

# A Hierarchy-Aware Video-Language Model Evaluation and Hyperbolic Baseline for Surgery

Ana Manzano Rodriguez<sup>1,2,3,4</sup> Pascal Mettes<sup>1,2</sup> Marlies P. Schijven<sup>1,3,4,5</sup> Cees G. M. Snoek<sup>1,2</sup>

<sup>1</sup>Data Science Center HAVA-Lab, University of Amsterdam

<sup>2</sup>Video & Image Sense Lab, Informatics Institute, University of Amsterdam

<sup>3</sup>Amsterdam UMC Location University of Amsterdam, Surgery <sup>4</sup>Amsterdam Public Health, Digital Health

<sup>5</sup>Amsterdam Gastroenterology and Metabolism

## Abstract

Surgical procedures follow a phase-to-step hierarchy, yet the video-language models used to recognize them are evaluated with flat per-level metrics that ignore cross-level coherence and error structure. In this paper we make two contributions to address this problem, (i) we introduce SURGHIBENCH, the first hierarchy-aware evaluation suite for surgical video understanding, with three tasks measuring recognition, consistency, and severity across granularity levels. We evaluate a general-purpose CLIP model, a Euclidean surgical model, and, as second contribution: (ii) HYPERSURG, a new hyperbolic model that enforces phase-step containment via entailment cones, across four (existing) datasets spanning three procedure types. The suite reveals that two models with the same accuracy can produce predictions of very different error severity, ranging from sibling confusions within the correct phase to unrelated cross-phase predictions. Hyperbolic geometry shifts predictions toward the correct procedural neighborhood, and these gains scale with the tree-likeness of each dataset’s annotation hierarchy, providing a principled indicator when hierarchy-aware geometry helps.

## 1. Introduction

Developing assistance systems for the operating room requires understanding what is happening during a procedure [13]. Surgical workflow analysis provides this understanding by recognizing surgical activities at different levels of temporal granularity, which is inherently hierarchical: a gallbladder removal (cholecystectomy) comprises *phases* (e.g. Calot triangle dissection), each containing *steps* (e.g. dissecting the cystic duct), which in turn consist of fine-grained *actions* (e.g. grasping, cutting) (Figure 1). Because steps belong to broader procedural contexts, not all recognition errors are equal [3, 24]. Confusing two steps within the same phase means the prediction is at least in the right

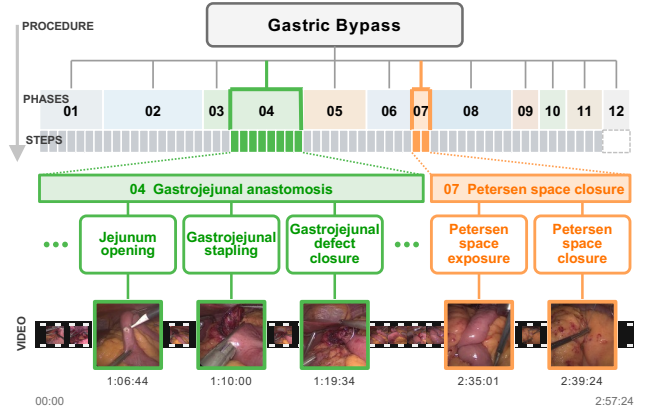


Figure 1. **Surgical procedures are hierarchically structured.** A gastric bypass decomposes into 12 phases containing 46 steps. For surgical AI, current video-language models embed all levels in a flat space and evaluate each independently, ignoring underlying structure. To address this problem, we introduce (i) a hierarchy-aware evaluation suite and (ii) a hyperbolic baseline model.

part of the procedure, whereas predicting a step from an entirely different phase places it in an unrelated procedural context. Effective surgical support therefore requires recognizing both the broad phase and the specific step, and ensuring that these predictions are consistent.

Despite this clinical need, current surgical video-language models do not explicitly capture hierarchical relationships. State-of-the-art models such as SurgVLP [33] learn from narrated surgical video without hierarchical structure, and while the SurgLaVi model [18] learns from hierarchically annotated data, all levels still coexist in the model’s flat embedding space without explicit hierarchical distinction. Evaluation has not addressed this gap either, as all current models are measured with accuracy or F1 at each level *independently*, without assessing whether predictions across levels are consistent or whether errors respect the surgical procedure structure. In practice, two models with the same score may produce errors of very different

severity, with standard metrics not being able to distinguish between them. This disconnect between what we evaluate and what clinical deployment demands has been identified as a key barrier for AI adoption in surgery [23].

