Hybrid-Tower: Fine-grained Pseudo-query Interaction and Generation for Text-to-Video Retrieval

Supplementary Material

In this supplementary material, we report additional experimental results which are not included in the main paper due to space limits.

S1. Evaluation on Million-level Video Datasets

We apply the TRECVID evaluation [1] to our method, which consists of million-scale video datasets. s shown in Tab. S1, our method generalizes well.

Model	V3C1			V3C2		Mean
	TV19	TV20	TV21	TV22	TV23	
CLIP4Clip	0.142	0.161	0.183	0.127	0.139	0.150
CLIP-ViP	0.143	0.148	0.175	0.109	0.088	0.133
F-Pig(Ours)	**0.159**	**0.186**	**0.216**	**0.158**	**0.146**	**0.173**

Table S1. **Performance comparison.** Metric: inf P (higher is better). Training Data: MSRVTT-9k. The results for CLIP4Clip and CLIP-ViP are sourced from LPD[2].

S2. Sensitivity nalysis

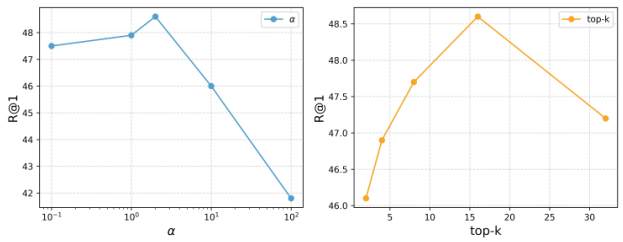

Figure S1. Sensitivity analysis on MSRVTT-1k. : generation loss weight; top-k: number of selected patch tokens in ITS module.

S3. Video-to-Text Retrieval Results

We report video-to-text retrieval results on MSRVTT-1k and MSRVTT-3k, where our method also achieves improved performance, see Tab. S2.

Model	MSRVTT-1k			MSRVTT-3k		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP4Clip	41.4	70.6	80.5	50.8	78.5	87.0
TS2-Net*	41.4	68.0	77.8	53.0	82.5	89.8
X-CLIP*	43.4	72.2	82.6	56.1	84.3	92.3
CLIP-ViP	44.5	73.1	81.7	54.6	84.1	**92.7**
TeachCLIP	44.3	**73.6**	**83.7**	55.2	81.9	90.3
PIG	**47.5**	72.9	82.2	**56.6**	**84.5**	92.1

Table S2. **Video-to-text retrieval results.** Models marked by belong to Single-Tower methods. Backbone: CLIP-ViT-B/32.

Item	Setting
GPUs	8 NVIDI 3090
Backbone	CLIP (ViT-B/32, ViT-B/16)
Initialization	Open I-released CLIP
Learning rate (stage 1)	9 1 (generator pretraining)
Learning rate (stage 2)	1 1 ^{6} (full fine-tuning)
Weight decay	0.2
Optimizer	damW [3]
Learning rate schedule	Cosine annealing [4] + warmup (0.01)
Epochs	100
Input frame size	224 224
Max. frame/word tokens	12 / 50
Batch size	128
DiDeMo tokens	32/64 [5]
	2
Top-k	16

Table S3. Implementation Details.

S4. Implementation Details

We list more implementation details, see Tab. S3. Our models are initialized from: https://huggingface.co/openai/clip-vit-base-patch32, https://huggingface.co/openai/clip-vit-base-patch16.

References

[1] George wad, Keith Curtis, sad Butt, Jonathan Fiscus, fzal Godil, Yooyoung Lee, ndrew Delgado, Eliot Godard, Lukas Diduch, Yvette Graham, et al. Trecvid 2023-a series of evaluation tracks in video understanding. In *Proceedings of TRECVID*, 2023. 1

[2] Fan Hu, Zijie Xin, and Xirong Li. Learning partially-decorrelated common spaces for ad-hoc video search. In *Pro-*

ceedings of the 33rd CM International Conference on Multimedia, 2025. 1

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[4] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2022. 1

[5] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: dapting pre-trained image-text model to video-language alignment. In *ICLR*, 2023. 1