HyperVLP: Enhancing Hierarchical Surgical Video-Language Pre-training in Hyperbolic Space

Yaojun Hu[1], Kun Yuan[2,3,4†], Nassir Navab[4], Haochao Ying[1†], Jian Wu[1], and Nicolas Padoy[2,3]

[1] Zhejiang University, Hangzhou, China
[2] University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France
[3] IHU Strasbourg, Strasbourg, France
[4] Technical University of Munich, Munich, Germany
yaojunhu@zju.edu.cn, kyuan@unistra.fr, haochaoying@zju.edu.cn

Abstract. Surgical vision-language foundation models typically adopt educational materials, such as surgical lecture videos, to transfer surgical knowledge encoded in language into visual representations. These knowledge are multi-dimensional and hierarchical: fine-grained action cues appear in narration, mid-level key steps are summarized in subsection headings, and global procedural context, such as patient history and surgical strategy, is described in abstract texts. Prior work largely collapses these heterogeneous signals into a single flat embedding space, implicitly assuming independence across hierarchy levels. However, this is suboptimal because it ignores cross-level semantic containment, e.g., actions belong to steps, steps compose phases, weakens long-range dependency modeling. To this end, we propose a hyperbolic surgical video-language pre-training framework that explicitly preserves the hierarchical structure by mitigating structural false negatives induced by procedural context and enforcing semantic consistency between parent phases and their constituent child steps. Extensive experiments on multiple surgical benchmarks show consistent gains in zero- and few-shot phase recognition across procedures and institutions.

Keywords: Surgical Video · Hyperbolic Geometry · Phase Recognition.

1 Introduction

Surgical workflow analysis aims to automatically interpret intraoperative activities from video, with surgical phase recognition as a core task to model procedural progression and provide support for clinical decision-making and objective surgical assessment [14, 13]. Although supervised methods achieve strong performance [19, 26], they depend on dense annotations and generalize poorly across institutions and unseen procedures. Vision-language pretraining (VLP) [20, 11, 24, 12] has emerged as a powerful paradigm for zero-/few-shot

[1] Corresponding authors: Kun Yuan and Haochao Ying.

generalization, which allows models to transfer knowledge from the educational corpora to new tasks by aligning visual and textual representations.

Several important contributions in vision-language pre-training have been made in surgical workflow analysis [21, 23, 6]. SurgVLP [29] pioneers CLIP-style pretraining for surgical workflows with large-scale clip-caption datasets spanning multiple surgical types, while HecVL [28] and PeskaVLP [27] further introduce hierarchical supervision in euclidean embedding space and LLM-based language refinement to enrich structured supervision. Related efforts such as SurgLaVi [18] further constructs a hierarchical large-scale surgical video-language dataset to support domain-specific pretraining.

Surgical workflows exhibit a hierarchical tree structure [18, 27, 28], where phases decompose into ordered steps and fine-grained actions. Existing methods embed these relations in Euclidean space, which lacks the geometric inductive bias to preserve cross-level entailment, weakening parent-child dependencies and limiting structured reasoning and cross-procedure generalization. In contrast, hyperbolic space [1, 2, 17], with negative curvature and exponential capacity, naturally supports low-distortion embedding of hierarchical data and semantic containment. Moreover, current contrastive objectives [29, 18, 20] treat all non-matching video-text pairs as negatives, even when segments are semantically adjacent or temporally correlated. Such design ignores the inherent tree structure and pushes clinically related nodes apart, introducing structural false negatives and distorting workflow hierarchical tree structure.

To resolve the mismatch between tree-structured surgical knowledge and flat Euclidean embeddings, we propose HyperVLP, a hyperbolic video-language pre-training framework. We project video and text representations into a shared Lorentz hyperbolic manifold that naturally models hierarchical containment. To mitigate structural false negatives introduced by conventional contrastive learning, we design a geometry-aware hyperbolic contrastive loss that down-weights semantically related samples within the same procedural context. To further enforce clinically meaningful parent-child dependencies, we introduce a cone-based hyperbolic entailment objective that imposes geometric containment between parent and child embeddings across both inter-modal and intra-modal relations. Since hierarchical constraints in hyperbolic space can destabilize early alignment [3, 10], we adopt a progressive two-stage optimization strategy that first learns cross-modal alignment via geometry-aware contrastive learning and then enforces hierarchical containment through entailment-based refinement. We evaluate HyperVLP on Cholec80 [22], AutoLaparo [25], and MultiBypass140 [9], demonstrating consistent gains in phase recognition under both zero- and few-shot settings. Overall, our contributions are summarized as follows:

- We propose HyperVLP, a hierarchical surgical video-language pretraining framework in hyperbolic space that explicitly encodes the intrinsic tree structure of surgical workflows.
- We identify structural false negatives in surgical video-language contrastive learning and develop a geometry-aware hyperbolic contrastive strategy for hierarchy-aware alignment.

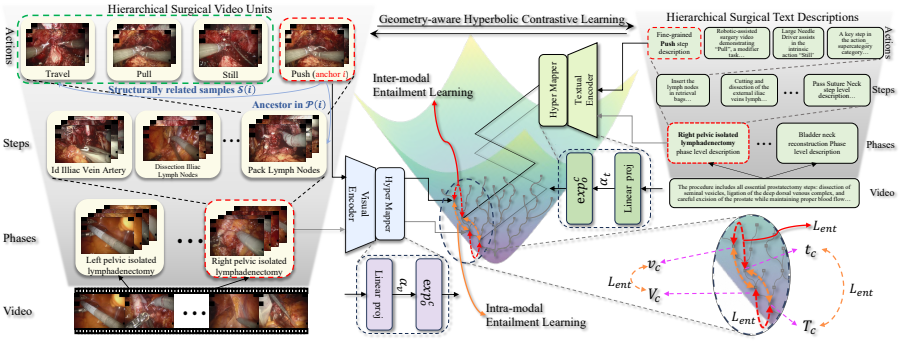

Fig. 1: Overview of HyperVLP. Red dashed boxes denote hierarchical units (V, T, v, t), which are mapped to hyperbolic embeddings (V_c, T_c, v_c, t_c) shown on the right. Green boxes indicate structurally related samples $\mathcal{S}(i)$ used in geometry-aware contrastive learning.

– We introduce a cone-based hyperbolic entailment objective that enforces geometric containment between parent and child embeddings across both inter- and intra-modal relations.

2 Method

2.1 Preliminaries: Hyperbolic Geometry

Hyperbolic space naturally models hierarchical data due to its constant negative curvature and exponential volume growth. This geometry enables low-distortion embedding of tree-structured surgical workflows, where phases decompose into steps and fine-grained actions. We adopt the Lorentz model $\mathbb{L}^n$, defined as the upper sheet of a two-sheeted hyperboloid $\mathbb{L}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{L}} = -\frac{1}{c}, x_0 = \sqrt{\frac{1}{c} + \|\mathbf{x}_n\|^2}, c > 0\}$, where $c \in \mathbb{R}$ represents the curvature and $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}} = \langle \mathbf{x}_n, \mathbf{y}_n \rangle_{\mathbb{E}} - x_0 y_0$ denotes the Lorentzian inner product with $\langle \cdot, \cdot \rangle_{\mathbb{E}}$ indicating the Euclidean inner product. $\mathbf{x}_n$ is the n-dimensional spatial component of $\mathbf{x}$. The geodesic distance is $d_{\mathbb{L}}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{c}} \cosh^{-1}\left(-c \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}}\right)$. Euclidean features are projected to $\mathbb{L}^n$ by mapping tangent vectors at the origin $\mathbf{o} = [\sqrt{1/c}, \mathbf{0}]$ through the exponential map $\exp_{\mathbf{o}}^c(\mathbf{x}) = \cosh\left(\sqrt{c}\,\|\mathbf{x}\|_{\mathbb{L}}\right) \mathbf{o} + \frac{\sinh\left(\sqrt{c}\,\|\mathbf{x}\|_{\mathbb{L}}\right)}{\sqrt{c}\,\|\mathbf{x}\|_{\mathbb{L}}} \mathbf{x}$.

