SUPPLEMENTARY MATERIALS OF *HOVER*: HYPERBOLIC VIDEO-TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

1 HIERARCHICAL DATASETS

We build three hierarchical datasets to validate the benefits of learning hierarchical semantic structures. ActivityNet (Hierarchy) and ActivityNet (Few-shot) are derived from ActivityNet Caba Heilbron et al. (2015), and Charades (Hierarchy) is built from Charades Gao et al. (2017).

ActivityNet (Hierarchy) contains binary video trees with a maximum depth of 6. The leaf nodes are individual video clips, and adjacent video clips are concatenated to form their parent node. Train, validation, test splits correspondingly have 8846, 4269, 4512 video trees, and 50982, 23214, 21711 nodes. Details are shown in Tab. 1.

ActivityNet (Few-shot) contains 10% of all the video trees from ActivityNet (Hierarchy). Train split consists of 885 video trees, and 3461 nodes. Details are shown in Tab. 2.

Charades (Hierarchy) contains binary video trees with a maximum depth of 3, constructed in the same manner as ActivityNet (Hierarchy). Train, test splits consists of 1984, 572 video trees, and 5909, 1828 nodes, respectively. Details are shown in Tab. 3.

Table 1: Numbers of different structures in ActivityNet (Hierarchy).

Split	Single-leaf		Multi-leaf		Number of leaf nodes				
	Trees	Nodes	Trees	Nodes	1	2	3	4	$\geq$5
train	1282	2564	7564	48418	1282	3094	2320	1142	1008
validation	689	1378	3580	21836	689	1512	1176	515	377
test	1201	2402	3341	19309	1201	1466	1258	365	252

Table 2: Numbers of different structures in ActivityNet (Few-shot).

Split	Single-leaf		Multi-leaf		Number of leaf nodes				
	Trees	Nodes	Trees	Nodes	1	2	3	4	$\geq$5
train	128	256	743	4665	128	325	222	95	101
validation	82	164	327	1919	82	146	104	45	32
test	111	222	353	2105	111	140	150	33	30

Table 3: Numbers of different structures in Charades (Hierarchy).

Split	Single-leaf		Multi-leaf		Number of leaf nodes			
	Trees	Nodes	Trees	Nodes	1	2	3	4
train	779	1558	1205	4351	779	1021	167	17
test	188	376	384	1452	188	309	65	10

2 CROSS-ORDER ANALYSIS

We provide detailed comparisons between the proposed HOVER and Euclidean baseline CLIP4Clip Luo et al. (2022) in regard to action-composition orders. Tab. 4 and Tab. 5 report the performance on ActivityNet (Few-shot) and Charades (Hierarchy), respectively. The second column "leaf" represents the number of leaf nodes in a single video tree. On ActivityNet (Few-shot), the best performance is achieved in high-order trees containing 3 leaf nodes, with $+22.9\%$ in MRR for text-to-video retrieval over the Euclidean baseline. On Charades (Hierarchy), the best performance is achieved in high-order trees containing 3 leaf nodes, with $+39.0\%$ in MRR for text-to-video retrieval over the Euclidean baseline.

Table 4: Comparison in terms of different composition orders of actions on ActivityNet (Few-shot).

Leaf	Method	t2v R@1	R@5	R@10	v2t R@1	R@5	R@10	MRR(t2v)	ΔMRR
1	CLIP4Clip	13.8	40.4	54.7	14.0	40.3	54.4	0.266	+15.04%
	HOVER	**16.3**	**45.9**	**61.1**	**17.6**	**45.9**	**60.1**	**0.306**	
2	CLIP4Clip	9.1	30.3	42.2	8.7	30.4	43.3	0.200	+17.50%
	HOVER	**11.0**	**36.1**	**48.4**	**12.1**	**35.8**	**49.9**	**0.235**	
3	CLIP4Clip	4.8	20.3	31.7	5.5	22.0	33.5	0.131	**+22.90%**
	HOVER	**6.2**	**25.9**	**38.7**	**7.9**	**29.4**	**41.7**	**0.161**	
4	CLIP4Clip	10.7	36.8	52.3	11.0	37.3	53.6	0.233	+15.45%
	HOVER	**12.7**	**44.4**	**59.1**	**15.6**	**46.9**	**62.1**	**0.269**	
≥5	CLIP4Clip	10.7	34.9	51.2	9.8	34.1	50.4	0.230	+18.26%
	HOVER	**13.4**	**42.1**	**60.9**	**14.0**	**43.1**	**60.5**	**0.272**	

Table 5: Comparison in terms of different composition orders of actions on Charades (Hierarchy).

