0	**1.6**

Table 2: We employ EDA augmentation to train CLAP with diverse prompt sources on the VLCS dataset. Each prompt source contributes to improvements in CLIP's zero-shot performance, with "Random" and "Template" prompts, in their simpler forms, yielding better outcomes.

Metrics	CLIP (base)	Avg. top-1 acc. (%) of different sources			
		LLM	Random	Prm.Stl.	Template
ZS(Avg.) (↑)	77.3	78.2	**81.6**	81.2	**81.6**
R (↓)	6.1	3.2	**0.7**	2.7	1.9
δ (↓)	2.8	1.5	**0.3**	1.2	0.9
$\Delta_{(NC)}$ (↓)	8.1	3.3	**2.3**	3.0	**2.3**

Our experimental results indicate that CLAP, when trained across these varied prompt formats, enhances the performance of CLIP. Notably, despite the complex generation mechanisms of "LLM" and "Prm.Stl." prompts, the simpler, random-styled and template-based prompts demonstrate superior efficacy. However, it is important to highlight that the improvements attributed to these diverse prompt formats, trained with EDA, do not surpass the best performance of the prompt augmentations tailored for template-based prompts.

A.4 Detailed Results on ViT-B/16

Details on Zero-Shot Evaluations We present the domain-level zero-shot performance with various prompts across each dataset in Tab. 3. CLAP consistently enhances CLIP's zero-shot performance across these different prompts. Given that CLAP exclusively utilizes text data for training, it does not compromise CLIP's inherent ability to generalize across domains, which is acquired from its extensive training dataset. Rather, by achieving a more effective disentanglement of content, it unequivocally enhances CLIP's zero-shot performance across all dataset domains.

Table 3: Domain-level zero-shot results of the ViT-B/16 model on the test datasets.

Dataset Domains		Domain-level avg. top-1 acc. (%) of zero-shot performance usig ViT-B/16 (↑)											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	96.4	96.9	97.5	93.4	97.0	97.6	97.4	97.6	97.6	87.8	93.5	97.1
	C.	98.9	99.0	98.9	99.0	99.2	99.0	99.1	99.0	98.9	95.4	97.6	98.8
	P.	99.9	99.9	99.9	99.3	99.6	99.9	99.9	99.9	99.9	93.1	99.0	99.9
	S.	87.7	90.1	92.5	89.2	89.6	92.5	88.1	89.4	92.3	87.1	89.3	93.1
VLCS	C.	99.7	99.8	99.9	99.9	99.9	99.9	99.9	99.9	99.9	87.0	96.0	99.9
	L.	61.8	66.2	67.7	69.9	70.4	70.4	70.2	70.2	70.7	55.9	59.9	65.9
	S.	70.1	74.8	78.0	73.3	76.0	77.2	73.6	76.4	76.9	61.4	66.2	75.3
	V.	73.9	77.1	84.9	84.8	85.4	86.0	86.1	85.6	86.2	68.9	70.3	82.9
OfficeHome	A.	80.5	79.0	81.8	80.1	76.0	81.6	83.2	78.7	83.2	73.0	69.2	73.6
	C.	64.6	59.6	66.4	63.7	58.9	65.4	68.1	61.9	69.0	57.0	52.0	60.4
	P.	86.3	83.6	87.5	86.6	83.4	87.2	89.1	86.6	89.7	77.2	72.3	78.9
	R.	88.0	85.9	88.5	87.6	84.8	87.7	89.8	87.2	90.0	79.0	76.5	81.1
DomainNet	C.	71.0	64.3	71.9	70.5	62.1	72.0	71.3	63.4	72.8	63.2	53.9	64.6
	I.	48.6	40.5	50.6	47.7	40.7	49.5	47.8	40.0	50.5	42.9	35.0	45.1
	P.	66.6	59.1	67.7	66.0	59.0	67.3	66.5	59.8	68.4	57.2	50.4	59.4
	Q.	14.9	12.4	15.2	13.3	11.5	13.8	14.1	11.8	14.3	12.0	9.2	13.1
	R.	82.6	76.6	83.1	82.2	75.8	82.2	83.4	78.2	83.7	75.2	67.9	75.6
	S.	63.1	56.1	63.7	62.2	55.0	63.1	63.4	56.4	64.4	55.7	47.5	57.6

Details on Few-Shot Evaluations We display the quantitative results of few-shot performance in Tab. 4. CLAP consistently enhances the few-shot capabilities, showcasing improvements across test datasets at a closer domain level.