2.2 Proposed Methodology: HyperVLP

Hierarchical Embedding in Hyperbolic Space. Our data contains three hierarchical levels (e.g., phase, step, action). During training, we sample any two adjacent levels to form a parent-child pair for hierarchical modeling. For each unit

(V, T, v, t), visual and textual features are extracted by encoders $\mathcal{F}_v$ and $\mathcal{F}_t$, linearly projected to the tangent space at the origin $\mathbf{o} = [\sqrt{1/c}, \mathbf{0}]$, and mapped to the Lorentz manifold $\mathbb{L}^n$ via the exponential map: $V_c = \exp_{\mathbf{o}}^c(W_v\mathcal{F}_v(V)), \quad T_c = \exp_{\mathbf{o}}^c(W_t\mathcal{F}_t(T))$. The same transformation is applied to the step-level features v and t to obtain v_c and t_c. Learnable scaling in the tangent space stabilizes optimization under negative curvature.

Geometry-Aware Hyperbolic Contrastive learning. Standard contrastive objectives [16, 20, 24] always treat all non-matching samples as true negatives. This assumption breaks in hierarchical surgical workflows. We identify two sources of *structural false negatives*: (i) **contextual proximity**, where clips correspond to adjacent steps within the same phase; (ii) **hierarchical entailment**, where fine-grained clips are semantically subsumed by higher-level segments. Treating such samples as independent negatives enforces artificial separation and distorts the workflow topology.

For each anchor i, we partition candidates into three sets (illustrated on the left side of Fig. 1): (i) $\mathcal{S}(i)$: structurally related samples from the same procedure grouped by temporal containment (including contextual neighbors); (ii)$\mathcal{P}(i)$: matched positive pairs (including hierarchical ancestors and descendants); (iii) $\mathcal{N}(i)$: remaining cross-procedure samples. We define the geometry-aware hyperbolic contrastive loss between parent-level visual and textual embeddings as

$$\mathcal{L}_{geohc}(\mathbf{V_c}, \mathbf{T_c}) = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\sum_{k\in\mathcal{P}(i)} d_{ik}}{\sum_{k\in\mathcal{P}(i)} d_{ik} + \sum_{k\in\mathcal{S}(i)} \alpha_{ik} d_{ik} + \sum_{k\in\mathcal{N}(i)} d_{ik}}$$

where $d_{ik} = -d_{\mathbb{L}}(V_c^i, T_c^k)/\tau$ denotes the hyperbolic similarity between anchor i and sample k. To mitigate structural false negatives, we introduce an adaptive weight $\alpha_{ik} = \sigma\Big(-\frac{d_{ik}}{\sum_{k\in\mathcal{P}(i)} d_{ik}+\epsilon}\Big)$ where $\sigma(\cdot)$ is the sigmoid function and ϵ ensures numerical stability. This design suppresses structurally related samples that exhibit high similarity relative to true positives, thereby preventing over-separation of semantically correlated clips and preserving hierarchical topology. We further construct the symmetric formulation by anchoring textual embeddings against visual candidates, resulting in $\mathcal{L}_{geohc}(\mathbf{T}_c, \mathbf{V}_c)$. The same geometry-aware mechanism is applied at the child level to enforce hierarchical generalization. The overall hyperbolic contrastive objective is defined as:

$$\mathcal{L}_{GeoHCL} = \frac{1}{4}(\mathcal{L}_{geohc}(\mathbf{V}_c, \mathbf{T}_c) + \mathcal{L}_{geohc}(\mathbf{T}_c, \mathbf{V}_c) + \mathcal{L}_{geohc}(\mathbf{v}_c, \mathbf{t}_c) + \mathcal{L}_{geohc}(\mathbf{t}_c, \mathbf{v}_c)).$$

Hierarchical Entailment Learning. Some prior studies [3] investigate hierarchical representation learning in hyperbolic space. In our work, we introduce a hierarchical-aware hyperbolic entailment mechanism that explicitly models multi-level semantic containment across modalities. Specifically, we adopt hyperbolic entailment cones in the Lorentz space, where each point $\mathbf{x}$ defines a conical region $\mathbb{R}_{\mathbf{x}}$ such that $\mathbf{y} \in \mathbb{R}_{\mathbf{x}}$ is semantically subsumed by $\mathbf{x}$. The half-aperture is given by $\theta(\mathbf{x}) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\|\mathbf{x}_n\|}\right)$, where c denotes curvature and K controls