To address the lack of hierarchy-awareness in surgical AI, this paper makes two core contributions:

1. **Contribution I: SURGHIBENCH** (Sec. 3), the first hierarchy-aware evaluation suite for surgical video-language models, with three tasks measuring per-level and joint recognition, cross-level consistency, and error severity, evaluated across four (existing) video datasets spanning three procedure types. SURGHIBENCH comes with three baseline models: CLIP, SurgCLIP<sup>\*</sup><sub>( $\beta$ )</sub>, and
2. **Contribution II: HYPERSURG** (Sec. 4), a new hierarchy-aware surgical video-language baseline model that enforces entailment cone constraints over a hyperbolic embedding space, where the exponentially growing volume naturally fits the tree-like structure of surgical taxonomies, paired with a hierarchy-aware triplet batch sampler.

SURGHIBENCH reveals (Sec. 5) that models with comparable accuracy differ substantially in how coherent their predictions are. With HYPERSURG, predictions shift from unrelated phases toward the correct procedural neighborhood, and cross-level contradictions decrease by up to 13 pp. Step recognition benefits most from the hierarchical inductive bias (+9.3 pp on average), as the exponentially growing space provides room for fine-grained distinctions that flat geometry collapses. Before detailing our contributions we first discuss related work.

## 2. Related Work

**Surgical Vision-Language Models.** While vision-language models such as CLIP [20] have transformed general visual recognition for quite a while, their adaptation to surgery is still in its early stages. SurgVLP [33] was the first to apply contrastive pretraining to narrated surgical video, using a ResNet-50 vision encoder aligned with a BioClinicalBERT text encoder. SurgLaVi [18] contributed a large-scale dataset of hierarchically annotated surgical clips spanning actions, steps, and phases, together with SurgCLIP, a dual-encoder model trained on this data. Subsequent work has aimed to make these representations hierarchy-aware: HecVL [31] constructs three separate embedding levels (clip, phase, video) via progressive aggregation of fine-grained representations into coarser ones, while PeskaVLP [32] enriches textual descriptions with LLM-generated hierarchical knowledge and procedural alignment. These methods rely on increasingly complex architectures (separate embedding spaces, progressive aggregation, LLM-augmented knowledge) yet all remain in Euclidean space. HYPERSURG takes an alternative approach: rather than adding architectural complexity, we change the

underlying geometry, unifying all levels in a single hyperbolic space where the structure itself encodes hierarchical granularity.

**Hierarchical Evaluation.** Not all misclassifications are equally costly: confusing adjacent categories in a semantic hierarchy is less severe than confusing distant ones, and evaluation metrics should reflect this. Bertinetto *et al.* [3] showed that classifiers trained to minimize tree distance to the ground truth sacrifice minimal flat accuracy while producing substantially milder errors. This principle has found clinical uptake. In dermatology, hierarchy-aware prototypes penalize confusing melanoma with a benign nevus (a distant branch) more heavily than confusing it with basal cell carcinoma (a sibling malignancy), reducing dangerous misdiagnoses [30]. In radiology, Srivastava and Mishra [24] showed that models with comparable recognition ability on chest X-rays can differ substantially in error severity. Moreover, existing hierarchy-aware training losses and post-inference reranking methods, designed for taxonomies where only leaf nodes are labels, actually *increase* severity on medical hierarchies where every node is a valid diagnosis. This suggests that reducing error severity in multi-level clinical hierarchies requires representations that intrinsically respect the hierarchy. For action recognition, Long *et al.* [12] introduced sibling-mAP and cousin-mAP, which relax correctness criteria along a category tree. However, all of these frameworks target natural-image classification or general action recognition; none has been applied to surgical video, where the procedural hierarchy provides an equally natural tree structure for grading error severity. Within surgical workflow analysis, temporal hierarchy has been studied through phase recognition [4], step recognition [11], and instrument-action detection [22], yet these tasks are invariably evaluated independently with flat metrics. Our cross-level consistency evaluation is, to our knowledge, the first to explicitly measure whether a model’s predictions *across* granularity levels are mutually consistent.

