Fine-tuned Hyperbolic CLIP Models are Good Video Learners

Supplementary Material

A. Full Configuration Results

Tab. 11 presents the complete zero-shot results across all $2 \times 4 \times 3 = 24$ configurations. In the main paper, we report condensed tables isolating the effects of freezing strategy (Tab. 2) and temporal aggregation (Tab. 3).

B. Embedding Norm Analysis

Following MERU [1], we analyze embedding norms (distance from the natural origin) to understand how features are distributed in each geometry. Fig. 5 shows the norm distributions for video and text embeddings across all evaluation datasets.

C. Detailed Hierarchical Analysis

Tab. 12 provides the full per-parent accuracy breakdown across the K400 action hierarchy from Long *et al*. [11]. Tab. 13 reports per-grandparent accuracy when predictions are recomputed within each cluster, along with within-parent misclassification rates.

D. Temporal and Static Class Lists

K400 temporal classes (32). bouncing on trampoline, breakdancing, busking, cartwheeling, cleaning shoes, country line dancing, drop kicking, gymnastics tumbling, hammer throw, high kick, jumpstyle dancing, kitesurfing, parasailing, playing cards, playing cymbals, playing drums, playing ice hockey, robot dancing, shining shoes, shuffling cards, side kick, ski jumping, skiing (not slalom or crosscountry), skiing crosscountry, skiing slalom, snowboarding, somersaulting, tap dancing, throwing ball, throwing discus, vault, wrestling.

K400 static classes (32). belly dancing, bending back, blasting sand, blowing nose, changing wheel, clapping, curling hair, deadlifting, dining, doing aerobics, dribbling basketball, eating doughnuts, filling eyebrows, getting a tattoo, laying bricks, long jump, lunge, making bed, moving furniture, mowing lawn, peeling apples, playing badminton, playing controller, playing cricket, pull ups, riding camel, shot put, testifying, trimming trees, waxing eyebrows, yawning, yoga.

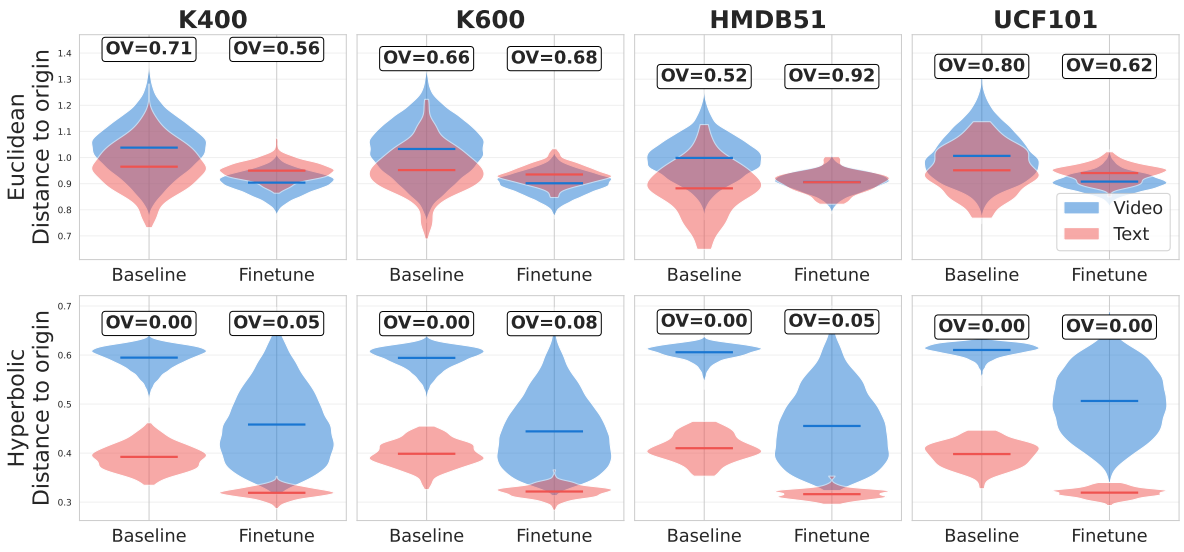

Figure 5. **Embedding norm distributions.** Distribution of video and text norms (distance from origin) for Euclidean and hyperbolic features across all evaluation datasets.

Table 11. **Full ablation: zero-shot accuracy across freeze strategies and temporal aggregations.** All 24 configurations trained on K400 with ViT-B/16. Deltas show hyperbolic − Euclidean difference.

