

Supplementary Material for Context-Adaptive Multi-Prompt Embedding with Large Language Models for Vision-Language Alignment

Anonymous authors

Paper under double-blind review

1 A Limitations

2 Our experiments were constrained by available compute resources. Most of our training
3 runs used batch sizes of 1K or 4K, and we scaled to 16K only for comparisons with other
4 CLIP methods. This remains smaller than the batch sizes of up to 32K used in large-scale
5 models like OpenAI CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). Additionally,
6 we use ViT-B/16 as our visual encoder throughout all experiments. Exploring stronger
7 backbones such as ViT-L/14 may yield further improvements.

8 In the video domain, recent methods often train on large mixtures of diverse video-text
9 datasets. For instance, VideoCoCa (Yan et al., 2022) leverages both HowTo100M (Miech
10 et al., 2019) and VideoCC3M (Nagrani et al., 2022), while UMT (Li et al., 2023) combines
11 Kinetics (Carreira et al., 2019), CC3M (Sharma et al., 2018), and WebVid (Bain et al., 2021).
12 In contrast, we use only VideoCC3M in our experiments to simplify the setting and better
13 understand the impact of our proposed context-adaptive multi-prompt method. Expanding
14 to multi-source training remains a promising avenue for future work.

15 B Reproducibility: Additional Implementation Details

16 B.1 Visual Encoder.

17 Our image and video encoders use a ViT backbone with an attention pooling layer applied
18 after the final ViT layer. This attention pooling layer is a single multi-head attention module
19 (12 heads) with one learnable query. The ViT patch features serve as both keys and values.
20 The output is a global visual representation used in the contrastive loss.

21 B.2 Text Projection Layer

22 The output of the LLM text encoder is first passed through a linear projection layer to
23 reduce its feature dimension to D/K , where D is the embedding dimension of the ViT visual
24 encoder and K is the number of adaptive prompts.

25 B.3 Image and Video CLIP Training.

26 We summarize the hyperparameters used in our training runs for both image-text and
27 video-text contrastive learning.

28 C Additional Ablations

29 **Negation-augmented contrastive loss.** We ablate the effect of the loss weight β for the
30 negation-aware contrastive loss. As shown in Table 2, incorporating L_{neg} with $\beta = 0.1$
31 consistently improves retrieval performance. However, setting $\beta = 1.0$ does not yield
32 additional gains, and thus we use $\beta = 0.1$ in all main experiments.

33 As mentioned in the main paper, the loss weight α for the prompt diversity regularization
34 loss L_{div} is also set to 0.1.

	Image CLIP pretraining	Video CLIP training
optimizer	AdamW	AdamW
momentum	$\beta=0.9$	$\beta=0.9$
weight decay	0.01	0.01
batch size	1024, 4096, 16384	128
learning rate	5e-4 (for batch 1024, 4096) 1e-3 for (batch 16384)	1e-5
warmup iterations	5k	5k
total iterations	500000	50000
image size	224× 224	224× 224
number of frames	N/A	16 uniformly sampled per video

Table 1: Training hyperparameters for Image and Video CLIP.

method	Flickr R@1		MSCOCO R@1	
	img-to-txt	txt-to-img	img-to-txt	txt-to-img
$\beta = 0$ (w/o L_{neg})	66.0	47.1	41.0	25.2
$\beta = 0.1$	67.2	48.0	41.8	26.0
$\beta = 1.0$	66.8	47.5	41.6	25.8

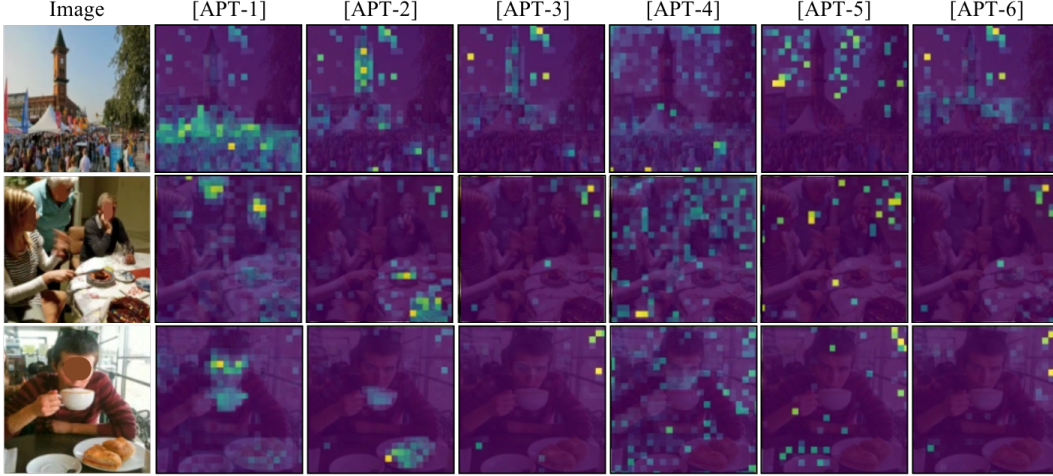
Table 2: Negation-aware contrastive loss (L_{neg}).

Figure 1: Additional attention visualization.

35 D Additional Attention Visualization

36 We provide additional visualizations of the attention maps extracted from the attention
 37 pooling layer of the ViT encoder. For each image (row), we visualize how each of the
 38 six adaptive prompt tokens [APT-i] (columns) attends to different regions of the image.
 39 These maps highlight the specialization of each prompt token: some focus on subjects or
 40 foreground objects, while others attend to broader background elements. Attention maps
 41 are averaged over heads within the respective channel segment of the visual embedding.

42 References

43 Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video
 44 and image encoder for end-to-end retrieval. In *ICCV*, pp. 1728–1738, 2021.

- 45 Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the
46 kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.
- 47 Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le,
48 Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language
49 representation learning with noisy text supervision. In *ICML*, 2021.
- 50 Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yanan He, Limin Wang, and Yu Qiao. Un-
51 masked teacher: Towards training-efficient video foundation models. In *ICCV*, 2023.
- 52 Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev,
53 and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred
54 million narrated video clips. In *ICCV*, 2019.
- 55 Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen
56 Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In
57 *ECCV*, pp. 407–426. Springer, 2022.
- 58 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agar-
59 wal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and
60 Ilya Sutskever. Learning transferable visual models from natural language supervision.
61 In *ICML*, 2021.
- 62 Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions:
63 A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*,
64 2018.
- 65 Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and
66 Jiahui Yu. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive
67 captioners. *arXiv preprint arXiv:2212.04979*, 2022.