Details on Adversarial Evaluations In Tab. 5, we detail our adversarial performance evaluations for PACS, VLCS, OfficeHome, and DomainNet, respectively. CLAP enhances both zero-shot and one-shot performance across all domains of the tested datasets. While Im.Aug boosts one-shot robustness against adversarial tasks, its impact on zero-shot adversarial robustness is inconsistent.

Details on Ablative Analysis In Tab. 6, we provide detailed results from our analysis on zero-shot performance using various combinations of prompt augmentations. Additionally, in Tab. 7, we present the outcomes of our ablative studies focusing on the hyperparameters τ, latent dimension, and α, respectively, each evaluated domain-wise. The results indicate that CLAP is effective across a wide range of hyperparameters.

B Data Synthesis

B.1 Synthetic Image Generation

We employ the stable diffusion [39] v2.1 model for generating synthetic images used in our comparing experiments, specifically utilizing the Stable Diffusion

Table 4: Domain-level few-shot results of the ViT-B/16 model using the test datasets.

Dataset	Domains	Domain-level avg. top-1 acc. (%) of few-shot performance of ViT-B/16 (↑)														
		1-shot			4-shot			8-shot			16-shot			32-shot		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	79.5	84.1	94.5	92.4	96.4	97.2	95.1	97.2	98.4	97.9	98.1	98.4	98.8	99.1	98.9
	C.	86.7	96.1	98.3	96.8	98.6	99.2	98.8	98.9	99.3	99.5	99.2	99.5	99.6	99.6	99.6
	P.	97.4	99.8	99.9	99.6	99.8	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9	99.9
	S.	75.1	80.0	87.3	91.1	92.3	92.5	92.3	92.3	92.9	92.4	92.6	93.1	93.9	94.2	94.1
VLCS	C.	99.2	99.7	99.8	99.9	99.8	99.9	99.8	99.7	99.9	99.7	99.9	99.9	99.9	100.0	99.9
	L.	41.3	41.3	41.1	56.7	57.0	59.8	46.2	36.8	48.3	59.4	60.4	62.6	60.4	60.7	61.9
	S.	45.3	46.1	50.8	61.9	63.7	69.0	67.4	67.7	71.3	75.9	76.8	80.9	77.4	78.6	81.0
	V.	50.9	53.4	59.0	64.5	66.7	76.1	75.4	74.1	78.7	72.6	73.9	77.7	85.7	86.1	87.9
OfficeHome	A.	42.6	45.1	43.9	76.8	77.6	77.7	84.8	86.0	85.5	91.8	92.1	92.1	97.4	97.5	97.5
	C.	40.1	45.0	43.8	69.9	70.2	70.5	75.8	75.9	76.6	81.6	81.6	81.6	89.0	89.0	89.2
	P.	70.2	73.3	73.4	89.7	90.3	90.2	93.8	93.7	93.9	95.7	95.7	95.8	97.7	97.6	97.6
	R.	58.4	59.3	59.4	81.7	83.1	82.9	89.7	89.5	89.9	92.9	92.7	93.2	95.8	95.8	95.8
DomainNet	C.	42.1	43.6	43.8	66.8	67.5	67.8	74.2	74.3	74.6	78.5	78.6	78.8	82.8	82.8	82.7
	I.	19.5	20.8	21.0	38.5	39.3	39.7	46.7	47.0	47.3	53.2	53.2	53.6	60.0	59.9	60.1
	P.	32.1	33.5	34.2	60.5	60.9	61.5	68.0	68.0	68.7	72.5	72.6	73.0	76.7	76.6	76.8
	Q.	15.2	15.3	15.3	30.0	29.6	29.9	37.1	36.4	36.8	43.8	43.4	43.5	49.4	49.1	49.0
	R.	50.8	52.1	52.7	76.7	77.0	77.6	81.7	81.9	82.2	84.0	83.9	84.3	85.9	85.9	86.0
	S.	33.1	33.9	34.8	56.2	56.6	57.2	62.9	62.9	63.7	67.8	67.7	68.1	72.5	72.3	72.6

Fig. 1: Examples of synthetic images created with SDv2.1 and associated prompts.

v2-1 Model Card available on Hugging Face[1]. For each class across the four datasets, we produce 480 images using our synthetic template prompts as input for the stable diffusion model. All generated images are of 512×512 resolution. Examples of these synthetic images alongside their corresponding text prompts are displayed in Fig. 1.