behavior near the origin. The aperture shrinks with increasing $\|\mathbf{x}_n\|$, naturally enforcing tighter semantic specificity away from the origin. During training, we constrain a specific concept $\mathbf{y}$ to lie within the aperture $\theta(\mathbf{x})$ of its parent concept $\mathbf{x}$ by penalizing the angular residual of an outward point $\mathbf{y}$ with exterior angle denoted as $\varphi(\mathbf{y}, \mathbf{x}) = \frac{y_0 + c x_0 \langle \mathbf{y}, \mathbf{x} \rangle_{\mathcal{L}}}{\|\tilde{x}\| \sqrt{c \langle \mathbf{y}, \mathbf{x} \rangle_{\mathcal{L}}^2 - 1}}$, where y_0 and x_0 denoting the time dimension. Thus, the entailment loss is formulated as: $L_{ent}(\mathbf{y}, \mathbf{x}) = \max(0, \varphi(\mathbf{y}, \mathbf{x}) - \eta\theta(\mathbf{x}))$, where η is a threshold to adjust the distance region.

To explicitly model semantic containment in surgical workflows, we introduce both inter-modal and intra-modal entailment supervision in hyperbolic space. Across modalities, we treat textual concepts as semantically containing their corresponding videos (e.g., "cut" entails all cutting clips), and enforce containment at both child and parent levels via $L_{ent}(v_c, t_c)$ and $L_{ent}(V_c, T_c)$, yielding the inter-modal loss $\mathcal{L}_{inter} = L_{ent}(v_c, t_c) + L_{ent}(V_c, T_c)$. Within each modality, we further impose parent-child structural constraints to preserve the inherent tree-like organization of surgical workflows, formulated as $\mathcal{L}_{intra} = L_{ent}(t_c, T_c) + L_{ent}(v_c, V_c)$. Together, these objectives enforce hierarchical consistency both across and within modalities.

Training Paradigm. We adopt a two-stage optimization strategy to progressively stabilize hierarchical alignment in hyperbolic space. In *Stage I*, all model parameters are optimized using only the proposed geometry-aware hyperbolic contrastive loss, aiming to establish robust cross-modal alignment while mitigating structural false negatives under hierarchical semantics. In *Stage II*, we freeze the visual and textual encoders and fine-tune only the hyperbolic mapping layers. In this stage, the model is jointly optimized with the geometry-aware hyperbolic contrastive loss and the cone-based hyperbolic entailment losses (both inter- and intra-modal), explicitly enforcing hierarchical containment constraints.

3 Experiments

3.1 Experimental Details

Training Dataset. Following prior work [29], we utilize the Surgical Video Lectures (SVL) dataset shared by the authors of [29] and reorganize it into an explicit hierarchical video-text corpus for structured pretraining. SVL contains large-scale surgical lecture videos with ASR-generated transcripts and structured annotations such as keystep descriptions and video-level abstracts [28]. Instead of treating video-text pairs as flat clip-caption alignments, we explicitly construct parent-child units at three levels: fine-grained clips with local descriptions, phase-level segments with procedural summaries, and full videos with global abstracts. Parent segments are temporally composed of child clips, forming a tree-structured workflow hierarchy. This design enables supervision across modalities and across semantic levels through explicit containment relations.

Downstream Datasets and Evaluation. To validate the generalization capability of our method, we evaluate on three established surgical benchmarks:

Table 1: Zero-shot Phase Recognition across Procedures and Centers.

Model	Dataset	Cholec80	AutoLaparo	Stras70	Bern70
MIL-NCE	HowTo100M	7.8 / 7.3	9.9 / 7.9	5.6 / 3.1	2.4 / 2.1
CLIP	CLIP400M	30.8 / 13.1	17.4 / 9.1	16.9 / 5.5	14.8 / 4.1
	Scratch	29.4 / 10.4	15.3 / 10.9	6.3 / 3.5	4.9 / 2.3
	SVL	33.8 / 19.6	18.9 / 16.2	15.8 / 8.6	17.8 / 7.1
SurgVLP	SVL	34.7 / 24.4	21.3 / 16.6	10.8 / 6.9	11.4 / 7.2
HecVL	SVL	41.7 / 26.3	23.3 / 18.9	26.9 / 18.3	22.8 / 13.6
PeskaVLP	SVL	45.1 / 34.2	26.5 / 23.6	46.7 / 28.6	45.7 / 22.6
HyperVLP	SVL	**49.9 / 37.2**	**42.9 / 32.9**	**52.7 / 41.1**	**50.0 / 30.3**

