Supplementary Materials
Beyond Coarse-Grained Matching in Video-Text Retrieval

A Further Analysis of the Fine-Grained Ability of Current Models

We show the *PoSRank* performance for individual parts-of-speech for all methods on VATEX and VLN-OOPS in Fig. A. We see similar trends as in Figure 6 of the main paper. Particularly, current models are better at distinguishing differences in adjectives and nouns and worse at distinguishing differences in adverbs and prepositions.

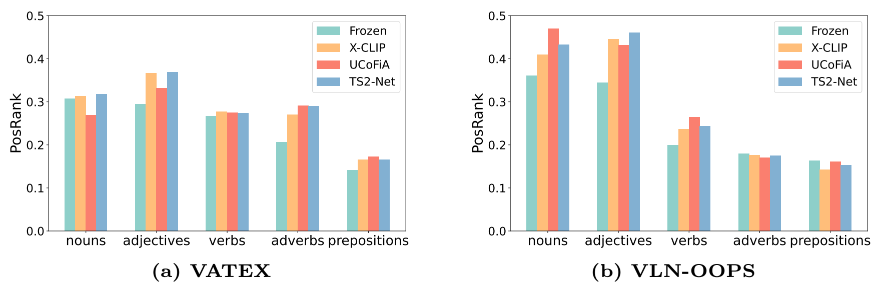

Fig. A: Fine-Grained Evaluation Per Part of Speech. Models find fine-grained differences in adverbs and prepositions the most difficult to distinguish.

B Implementation Details of our Fine-grained Training

X-CLIP, TS2Net, and UCoFiA use the visual and text encoder from CLIP (ViT-B/32) [2], while Frozen uses a variant of ViT and DistilBERT [3] pretrained on WebVid-2M [1]. The training configurations are consistent with the original publications, All models are optimized with Adam. For X-CLIP, TS2Net and UCoFiA are trained for 5 epochs with a batch size of 64, the initial learning rate for visual and text encoder is $1e{-}7$, and the initial learning rate for other modules is $1e{-}4$. For Frozen, we set a learning rate of $1e{-}5$ with a maximum of 100 epochs, with early stopping if validation performance did not improve for 10 consecutive epochs, using a batch size of 32.

C Effect of the Number of Visual Prompts $\mathbb{T}$

We explore the impact of the number of visual prompts in Tab. A. With more than one visual prompt we mean-pool the embedding of all prompts for the fine-grained representation. Using any number of visual prompts $\mathbb{T}>0$ to better separate coarse and fine-grained objectives provides an increase in results. There is a small increase for coarse-grained evaluation with $\mathbb{T}>1$ however, $\mathbb{T}=1$ provides the best balance between coarse and fine-grained performance and training efficiency.

Table A: Number of visual prompts $\mathbb{T}$. $\mathbb{T}=1$ provides a reasonable balance between performance and training efficiency.

$\mathbb{T}$	Coarse-Grained (↑)		Fine-Grained (↑)				
	V2T	T2V	noun	adj	verb	adv	prep
0	47.5	40.0	0.894	0.864	0.969	0.468	0.701
1	53.1	39.3	0.858	0.832	0.960	0.501	0.668
2	54.0	39.8	0.859	0.796	0.953	0.472	0.519
4	54.0	40.2	0.854	0.798	0.950	0.424	0.477
8	52.5	39.7	0.841	0.783	0.954	0.419	0.512

D Variance of Results

The table below shows the variance from 3 runs of our full model on VLU-UVO, verifying that the variance of our approach is low.

Coarse-Grained (↑)		Fine-Grained (↑)				
V2T	T2V	noun	adj	verb	adv	prep
0.03	0.2	0.000007	0.000009	0.000001	0.0002	0.0009

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV (2021) 1
2. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1
3. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: EMC2@NeurIPS (2019) 1