B.2 LLM Prompts Generation

We utilize ChatGPT-3.5 [3] to create the LLM prompts employed in our comparative analysis of different prompt sources. Fig. 2 illustrates the process of prompting ChatGPT-3.5 to generate text prompts for specific class names. For each class, we produce 120 samples, and below are a few examples from the generated prompts:

– Bird:

[1] `https://huggingface.co/stabilityai/stable-diffusion-2-1`

Table 5: Domain-level results under adversarial attacks of ViT-B/16 on the datasets.

Dataset	Domains	Domain-level avg. top-1 acc. (%) under adversarial attackings using ViT-B/16 (↑)																	
		FGSM						PGD-20						CW-20					
		ZS-C			1-shot			ZS-C			1-shot			ZS-C			1-shot		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	76.3	79.3	79.3	61.2	78.0	87.3	1.7	2.2	1.8	16.0	42.1	63.1	1.5	2.0	2.3	0.5	1.1	1.7
	C.	94.9	95.0	94.0	66.5	84.2	95.1	33.3	37.7	35.6	33.3	57.2	86.1	28.8	34.0	33.2	11.9	23.6	31.8
	P.	91.6	90.3	91.7	67.4	80.8	92.1	5.7	7.0	6.7	27.1	55.0	69.8	4.7	4.9	5.8	0.7	2.7	4.1
	S.	84.5	87.5	89.8	71.6	74.6	83.8	75.8	78.4	79.2	63.0	66.3	74.6	74.5	76.8	77.9	62.7	65.4	70.3
VLCS	C.	55.3	53.8	55.5	25.8	28.8	25.3	4.4	5.1	4.7	2.0	5.2	2.5	2.9	3.1	3.5	0.7	1.2	1.0
	L.	49.4	45.5	50.6	27.0	32.6	30.4	15.2	14.9	16.0	6.4	8.9	8.0	12.4	11.2	13.0	6.1	8.3	7.7
	S.	61.7	58.1	62.5	48.0	46.9	51.6	13.2	13.9	14.0	8.6	10.7	10.0	9.2	8.8	10.2	8.3	7.9	8.4
	V.	65.3	63.2	65.6	36.5	40.1	41.0	7.5	7.9	7.9	5.3	9.4	8.9	5.2	4.8	5.6	2.9	2.8	2.9
OfficeHome	A.	55.3	53.8	55.5	25.8	28.8	25.3	4.4	5.1	4.7	2.0	5.2	2.5	2.9	3.1	3.5	0.7	1.2	1.0
	C.	49.4	45.5	50.6	27.0	32.6	30.4	15.2	14.9	16.0	6.4	8.9	8.0	12.4	11.2	13.0	6.1	8.3	7.7
	P.	61.7	58.1	62.5	48.0	46.9	51.6	13.2	13.9	14.0	8.6	10.7	10.0	9.2	8.8	10.2	8.3	7.9	8.4
	R.	65.3	63.2	65.6	36.5	40.1	41.0	7.5	7.9	7.9	5.3	9.4	8.9	5.2	4.8	5.6	2.9	2.8	2.9
DomainNet	C.	57.8	50.9	58.8	33.3	34.3	35.0	21.6	18.7	22.8	18.4	19.6	20.0	15.8	12.5	16.6	7.0	7.5	7.8
	I.	35.8	28.0	37.0	12.2	13.3	13.2	6.1	3.7	6.7	4.6	5.3	5.1	3.3	1.9	3.7	0.9	0.9	0.9
	P.	43.9	39.0	44.3	18.4	20.6	20.3	3.1	2.8	3.3	8.6	10.4	9.9	1.8	1.3	1.9	0.3	0.3	0.3
	Q.	12.9	10.3	13.2	10.9	10.8	11.1	8.4	6.8	8.6	5.4	5.4	5.6	7.1	5.4	7.4	4.9	4.8	5.1
	R.	62.1	55.9	62.4	34.5	35.9	36.5	7.1	6.5	7.5	17.6	19.7	19.6	4.5	3.4	4.7	1.2	1.4	1.4
	S.	49.1	43.3	49.7	25.7	26.0	27.5	17.8	15.5	18.6	13.6	14.4	15.1	13.4	10.2	13.9	5.0	5.2	5.6