**Hyperbolic Representations.** Hyperbolic spaces are non-Euclidean geometries of constant negative curvature whose volume grows exponentially with radius, making them a natural fit for encoding hierarchies [14]. Poincaré embeddings [15] showed that hyperbolic spaces can embed taxonomies with orders-of-magnitude fewer dimensions than Euclidean ones, and Ganea *et al.* [7] introduced a geometric notion of containment in hyperbolic space, where broader concepts enclose more specific ones within cone-shaped regions. Building on these ideas, MERU [5] was the first to apply hyperbolic geometry to vision-language pretraining, learning a two-level cross-modal hierarchy (text  $\supset$  image) on natural images. HyCoCLIP [16] extended entailment to compositional sub-parts, and PHYCLIP [29] proposed products of hyperbolic spaces for joint hierarchy and compositionality. In medical imaging, HYDEN [19] applied hy-

perbolic representations to radiology reports. We propose HYPERSURG, a hyperbolic surgical video-language model that enforces the procedural phase-step hierarchy through entailment cones. Concurrent to our work, HyperVLP by Hu *et al.* [9] also represents surgical video and language in a Lorentz manifold and use hyperbolic entailment to encode hierarchical relations. However, both HyperVLP and prior work evaluate with standard flat metrics, leaving open the hierarchy-aware questions our benchmark addresses.

### 3. Hierarchy-Aware Evaluation for Surgery

We introduce SURGHIBENCH, the first hierarchy-aware evaluation suite for surgical video-language models. SURGHIBENCH combines three complementary tasks that together provide a complete picture of hierarchical awareness: whether predictions at each granularity level are correct, whether they are mutually consistent along the levels, and whether predictions land on related or unrelated surgical concepts (Sec. 3.1). These tasks are evaluated across four (existing) surgical video datasets that span a spectrum of tree-likeness in their annotation structure (Sec. 3.2) and three models ranging from general-purpose to hierarchy-aware (Sec. 3.3). SURGHIBENCH is model-agnostic and can evaluate any surgical video-language model that produces phase and step predictions. Each task is defined below together with its metrics, all computed per-video and averaged. All data, training details and source code will be made available on our benchmark website.

#### 3.1. Hierarchy-Aware Tasks

**Task 1: Single- and Multi-Level Recognition.** *How well does a model recognize phases and steps, individually and jointly?* We first evaluate how well a model identifies phases and steps independently, and report video-averaged *accuracy* and *macro-averaged F1* at each level. While accuracy measures overall correctness, macro-F1 better captures performance on rare but clinically important classes by weighting all classes equally, which is important given the severe class imbalance typical of surgical datasets. A model may achieve high phase accuracy while performing poorly on steps, or vice versa; to detect such imbalances we also report a *multi-level* score, defined as the geometric mean of phase and step accuracy:

$$\text{Multi-level} = \sqrt{\text{Acc}_{\text{phase}} \times \text{Acc}_{\text{step}}}, \quad (1)$$

which remains low unless both levels are strong (*e.g.* 30% phase with 10% step yields 17.3%, while 20% on both yields 20%). These per-level metrics do not show, however, whether predictions are coherent across levels, which the remaining tasks address.

**Task 2: Cross-Level Consistency.** *Are the model’s phase and step predictions mutually consistent?* A system that

predicts the *Preparation* phase alongside a *Clipping* step produces contradictory outputs that could mislead clinical decision support [28]. We measure joint correctness across both levels with three metrics. The *Both-wrong* and *Both-right* rates measure how often both levels are simultaneously wrong or right, revealing whether errors are isolated to one level or propagate across both. *Step|Phase* is step accuracy conditioned on the phase being correct, which isolates step discrimination from phase errors. These metrics are inspired by hierarchical consistency measures in image classification [10, 17], adapted to surgical video, where only joint phase-step correctness has been reported [21] without analyzing error patterns across levels.