Cholec80 [22] (cholecystectomy), AutoLaparo [25] (hysterectomy), and Multibypass140 [9] (gastric bypass). Multibypass140 contains 140 laparoscopic Roux-en-Y gastric bypass surgeries annotated with 12 phases and 46 steps. Following prior work, we report results separately on its two center-specific subsets, StrasBypass70 (Stras70) and BernBypass70 (Bern70), each comprising 70 cases collected from the University Hospital of Strasbourg and Bern University Hospital, respectively, to assess cross-center generalization. All these datasets differ from our training data in both surgical type and medical center, introducing substantial domain shifts. Performance is evaluated on the official test splits. In all tables, results are reported as Top-1 accuracy and F1-score (Acc/F1).

Implementation Details. We adopt ResNet-50 [5] as the visual encoder $\mathcal{F}_v$ and BioClinicalBERT [7] as the textual encoder $\mathcal{F}_t$, both initialized from pretrained weights. In Stage I, training is performed with a global batch size of 256 on 8 NVIDIA A100 GPUs using the AdamW optimizer, with an initial learning rate of 5×10^{-5}. In Stage II, training is conducted on a single GPU with batch size 128 and a reduced learning rate of 1×10^{-5}. For fair comparison, all baselines follow the same SVL dataset and protocol as PeskaVLP [27].

3.2 Results

Zero-shot Phase Recognition. Table 1 reports zero-shot phase recognition across procedures (Cholec80, AutoLaparo) and centers (Stras70, Bern70). General-purpose models (MIL-NCE [15], CLIP [20]) show limited transferability, especially under cross-procedure and cross-center settings. Surgical VLMs pretrained on SVL substantially improve performance, confirming the benefit of domain-specific pretraining. Among all methods, HyperVLP consistently achieves the best results on all four benchmarks, with clear margins on both Accuracy and F1-score, particularly under cross-center evaluation (Stras70 and Bern70). Notably, HecVL and PeskaVLP are themselves Euclidean *hierarchical* surgical VLP methods built on the same SVL data and backbone. The consistent improvement of HyperVLP over them isolates the benefit of modeling the workflow hierarchy in hyperbolic rather than Euclidean space, beyond non-hierarchical

Table 2: Linear-probing evaluation results. V: supervision is from visual frames. L: supervision is from natural languages. VL: supervision is from both visual and language entities.

Model	Dataset	shot	Cholec80	Autolaparo	Stras70	Bern70
ImageNet	ImageNet	100	66.4 / 54.9	57.5 / 44.9	66.2 / 53.6	64.7 / 31.6
		10	57.4 / 42.3	44.9 / 30.4	53.3 / 42.1	53.3 / 25.6
Moco [4]	SVL(V)	100	68.2 / 55.8	59.5 / 48.4	71.6 / 58.1	69.6 / 36.5
		10	57.6 / 43.5	49.9 / 34.6	63.1 / 49.3	59.1 / 29.9
Moco [4]	Cholec80	100	73.4 / 62.8	51.3 / 37.4	67.8 / 55.4	66.0 / 33.1
		10	69.6 / 56.9	45.4 / 31.7	58.1 / 45.2	52.7 / 25.7
CLIP	NA(L)	100	64.8 / 50.7	58.5 / 46.1	65.4 / 50.6	64.1 / 33.3
		10	57.5 / 40.0	46.2 / 31.4	54.3 / 42.1	52.8 / 27.9
CLIP	SVL(L)	100	64.9 / 55.0	53.1 / 42.1	69.1 / 55.7	68.2 / 35.2
		10	58.9 / 42.3	45.3 / 35.3	58.2 / 45.2	56.5 / 29.8
SurgVLP	SVL(L)	100	63.5 / 50.3	54.3 / 41.8	65.8 / 50.0	66.5 / 34.3
		10	55.0 / 39.9	48.5 / 32.0	57.0 / 44.0	57.7 / 28.5
HecVL	SVL(L)	100	66.0 / 53.2	56.9 / 44.2	69.8 / 54.9	70.0 / 34.4
		10	56.1 / 40.3	46.9 / 32.1	60.2 / 46.8	59.3 / 31.2
PeskaVLP	SVL(VL)	100	69.9 / 59.8	63.1 / 49.7	71.4 / 59.5	71.5 / 37.4
		10	61.9 / 50.6	53.1 / 36.8	63.8 / 50.4	62.9 / 32.7
HyperVLP	SVL(VL)	100	**71.3 / 60.2**	**65.3 / 51.1**	**75.6 / 62.3**	**76.9 / 45.1**
		10	**64.2 / 51.3**	**60.1 /43.8**	**66.9 / 53.4**	**72.3 / 44.8**