- A pair of vibrant macaws converse in a lush, tropical rainforest, depicted in a lively, exotic wildlife painting.
- A solitary eagle watches over a vast, rugged canyon at sunrise, portrayed in a majestic, wilderness landscape photograph.

– Dog:

- A sleek Whippet races in a competitive dog track, illustrated in a fast-paced, dynamic sports style.
- A sturdy and reliable English Bulldog watching over a small shop, its solid presence reassuring to the owner.

– Car:

- A quirky art car parades through the streets in a colorful festival, captured in a fun, expressive style illustration.
- A high-tech, autonomous car maneuvers through a smart city environment, portrayed in a futuristic, sci-fi digital art piece.

– Chair:

- A folding chair at an outdoor wedding, elegantly decorated and part of a beautiful ceremony.
- A high-end executive chair in a law firm, projecting authority and professionalism.

– Person:

- An energetic coach motivates a team on a sports field, illustrated in an inspiring, leadership-focused painting.
- A graceful figure skater glides across an ice rink, captured in a delicate, winter-themed pastel drawing.

Table 6: Zero-Shot Performance on VLCS Dataset Across Varied Augmentation Combinations and Prompt Sources: ① Random Object Size Deletion, ② Random Object Color Deletion, ③ Random Image Type Deletion, ④ Random Art Style Deletion, ⑤ Random Swapping Order, ⑥ Addition of Gaussian Noise.

Method	Domains	Avg. top-1 acc. (%) (↑) of different augmentations and prompts on VLCS										
		CLIP (base)	①②③④⑤⑥	①②③④⑤	①②③④⑥	①②③④	③④⑤⑥	①②⑤⑥	EDA LLM	EDA Rand.	EDA Pr.St.	EDA Temp.
ZS(C)	C.	99.7	99.9	99.8	99.9	99.8	99.9	99.9	97.9	99.7	99.9	99.9
	L.	61.8	66.6	62.3	67.0	62.2	66.2	67.7	66.2	69.0	67.3	66.5
	S.	70.1	78.1	75.5	78.0	74.3	78.5	78.0	73.2	76.9	73.5	76.9
	V.	73.9	82.8	80.6	83.2	79.3	82.7	84.9	72.6	81.8	81.8	81.9
ZS(CP)	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9
	L.	69.9	69.3	67.9	69.6	68.4	70.0	70.4	69.3	70.4	71.2	69.7
	S.	73.3	77.6	76.4	76.7	75.9	78.8	77.2	76.2	75.2	75.1	78.0
	V.	84.8	85.3	84.0	85.3	84.2	85.1	86.0	77.0	84.2	86.0	84.6
ZS(PC)	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
	L.	70.2	70.0	68.0	70.1	68.5	70.0	70.7	67.5	70.6	71.8	70.0
	S.	73.6	76.6	75.6	76.0	74.8	77.8	76.9	76.9	75.1	74.9	78.2
	V.	86.1	85.7	84.7	85.7	84.5	85.5	86.2	78.2	84.6	86.8	84.8
ZS(NC)	C.	87.0	99.8	99.6	99.8	99.4	99.7	99.9	95.3	98.6	99.6	99.8
	L.	55.9	65.2	61.3	65.6	60.5	65.4	65.9	63.0	66.7	64.0	64.7
	S.	61.4	75.6	70.3	75.2	68.3	74.9	75.3	68.9	73.3	69.8	73.0
	V.	68.9	80.1	75.2	80.4	73.8	79.4	82.9	69.3	79.6	77.2	78.6

C Experiments on Other CLIP Model Scales

C.1 Experiments on ViT-L/14

We refined the output dimension to align with the input dimension of 768. The chosen latent dimensions were 448 and 640 for PACS and VLCS, respectively, and 768 for both OfficeHome and DomainNet. The inference weighting α was set to 0.1 for PACS, 0.03 for VLCS, 0.14 for OfficeHome, and 0.2 for DomainNet. All other training configurations remained consistent with the ViT-B/16 experiments across each dataset. The training configuration for Im.Aug was set the same as CLAP for each dataset, with the inference weighting α being 0.1 for PACS and 0.03 for the other three datasets.