**Task 3: Hierarchy-Aware Error Severity.** *Does the model confuse related or unrelated surgical concepts?* We capture this with two complementary measures. *Sibling step accuracy* [12] measures how often the predicted step falls within the same phase as the ground-truth step, even if the exact step is wrong. *Cousin step accuracy* measures the opposite: how often the prediction lands in an entirely different phase. A model with high sibling and low cousin accuracy predicts within the correct procedural context, indicating that it captures the coarse structure of the procedure. We also introduce *semantic step accuracy*, which measures whether the model recognizes the type of surgical action being performed regardless of anatomical site (*e.g.* *Jejunojejunal Defect Closure* vs. *Gastrojejunal Defect Closure*), identifying *what* is being done even when it confuses *where*. We manually group steps that share the same surgical action across different anatomical sites; full groupings are in the supplemental (Sec. A.1).

#### 3.2. Surgical Video Datasets

**Evaluation Datasets.** Our evaluation focuses on phase and step recognition, the two granularity levels that enable the cross-level analysis required by our three tasks. We evaluate on four video datasets (Tab. 1): MIPO [8] (50 radius fracture fixation videos), StrasbyPass70 and BernBypass70 [11] (70 gastric bypass videos each, recorded at two institutions), and GraSP [1] (13 robot-assisted prostatectomy videos). All four provide phase and step annotations, but differ in how steps relate to phases. In some datasets, each step belongs to exactly one phase, forming a clean tree (MIPO, 100%). In others, steps recur across multiple phases, forming a directed acyclic graph (DAG) where the parent-child relationship is ambiguous (GraSP, 33%). This difference stems from how each dataset defines its steps. GraSP’s step vocabulary mixes procedural substages (*Seminal vesicle dissection*, *Prevesical dissection*) with reusable instrument actions (*Suction*, *Cut suture or tissue*, *Tie suture*). The latter are phase-agnostic—*Suction* occurs whenever there is blood or fluid (10 of 11 phases), *Cut suture or tissue* whenever tissue must be separated (10 of 11). By contrast, the bypass

Table 1. **SURGHIBENCH evaluates video-language models on hierarchy-awareness**, covering three tasks and four surgical video datasets. Datasets are ordered by tree-likeness, defined as the percentage of steps that belong to exactly one phase (100% = pure tree, lower = more DAG-like).

| Dataset       | Source                      | Procedure       | Videos | Phases <sup>†</sup> | Steps <sup>†</sup> | Prompts       | Tree-likeness |
|---------------|-----------------------------|-----------------|--------|---------------------|--------------------|---------------|---------------|
| MIPO          | Graeff <i>et al.</i> [8]    | Radius fracture | 50     | 5                   | 8                  | Ours          | 100%          |
| StrasbyPass70 | Lavanchy <i>et al.</i> [11] | Gastric bypass  | 70     | 11                  | 46                 | SurgLaVi [18] | 74%           |
| BernBypass70  | Lavanchy <i>et al.</i> [11] | Gastric bypass  | 70     | 11                  | 46                 | SurgLaVi [18] | 54%           |
| GraSP         | Ayobi <i>et al.</i> [1]     | Prostatectomy   | 13     | 11                  | 21                 | SurgLaVi [18] | 33%           |

<sup>†</sup>Classes used for evaluation; see supplementary Sec. A.2 for details.

datasets (54–74%) mostly define steps as procedural sub-stages (*Gastrojejunal Defect Closure*, *Retrogastric Dissection*, *Vertical Stapling*) that describe the surgical goal being accomplished, and are therefore largely phase-exclusive by construction. MIPO follows the same goal-level convention (*Distal fixation*, *Proximal fixation*). This spectrum allows our evaluation to also test whether the benefits of hierarchy-aware representations are tied to the structural properties of each dataset’s annotation hierarchy.

For zero-shot evaluation, we use the text descriptions provided by Perez *et al.* [18] for GraSP and the bypass datasets; for MIPO, we wrote new prompts following the same narration style (see supplemental Sec. A.2).

**Pretraining Dataset.** The surgical video-language models evaluated in this work are pretrained on a subset of the SurgLaVi dataset [18]. The full dataset contains 5,317 narrated surgical videos sourced from diverse surgical domains, across specialties such as general surgery, colorectal, gynecology, thoracic, and urology. Of these, 2,464 are part of the open-source subset, called SurgLaVi- $\beta$ . We use SurgLaVi- $\beta$  without modification.