Leaf	Method	t2v R@1	R@5	R@10	v2t R@1	R@5	R@10	MRR(t2v)	ΔMRR
1	CLIP4Clip	2.7	11.2	18.4	4.0	11.1	17.7	0.081	+29.63%
	HOVER	**4.5**	**14.9**	**21.5**	**5.8**	**16.5**	**22.1**	**0.105**	
2	CLIP4Clip	1.5	5.7	9.0	1.5	5.7	8.2	0.043	+23.26%
	HOVER	**1.7**	**6.9**	**12.1**	**1.5**	**6.3**	**11.8**	**0.053**	
3	CLIP4Clip	1.8	8.0	12.7	1.8	7.5	11.5	0.059	**+38.98%**
	HOVER	**2.3**	**13.3**	**21.1**	**2.4**	**9.4**	**14.9**	**0.082**	
4	CLIP4Clip	3.7	21.6	32.8	1.5	17.4	29.5	0.137	+16.79%
	HOVER	**4.5**	**22.4**	**44.0**	**3.8**	**22.0**	**33.3**	**0.160**	

3 EMBEDDING VISUALIZATION

We provide visualization of the learned video-text embedding vectors in Fig. 1.

REFERENCES

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pp. 961–970, 2015.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pp. 5267–5275, 2017.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

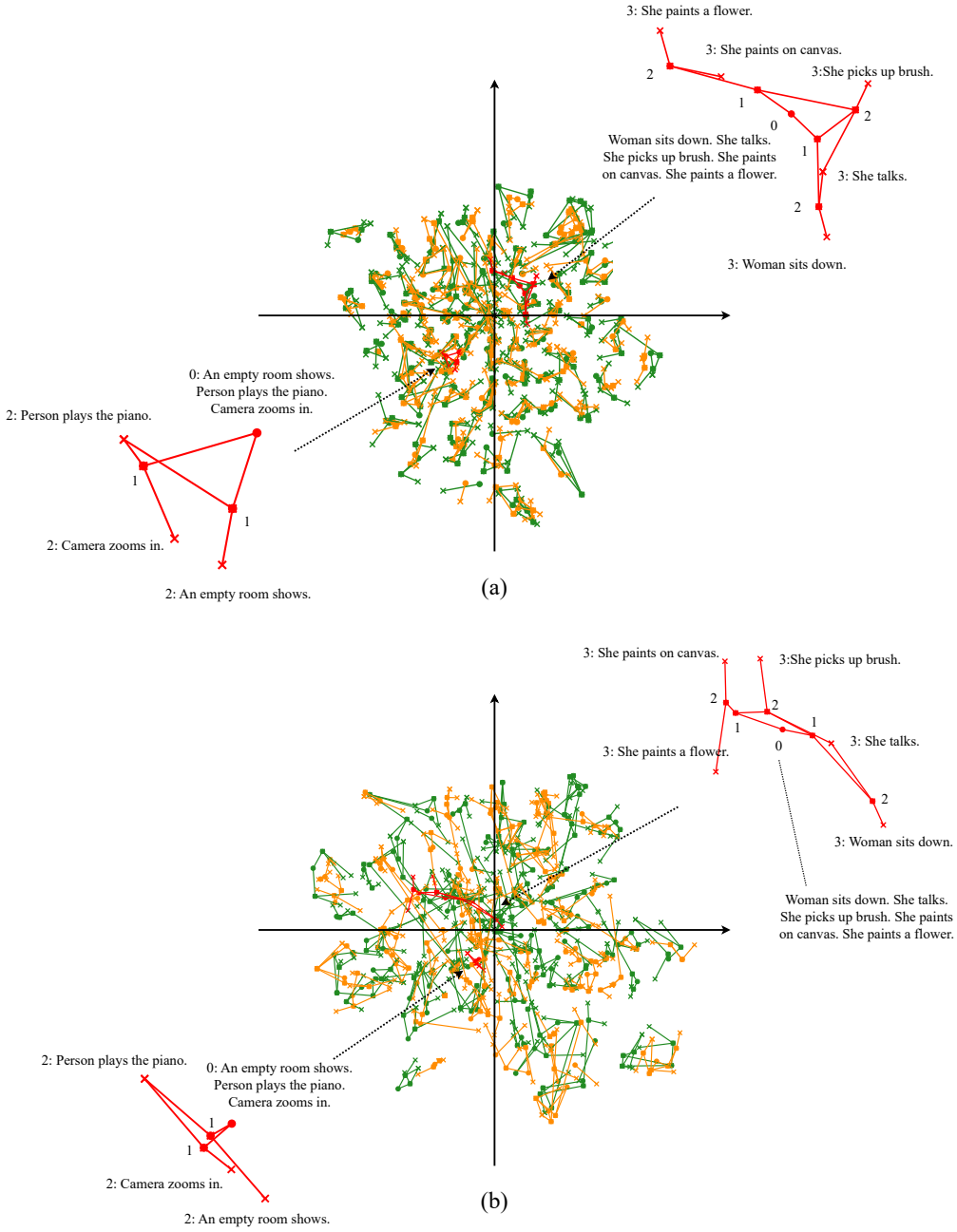

Figure 1: Hyperbolic embeddings of **videos** and **texts** when trained with (a) L_{joint}, (b) L_{align}. Leaf nodes are marked as $\times$, and other nodes are marked as $\bullet$. Nodes of **specific video examples** are annotated with their depths in the semantic tree and the corresponding text descriptions.