Table 8 showcases the zero-shot results for the ViT-L/14 model using four distinct prompts, following the protocol established for the ViT-B/16 experiments. These results demonstrate that CLAP is more efficient than Im.Aug in enhancing zero-shot performance. Moreover, Tab. 9 illustrates that CLAP significantly reduces variations in zero-shot performance across different prompts, thereby confirming CLAP's performance improvements over CLIP across a range of model sizes. Detailed domain-level results are presented in Tab. 10, offering an in-depth analysis.

C.2 Experiments on ResNet50x16

To validate our approach on different model structures, we repeated zero-shot experiments on the ResNet50x16 model pre-trained with CLIP. Since the output

Table 7: Ablative study of hyperparameters on VLCS dataset using ViT-B/16 model.

Hyper-parameters	Value	Avg. top-1 acc. (%) (↑) using ViT-B/16 on VLCS dataset											
		ZS (C)				ZS (CP)				ZS (PC)			
		C.	L.	S.	V.	C.	L.	S.	V.	C.	L.	S.	V.
τ	0.1	99.9	67.6	77.5	84.2	99.9	70.9	74.9	85.9	99.9	71.2	74.6	86.3
	0.3	99.9	66.3	77.2	82.4	99.9	69.9	76.7	85.2	99.9	69.9	76.4	85.4
	0.5	99.9	67.7	78.0	84.9	99.9	70.4	77.2	86.0	99.9	70.7	76.9	86.2
	0.7	99.9	65.9	77.7	83.1	99.9	68.9	77.9	84.9	99.9	69.6	77.7	85.0
	0.9	99.9	66.0	77.6	83.3	99.9	69.0	77.9	85.0	99.9	69.7	77.5	85.0
Lantent dim.	128.0	99.9	66.0	77.6	82.6	99.9	70.0	77.4	85.4	99.9	70.1	77.1	85.7
	192.0	99.9	64.9	77.9	83.0	99.9	68.9	78.0	85.6	99.9	69.0	77.8	86.0
	256.0	99.9	63.8	77.6	82.7	99.9	67.6	78.7	84.8	99.9	67.8	78.6	85.2
	320.0	99.9	66.0	77.8	82.9	99.9	69.2	78.1	85.3	99.9	69.7	77.7	85.5
	384.0	99.9	65.8	76.9	82.8	99.9	69.4	77.5	85.3	99.9	69.6	77.0	85.5
	448.0	99.9	65.8	77.4	82.1	99.9	69.7	77.6	84.9	99.9	69.9	77.1	85.6
	512.0	99.9	67.7	78.0	84.9	99.9	70.4	77.2	86.0	99.9	70.7	76.9	86.2
α	$10^{-1.5}$	99.9	66.5	77.9	83.1	99.9	70.4	77.1	86.0	99.9	70.3	76.6	86.1
	10^{-1}	99.9	69.5	77.5	85.7	99.9	70.4	77.1	86.2	99.9	70.9	76.5	86.1
	$10^{-0.5}$	99.9	70.6	75.2	85.5	99.9	70.7	75.7	85.9	99.9	71.0	75.1	85.7
	10^{0}	99.8	71.5	73.5	83.5	99.9	71.7	74.4	85.8	99.8	72.3	73.5	85.5
	$10^{0.5}$	99.8	72.0	73.1	85.5	99.8	72.2	73.7	85.7	99.8	72.5	72.9	85.6
	10^{1}	99.8	72.1	72.8	85.4	99.8	72.3	73.4	85.7	99.8	72.5	72.9	85.5
	$10^{1.5}$	99.8	72.1	72.8	85.4	99.8	72.2	73.3	85.7	99.8	72.6	72.7	85.5

dimension of CLIP is the same as ViT-B/16, we used the same training configuration as ViT-B/16 for training Im.Aug and CLAP. For inference, we refined the weighting coefficient α to 0.1, 1, 0.03, and 0.1 for Im.Aug, and 0.03, 0.2, 0.06, and 0.1 for CLAP, for PACS, VLCS, OfficeHome, and DomainNet respectively.