CLIP/SurgVLP-style baselines.

Few-/Full shot Linear Probing. Following the protocol of PeskaVLP [27], we evaluate HyperVLP under 10% and 100% settings with frozen encoders and a linear classifier using V, L, and VL supervision. As shown in Table 2, general-purpose models show limited cross-procedure transfer, especially in the few-shot regime. Surgical pretraining improves performance, while HyperVLP consistently achieves the best results across datasets and supervision settings. The pronounced gains in the 10-shot setting indicate that our hyperbolic hierarchical modeling yields more linearly separable and procedure-agnostic representations.

Qualitative Analysis. To analyze how different semantic granularities are organized in hyperbolic space, we examine the distribution of embedding distances to the origin, where the radial coordinate reflects the level of hierarchical abstraction. Figure 2 presents the radial distance distributions of step- and phase-level video embeddings, along with their paired textual representations. The textual descriptions are generated with the assistance of a large language model [8] to ensure semantic consistency. We observe a clear separation between semantic levels as well as between modalities. Importantly, the relative ordering of step- and phase-level representations remains consistent across StrasBypass70 (a) and

Table 3: Ablation study. A checkmark (✓) indicates the model trained with the loss function, while a dash ("-") indicates the model trained without it.

$\mathcal{L}_{GeoHCL}$	$\mathcal{L}_{inter}$	$\mathcal{L}_{intra}$	Cholec80	AutoLaparo	Stras70	Bern70
-	-	-	33.8 / 19.6	18.9 / 16.2	15.8 / 8.6	17.8 / 7.1
✓	-	-	43.1 / 28.2	30.3 / 24.2	45.7 / 27.6	36.9 / 21.7
✓	✓	-	45.1 / 27.2	39.8 / 27.2	47.5 / 38.3	46.5 / 26.4
✓	-	✓	47.6 / 29.8	40.9 / 29.6	47.3 / 37.3	44.1 / 28.8
✓	✓	✓	**49.9 / 37.2**	**42.9 / 32.9**	**52.7 / 41.1**	**50.0 / 30.3**

BernBypass70 (b), indicating that the learned hyperbolic space preserves a stable and procedure-invariant hierarchical structure.

Ablation Study. We evaluate the contribution of each loss component to phase recognition. As shown in Table 3, adding the geometry-aware contrastive loss ($\mathcal{L}_{GeoHCL}$), inter-modal entailment ($\mathcal{L}_{inter}$), and intra-modal entailment ($\mathcal{L}_{intra}$) progressively improves performance across datasets. $\mathcal{L}_{GeoHCL}$ and $\mathcal{L}_{inter}$ provide the primary gains, while $\mathcal{L}_{intra}$ further enhances hierarchical consistency. The best results are achieved when all components are jointly optimized.

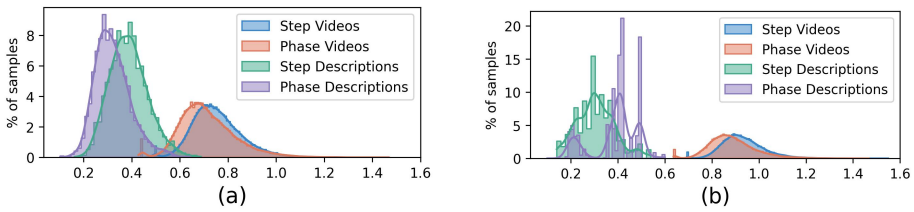

Fig. 2: Distribution of hyperbolic distances to the origin for video and text representations. The x-axis shows the hyperbolic radial distance of each embedding. Results are shown for Strasbypass70 (a) and BernBypass70 (b), including step-/phase-level videos and their corresponding descriptions.