Each video is divided into temporal segments at three hierarchy levels, with captions generated from the narrator’s descriptions, yielding  $\sim 113$ K caption–clip pairs. These levels differ in temporal granularity and scope: *Actions* (Level 1) capture brief, action-focused surgical moments (*e.g.* “the surgeon uses a stapler to make the first transverse staple line”). *Steps* (Level 2) aggregate multiple actions into intermediate procedural substages with more context (*e.g.* “the surgeon dissects into the lesser sac; the first transverse stapler firing is performed”). *Phases* (Level 3) span longer temporal segments with rich procedural detail (*e.g.* “the surgeon begins the gastric bypass by placing trocars at the angle of His, dissecting along the lesser curvature with a perigastric technique”). This hierarchical structure mirrors the procedural hierarchy of surgery itself, making hierarchy-aware training possible and allowing the model to learn representations that respect the natural granularity of surgical procedures.

### 3.3. Baseline Models

Our evaluation suite comes with three models. Two existing models and one model that we propose as part of this paper. (i) *CLIP ViT-B/16* [20], the original OpenAI CLIP model with no surgical pretraining. Since CLIP has no temporal modeling, each frame is encoded independently and similarities are averaged over the temporal window. It serves as a lower bound on what general-purpose vision-language representations achieve on surgical tasks. (ii) *SurgCLIP $^*_{(\beta)}$*  is our reimplementation of SurgCLIP $_{(\beta)}$  by Perez *et al.* [18] with identical architecture (TimeSformer [2] + BERT-base [6]), training procedure, and Euclidean cosine similarity. Since SurgLaVi released model and evaluation code but not the training pipeline, we reimplemented training from the details in their paper; the reproduction closely matches their reported results across all benchmarks (see Tab. 8 in supplemental), confirming a faithful reimplementation. We follow their zero-shot evaluation protocol and test splits, adapting the scripts for Lorentz distance in HYPERSURG. (iii) HYPERSURG is our hierarchy-aware model operating on a Lorentz hyperboloid, described next, using the same architecture, training data, and number of parameters as SurgCLIP $^*_{(\beta)}$ , ensuring a controlled experiment where any difference with our model is attributable to the embedding geometry alone.

## 4. Hyperbolic Baseline Model for Surgery

Having defined how to evaluate hierarchy-awareness, we now propose a baseline model that encodes surgical hierarchy by operating in hyperbolic space, where the temporal granularity of procedures is naturally represented. We cover the geometric background (Sec. 4.1), the architecture (Sec. 4.2), and the training objectives (Sec. 4.3).

### 4.1. Background: Lorentz Hyperbolic Space

Standard contrastive models, like CLIP and SurgCLIP, measure similarity via cosine distance, which depends only on the angle between embeddings and discards their norms. This leaves a single degree of freedom—angular direction—to simultaneously separate semantically distinct videos and distinguish hierarchy levels within the same

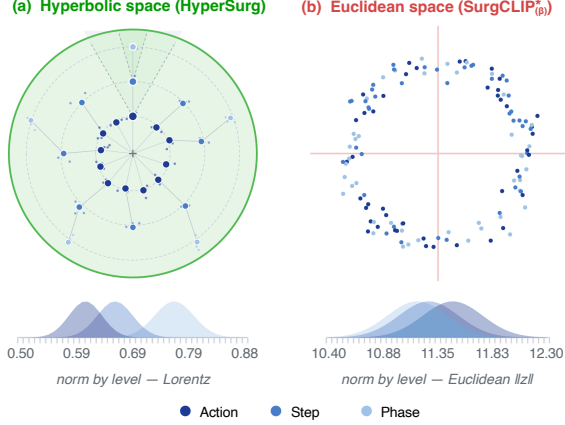


Figure 2. **Hyperbolic geometry organizes the surgical hierarchy by embedding norm; Euclidean cosine space collapses it.** (a) In HYPERSURG (green), embedding norms increase from fine-grained actions to coarse phases, reflecting the hierarchical structure. Gray wedges are entailment cones (Equation (7)). (b) In SurgCLIP $^*_{(\beta)}$  (red), L2 normalization for cosine similarity discards any norm signal, preventing cosine-based models from exploiting the hierarchy. **Bottom:** per-level norm distributions. Note distinguishability of hierarchical levels in hyperbolic space.