Table 11 showcases the zero-shot results for ResNet50x16 model across different prompts, substantiating that CLAP is more effective than Im.Aug in refining CLIP features. Moreover, Tab. 12 illustrates that both Im.Aug and CLAP reduce variations in zero-shot performance across different prompts, with the improvement of CLAP being more significant. The results validate our approach across different model scales, including both ViT-based and CNN-based structures. Domain-level results are detailed in Tab. 13.

D Discussion

D.1 Rationale behind CLAP's Foundation on CLIP

The primary challenge in cross-modal transferability lies in the significant domain gap between text and image data, which typically hinders the direct application of models trained in one modality to another. For a causal explaination, despite the consistency of the content variable that dictates the object label across modalities, the generative processes from latent variables to observations inherent to each modality differ markedly. The CLIP model, trained on a comprehensive dataset of image-text pairs with a symmetric InfoNCE loss, significantly ameliorates this issue. By aligning the features of text and images into similar patterns, it facilitates leveraging a network trained atop the CLIP encoder of

You

I have some class names, such as "dog", "car", "person", "chair". I want you to generate various style of these objects in different art-style pictures. Providing me with a list of different sentences that decribe a scene contaning the object, with different object attributes (such as object size, condition or age(if applicable), colors), different backgrounds the object may apear in; different image types (such as photo, painting, etc).

ChatGPT

Absolutely, I'll provide you with a list of descriptive sentences for each of the class names "dog", "car", "person", "chair," incorporating various object attributes, backgrounds, and image types. Each sentence will describe a unique scene:

Fig. 2: The prompting method we use for generating text prompts with ChatGPT-3.5.

Table 8: Zero-shot performance across four prompts ("C", "PC", "CP") and 1 noised prompts ("NC") with CLIP pre-trained ViT-L/14 model. CLAP demonstrates consistent gains in zero-shot performance across all datasets, validating its effectiveness.

Prompt	Method	Zero-shot performance, avg. top-1 acc. (%) (↑)				
		PACS	VLCS	OfficeHome	DomainNet	Overall
ZS(C)	CLIP	97.6	77.1	85.9	63.2	80.9
	Im.Aug	98.3	78.5	86.0	63.4	81.6
	CLAP	**98.5**	**80.7**	**87.5**	**64.2**	**82.7**
ZS(CP)	CLIP	97.3	80.6	86.0	62.0	81.5
	Im.Aug	98.3	81.1	86.1	62.4	82.0
	CLAP	**98.5**	**81.4**	**87.9**	**63.7**	**82.9**
ZS(PC)	CLIP	98.4	81.7	86.5	63.5	82.5
	Im.Aug	**98.6**	81.9	86.6	63.7	82.7
	CLAP	**98.6**	**82.2**	**88.0**	**64.5**	**83.3**
ZS(NC)	CLIP	91.0	65.5	77.1	55.4	72.3
	Im.Aug	95.6	69.3	77.1	55.7	74.4
	CLAP	**98.5**	**73.1**	**81.3**	**58.3**	**77.8**

one modality as a viable proxy for the other. Consequently, this allows for the direct application of the disentangled network trained in the text modality atop CLIP's image encoder to refine representations.

D.2 Impact of Image and Text Augmentations

Identifying pure content factors poses a significant challenge. This difficulty primarily arises from the need for finding effective augmentations of observational data to alter style factors significantly while preserving content integrity.

Through the cross-modal alignment provided by CLIP, we discovered that disentangling in one modality can seamlessly improve representations in both modalities. The impact of image augmentations has been well-explored and found effective at preserving content, but traditional methods do not impose sufficient changes to remove all style information. Our exploration of text augmentations

Table 9: CLAP reduces the variance in zero-shot performance across different prompts with CLIP pre-trained ViT-L/14 model.