Geom.	Freeze	Temporal	K400	K600	HMDB-51	UCF-101
Acc@1						
Euc.	None	Mean	63.1	47.0	37.3	60.7
		SeqLSTM	62.4	46.3	38.1	61.3
		SeqTransf	62.1	34.4	36.6	55.5
	Backbone	Mean	27.2	35.2	19.9	41.9
		SeqLSTM	36.1	43.3	30.2	48.4
		SeqTransf	45.1	31.9	28.9	47.1
	Text	Mean	55.4	42.9	35.8	59.6
		SeqLSTM	55.6	43.9	36.7	59.6
		SeqTransf	57.0	34.4	33.3	55.7
	Image	Mean	52.8	42.8	32.9	53.4
		SeqLSTM	52.7	42.3	33.2	53.6
		SeqTransf	52.2	28.9	28.7	47.6
Hyp.	None	Mean	66.9 (+3.8)	51.3 (+4.3)	39.6 (+2.3)	65.7 (+4.9)
		SeqLSTM	67.0 (+4.6)	51.2 (+4.9)	40.1 (+2.0)	65.7 (+4.4)
		SeqTransf	65.7 (+3.6)	42.5 (+8.1)	37.6 (+1.0)	60.5 (+5.0)
	Backbone	Mean	28.7 (+1.5)	38.7 (+3.5)	23.5 (+3.6)	44.5 (+2.6)
		SeqLSTM	38.9 (+2.8)	46.3 (+3.0)	28.5 (-1.7)	53.2 (+4.8)
		SeqTransf	47.9 (+2.8)	40.9 (+9.0)	31.3 (+2.5)	52.9 (+5.8)
	Text	Mean	60.4 (+5.0)	42.8 (-0.1)	37.2 (+1.4)	59.8 (+0.2)
		SeqLSTM	60.5 (+4.9)	44.7 (+0.8)	37.0 (+0.3)	60.7 (+1.1)
		SeqTransf	61.4 (+4.4)	42.1 (+7.6)	36.6 (+3.3)	60.9 (+5.1)
	Image	Mean	55.5 (+2.7)	46.5 (+3.7)	34.5 (+1.5)	57.0 (+3.6)
		SeqLSTM	55.3 (+2.7)	46.7 (+4.4)	35.0 (+1.8)	57.0 (+3.3)
		SeqTransf	55.2 (+3.0)	38.4 (+9.5)	30.5 (+1.8)	53.1 (+5.5)
Acc@5						
Euc.	None	Mean	86.3	73.1	64.1	85.7
		SeqLSTM	85.7	72.6	64.8	85.8
		SeqTransf	86.2	63.4	64.2	80.5
	Backbone	Mean	51.7	61.0	46.9	71.7
		SeqLSTM	63.1	69.5	55.9	76.0
		SeqTransf	73.2	59.3	54.1	76.5
	Text	Mean	81.0	69.0	62.3	82.7
		SeqLSTM	80.8	70.4	62.5	83.6
		SeqTransf	82.4	63.3	61.0	81.9
	Image	Mean	78.8	68.8	56.9	79.6
		SeqLSTM	78.6	68.7	58.0	79.9
		SeqTransf	79.3	56.3	54.3	73.5
Hyp.	None	Mean	88.5 (+2.2)	78.0 (+4.9)	63.8 (-0.3)	89.3 (+3.7)
		SeqLSTM	88.6 (+2.9)	77.8 (+5.2)	63.4 (-1.4)	89.2 (+3.4)
		SeqTransf	88.4 (+2.2)	71.1 (+7.6)	62.0 (-2.2)	86.0 (+5.5)
	Backbone	Mean	53.6 (+1.8)	63.9 (+2.9)	46.4 (-0.4)	70.8 (-1.0)
		SeqLSTM	66.8 (+3.7)	72.1 (+2.6)	54.5 (-1.4)	79.6 (+3.7)
		SeqTransf	75.6 (+2.4)	67.6 (+8.3)	56.6 (+2.6)	81.0 (+4.5)
	Text	Mean	83.9 (+2.9)	69.8 (+0.8)	60.8 (-1.4)	86.6 (+3.9)
		SeqLSTM	84.0 (+3.2)	71.5 (+1.1)	62.9 (+0.4)	86.5 (+2.9)
		SeqTransf	85.2 (+2.8)	70.4 (+7.1)	61.8 (+0.8)	85.8 (+3.9)
	Image	Mean	80.8 (+2.0)	73.0 (+4.2)	57.1 (+0.2)	83.4 (+3.9)
		SeqLSTM	80.8 (+2.1)	73.2 (+4.5)	57.6 (-0.4)	83.8 (+3.9)
		SeqTransf	81.3 (+2.0)	66.1 (+9.8)	54.7 (+0.4)	82.1 (+8.5)

Table 12. **Per-parent accuracy** on K400 (200 classes from k400_depth [11]). Full breakdown by parent category.