4 Conclusion

In this paper, we presented HyperVLP, a hyperbolic surgical video-language pretraining framework that explicitly models the intrinsic tree structure of surgical workflows. Unlike prior Euclidean approaches that treat workflow labels as flat categories, our method embeds visual and textual representations into a shared Lorentz manifold to preserve hierarchical containment. We introduced a geometry-aware hyperbolic contrastive objective to alleviate structural false negatives and a cone-based entailment mechanism to enforce parent-child consistency across and within modalities. A progressive two-stage optimization strat-

egy further stabilizes hierarchical alignment in hyperbolic space. Extensive experiments on three publicly available benchmarks demonstrate consistent improvements in zero-shot and few-shot phase recognition, particularly under cross-procedure and cross-center domain shifts.

5 Acknowledgments

This research was partially supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang under Grant No. 2025C01132. This work benefited from French state funds managed by the National Research Agency via the IHU Strasbourg (ANR-10-IAHU-0002) and the ENACT AI Cluster (ANR-23-IACL-0004).

6 Disclosure of Interests

The authors declare that they have no competing interests.

References

1. Cannon, J.W., Floyd, W.J., Kenyon, R., Parry, W.R., et al.: Hyperbolic geometry. Flavors of geometry **31**(59-115), 2 (1997)
2. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, R.: Hyperbolic image-text representations. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
3. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic entailment cones for learning hierarchical embeddings. In: International conference on machine learning. pp. 1646–1655. PMLR (2018)
4. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. He, Y., Zhu, Y., Fu, P., Yang, R., Chen, T., Wang, Z., Li, Q., Zhou, P., Yang, X., Wang, S.: Endo-clip: Progressive self-supervised pre-training on raw colonoscopy records. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 106–116. Springer (2025)
7. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
8. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
9. Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N.: Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. International journal of computer assisted radiology and surgery pp. 1–9 (2024)

10. Le, M., Roller, S., Papaxanthos, L., Kiela, D., Nickel, M.: Inferring concept hierarchies from text corpora via hyperbolic embeddings. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3231–3241. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1313, https://aclanthology.org/P19-1313/
11. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
12. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 525–536. Springer Nature Switzerland, Cham (2023)
13. Liu, Y., Cao, X., Chen, T., Jiang, Y., You, J., Wu, M., Wang, X., Feng, M., Jin, Y., Chen, J.: A survey of embodied ai in healthcare: Techniques, applications, and opportunities. arXiv preprint arXiv:2501.07468 (2025)
14. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al.: Surgical data science–from concepts toward clinical translation. Medical image analysis **76**, 102306 (2022)
15. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2630–2640 (2019)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
17. Pal, A., Van Spengler, M., di Melendugno, G.M.D., Flaborea, A., Galasso, F., Mettes, P.: Compositional entailment learning for hyperbolic vision-language models. arXiv preprint arXiv:2410.06912 (2024)
18. Perez, A., Nwoye, C., Kermani, R.R., Mohareri, O., Jamal, M.A.: Surglavi: Large-scale hierarchical dataset for surgical vision–language representation learning. Medical Image Analysis p. 103982 (2026)
19. Pérez, A., Rodríguez, S., Ayobi, N., Aparicio, N., Dessevres, E., Arbeláez, P.: MuST: Multi-Scale Transformers for Surgical Phase Recognition . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15006. Springer Nature Switzerland (October 2024)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
21. Schmidgall, S., Kim, J.W., Jopling, J., Krieger, A.: General surgery vision transformer: A video pre-trained foundation model for general surgery. arXiv preprint arXiv:2403.05949 (2024)
22. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
23. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International conference on medical image computing and computer-assisted intervention. pp. 101–111. Springer (2023)

24. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)
25. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 486–496. Springer (2022)
26. Yang, S., Luo, L., Wang, Q., Chen, H.: Surgformer: Surgical Transformer with Hierarchical Temporal Attention for Surgical Phase Recognition . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15006. Springer Nature Switzerland (October 2024)
27. Yuan, K., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. Advances in Neural Information Processing Systems **37**, 122952–122983 (2024)
28. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 306–316. Springer (2024)
29. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. Medical Image Analysis **105**, 103644 (2025)