Metric	Method	Zero-shot variance, avg. top-1 acc. (%) (↓)				
		PACS	VLCS	OfficeHome	DomainNet	Overall
R	CLIP	1.0	4.6	0.6	1.5	1.9
	Im.Aug	0.3	3.4	0.6	1.3	1.4
	CLAP	**0.1**	**1.5**	**0.4**	**0.7**	**0.7**
δ	CLIP	0.4	2.0	0.3	0.6	0.8
	Im.Aug	0.1	1.5	0.3	0.5	0.6
	CLAP	**0.0**	**0.6**	**0.2**	**0.3**	**0.3**
$\Delta_{(NC)}$	CLIP	6.6	11.5	8.8	7.8	8.7
	Im.Aug	2.7	9.2	8.9	7.7	7.1
	CLAP	**0.1**	**7.7**	**6.3**	**5.9**	**5.0**

Table 10: Domain-level zero-shot results of the ViT-L/14 model on the test datasets.

Datasets	Domains	Domain-level avg. top-1 acc. (%) of zero-shot performance using ViT-L/14 (↑)											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	97.2	98.0	98.8	96.8	98.0	98.5	98.7	98.8	98.9	85.6	91.6	98.6
	C.	99.5	99.6	99.8	98.3	99.6	99.7	99.5	99.6	99.7	95.9	98.1	99.6
	P.	99.9	100.0	100.0	99.4	99.5	100.0	99.9	100.0	99.9	91.1	97.5	99.9
	S.	93.8	95.7	95.5	94.8	96.0	95.7	95.4	95.9	95.8	91.5	95.2	95.8
VLCS	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	87.5	87.9	94.4
	L.	57.4	60.1	64.3	71.3	71.6	72.6	71.7	72.0	72.6	53.8	59.7	60.7
	S.	71.0	72.4	74.4	66.2	67.4	66.8	69.9	70.4	69.9	55.9	60.5	62.9
	V.	80.0	81.6	84.3	85.2	85.7	86.2	85.1	85.3	86.4	65.0	69.3	74.3
OfficeHome	A.	86.2	86.3	87.7	85.7	86.2	88.1	87.0	87.0	87.8	78.1	77.1	80.7
	C.	73.3	73.4	75.7	73.8	73.4	76.0	73.1	73.5	76.0	65.9	66.3	70.6
	P.	92.0	91.8	93.6	92.3	92.4	94.3	92.9	92.8	94.1	80.7	81.0	86.8
	R.	92.2	92.7	93.0	92.2	92.4	93.4	93.1	93.3	93.9	83.8	84.0	86.9
DomainNet	C.	78.4	78.5	79.1	77.5	77.7	78.8	79.4	79.4	79.7	70.0	70.4	72.8
	I.	52.9	53.0	54.6	50.4	50.7	53.6	51.7	52.0	53.9	45.3	45.2	48.8
	P.	70.4	70.8	72.4	68.9	69.9	72.1	69.9	70.6	72.7	59.9	60.3	64.8
	Q.	21.5	21.6	22.5	20.6	20.9	21.7	22.6	22.8	22.9	17.9	18.4	20.2
	R.	85.8	85.9	85.9	85.3	85.5	85.7	86.3	86.4	86.2	77.5	77.5	78.7
	S.	70.2	70.4	70.7	69.4	69.8	70.6	71.0	71.3	71.5	62.0	62.2	64.6

reveals that the logical structure of text and the relative ease of implementing style changes can have a significant impact on achieving disentanglement. However, more efficient methods are worthy of exploration.

A promising direction for future research is to explore efficient combinations of both modalities to enhance disentangled semantics. As each modality has its unique advantages—Text data recapitulates properties well since it is pre-processed by human intelligence, while image data is more precise in depicting the exact same objects or events due to its more detailed nature—the impact of combining augmentations of both modalities could be substantial.

Table 11: Zero-shot performance with CLIP pre-trained ResNet50x16 model. CLAP demonstrates consistent enhancement across all datasets, validating its effectiveness.