GP	Parent	n	Baseline		SeqLSTM	
			Euc.	Hyp.	Euc.	Hyp.
Arts & Ent.	arts and crafts	149	38.9	51.0 (+12.1)	86.6	89.9 (+3.4)
	body motions	149	20.1	20.8 (+0.7)	39.6	49.0 (+9.4)
	dancing	642	74.8	78.7 (+3.9)	86.6	88.2 (+1.6)
	martial arts	299	27.1	51.5 (+24.4)	70.9	78.6 (+7.7)
	music	949	51.9	55.2 (+3.3)	91.0	93.0 (+2.0)
Household	cleaning	150	28.0	41.3 (+13.3)	68.0	78.0 (+10.0)
	cooking	349	76.5	71.9 (-4.6)	94.0	96.3 (+2.3)
	garden + plants	99	30.3	17.2 (-13.1)	86.9	84.8 (-2.0)
	paper	349	48.1	56.7 (+8.6)	88.5	88.8 (+0.3)
	using tools	149	48.3	53.0 (+4.7)	82.6	86.6 (+4.0)
Sports/Rec.	athletics jumping	199	52.3	59.3 (+7.0)	86.9	91.5 (+4.5)
	athletics throwing	350	45.4	23.1 (-22.3)	75.4	80.3 (+4.9)
	ball sports	397	36.5	41.3 (+4.8)	86.1	87.7 (+1.5)
	golf	99	47.5	72.7 (+25.3)	93.9	96.0 (+2.0)
	gym	498	82.3	74.1 (-8.2)	90.2	90.6 (+0.4)
	gymnastics	99	33.3	36.4 (+3.0)	61.6	65.7 (+4.0)
	head + mouth	397	29.5	47.9 (+18.4)	61.7	66.0 (+4.3)
	heights	398	59.0	63.6 (+4.5)	89.7	94.0 (+4.3)
	juggling	197	46.2	14.2 (-32.0)	82.7	87.3 (+4.6)
	racquet + bat sports	400	39.2	42.0 (+2.8)	81.2	87.2 (+6.0)
	snow + ice	596	88.8	95.3 (+6.5)	96.8	97.7 (+0.8)
	swimming	100	51.0	59.0 (+8.0)	96.0	97.0 (+1.0)
	water sports	396	82.8	87.4 (+4.5)	94.9	96.0 (+1.0)
Personal Care	hair	250	44.0	46.4 (+2.4)	78.8	82.4 (+3.6)
	hands	99	2.0	2.0 (+0.0)	38.4	42.4 (+4.0)
	makeup	245	40.8	42.9 (+2.0)	90.2	90.2 (+0.0)
	personal hygiene	150	10.0	12.7 (+2.7)	83.3	88.0 (+4.7)
Relaxing/Leisure	animals	446	67.3	66.1 (-1.1)	93.7	96.0 (+2.2)
	mobility land	497	56.7	62.2 (+5.4)	92.4	93.0 (+0.6)
	mobility water	200	85.0	89.0 (+4.0)	92.5	95.5 (+3.0)
	playing games	249	44.2	47.0 (+2.8)	89.2	94.0 (+4.8)
Social	communication	249	55.8	49.8 (-6.0)	73.5	77.1 (+3.6)
	eating + drinking	149	43.6	45.6 (+2.0)	80.5	84.6 (+4.0)

Table 13. **Per-grandparent accuracy and within-parent misclassification rates** (K400, predictions recomputed within each cluster).

Config	Grandparent	Accuracy		W-Parent (%)	
		Euc.	Hyp.	Euc.	Hyp.
Baseline	Arts & Ent.	38.3	39.4	53.2	60.0
	Household	63.2	72.9	56.8	56.6
	Sports/Rec.	35.8	37.3	48.3	52.9
	Personal Care	43.1	41.8	36.6	37.4
	Relaxing/Leisure	70.6	73.1	66.0	63.6
	Social	67.3	69.8	65.4	57.5
SeqLSTM	Arts & Ent.	73.7	77.3	67.4	64.4
	Household	86.7	88.5	58.2	61.1
	Sports/Rec.	74.6	78.3	66.6	70.0
	Personal Care	81.6	84.3	48.9	49.6
	Relaxing/Leisure	89.4	91.7	77.0	79.1
	Social	86.2	91.0	70.9	69.4