Prompt	Method	Zero-shot performance, avg. top-1 acc. (%) (↑)				
		PACS	VLCS	OfficeHome	DomainNet	Overall
ZS(C)	CLIP	96.1	70.4	80.4	57.1	76.0
	Im.Aug	96.4	74.7	80.4	57.1	77.2
	CLAP	**97.0**	**79.9**	**81.6**	**58.0**	**79.1**
ZS(CP)	CLIP	95.0	73.5	79.0	56.1	75.9
	Im.Aug	95.7	75.8	79.3	56.5	76.8
	CLAP	**96.7**	**80.3**	**79.9**	**57.4**	**78.6**
ZS(PC)	CLIP	96.5	78.4	81.7	57.1	78.4
	Im.Aug	97.0	79.8	81.8	57.4	79.0
	CLAP	**96.8**	**80.1**	**82.5**	**58.2**	**79.4**
ZS(NC)	CLIP	86.4	61.2	69.3	48.2	66.3
	Im.Aug	88.3	71.3	69.5	48.7	69.4
	CLAP	**94.9**	**80.1**	**71.9**	**50.6**	**74.4**

Table 12: CLAP consistently reduces variances in zero-shot performance across different prompts with CLIP pre-trained ResNet50x16 model, validating its effectiveness.

Metric	Method	Zero-shot variance, avg. top-1 acc. (%) (↓)				
		PACS	VLCS	OfficeHome	DomainNet	Overall
R	CLIP	1.5	8.0	2.7	1.1	3.3
	Im.Aug	1.3	5.1	**2.5**	0.9	2.4
	CLAP	**0.3**	**0.4**	2.6	**0.8**	**1.0**
δ	CLIP	0.6	3.3	1.1	0.5	1.4
	Im.Aug	0.5	2.2	**1.0**	0.4	1.0
	CLAP	**0.1**	**0.2**	1.1	**0.3**	**0.4**
$\Delta_{(NC)}$	CLIP	9.7	9.3	11.1	8.9	9.7
	Im.Aug	8.1	3.5	10.9	8.5	7.7
	CLAP	**2.1**	**-0.1**	**9.7**	**7.5**	**4.8**

Table 13: Domain-level zero-shot results using RestNet50x16 on the test datasets.

Datasets	Domains	Domain-level avg. top-1 acc. (%) of zero-shot performance using RN50x16 (↑)											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	95.7	95.8	97.2	93.7	95.5	96.5	95.7	96.7	96.8	81.2	84.0	94.5
	C.	98.3	98.2	99.0	98.1	98.8	99.0	98.6	98.7	98.9	92.3	93.2	98.0
	P.	98.9	98.6	99.9	98.4	97.8	99.8	99.8	99.9	99.9	85.3	87.3	95.2
	S.	91.5	93.1	91.9	89.7	90.8	91.5	91.8	92.9	91.6	86.9	88.6	92.1
VLCS	C.	96.8	97.1	99.3	99.7	99.4	99.3	99.7	99.6	99.4	75.6	89.3	99.4
	L.	53.4	60.8	65.9	51.6	58.9	67.3	59.5	68.1	66.8	54.1	60.6	67.0
	S.	63.2	70.9	69.5	68.0	72.9	69.5	72.9	73.7	69.0	52.0	66.7	69.5
	V.	68.4	70.1	85.2	74.5	72.1	85.3	81.7	78.0	85.2	63.1	68.5	84.5
OfficeHome	A.	82.2	82.5	83.5	79.7	79.9	80.6	82.0	82.4	83.4	67.7	68.9	72.1
	C.	63.0	62.9	64.7	61.7	62.2	62.8	65.4	65.3	66.1	54.6	55.0	56.8
	P.	88.2	87.9	89.0	87.4	87.5	88.5	90.0	89.9	90.6	75.4	75.3	78.2
	R.	88.1	88.2	89.1	87.3	87.5	87.6	89.2	89.5	89.7	79.5	78.9	80.3
DomainNet	C.	69.0	68.9	69.6	68.6	68.6	69.4	69.9	70.0	70.4	59.5	60.1	61.4
	I.	51.0	51.1	52.7	48.2	49.0	50.6	48.2	48.9	50.7	41.2	41.6	44.3
	P.	65.2	65.6	66.5	63.7	64.4	65.6	65.4	65.9	67.0	53.5	54.3	56.8
	Q.	11.8	11.9	12.7	12.3	12.6	13.1	11.8	12.2	12.7	9.3	9.7	11.0
	R.	82.1	82.2	83.1	81.6	81.8	82.6	83.3	83.4	83.8	72.9	73.0	74.7
	S.	63.2	63.1	63.6	62.0	62.4	63.4	63.9	63.8	64.6	53.1	53.4	55.3