video, competing demands that one signal cannot satisfy. The Lorentz hyperboloid decouples these two roles: the distance from the origin encodes generality while the angular direction independently encodes semantic content, providing the missing degree of freedom. Formally, it is defined as the upper part of a two-sheeted hyperboloid in  $(n+1)$ -dimensional Minkowski space with curvature  $-1/c$ :

$$\mathbb{L}_c^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{L}} = -1/c, x_0 > 0\}, \quad (2)$$

where  $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}} = -x_0 y_0 + \sum_{i=1}^n x_i y_i$  is the Lorentz inner product. The geodesic distance between two points is:

$$d_{\mathbb{L}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{c}} \operatorname{acosh}(-c \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}}). \quad (3)$$

A key property is that the volume of a ball of radius  $r$  grows with ratio  $e^{(n-1)r/\sqrt{c}}$ , whereas in Euclidean space the same volume grows only polynomially ( $\propto r^n$ ). This exponential capacity makes the radial axis a natural encoding of hierarchy, separating granularity levels that flat geometry collapses. Figure 2 illustrates this property: in HYPERSURG, embedding norms increase monotonically from fine-grained actions to coarse phases, whereas L2 normalization in cosine-based models collapses this separation entirely.

## 4.2. Architecture

HYPERSURG follows a dual-encoder paradigm where a video encoder and a text encoder are jointly trained to align surgical clips with hierarchical captions in a shared hyperbolic embedding space (Figure 3). To isolate the effect of geometry, HYPERSURG uses the same encoders as

SurgCLIP $^*_{(\beta)}$ : a TimeSformer [2] (ViT-B/16, divided space-time attention, 12 layers) encodes  $T=16$  uniformly sampled frames into a 768-dimensional video representation via mean-pooled spatial patch tokens, and a BERT-base [6] text encoder maps captions to 768-dimensional text representations via the [CLS] token.

The two architectures diverge after the encoders. Both project the 768-dimensional representations to  $\mathbb{R}^{256}$  via learned linear layers, but SurgCLIP $^*_{(\beta)}$  then applies L2-normalization to measure cosine similarity, while HYPERSURG scales them by learnable per-modality log-space factors  $\alpha_v, \alpha_t$  and maps them onto the Lorentz hyperboloid  $\mathbb{L}_c^{256}$  via the exponential map at the origin  $\mathbf{o} = (1/\sqrt{c}, \mathbf{0})$ :

$$\exp_{\mathbf{o}}^c(\mathbf{v}) = \cosh(\sqrt{c} \|\mathbf{v}\|_{\mathbb{L}}) \mathbf{o} + \frac{\sinh(\sqrt{c} \|\mathbf{v}\|_{\mathbb{L}})}{\sqrt{c} \|\mathbf{v}\|_{\mathbb{L}}} \mathbf{v}, \quad (4)$$

where  $\mathbf{v} \in T_{\mathbf{o}}\mathbb{L}_c^n$  is the tangent vector at  $\mathbf{o}$ . Only the space components are stored; the time component is recovered via the hyperboloid constraint  $x_0 = \sqrt{1/c + \|\mathbf{x}_{\text{space}}\|^2}$  when needed for distance computations. The curvature  $c$  is a learnable parameter stored in log-space and clamped to  $[0.1, 10]$ .

## 4.3. Training Objectives

**Lorentz Contrastive Loss.** We measure similarity via negative Lorentz distance scaled by a learnable temperature  $\tau$ :

$$s(\mathbf{v}_i, \mathbf{t}_j) = -d_{\mathbb{L}}(\mathbf{v}_i, \mathbf{t}_j) / \tau. \quad (5)$$

The loss follows symmetric InfoNCE over all GPU-gathered pairs:

$$\begin{aligned} \mathcal{L}_{\text{contr}} &= \frac{1}{2} (\mathcal{L}_{v2t} + \mathcal{L}_{t2v}), \\ \mathcal{L}_{v2t} &= -\frac{1}{B} \sum_i \log \frac{e^{s(\mathbf{v}_i, \mathbf{t}_i)}}{\sum_j e^{s(\mathbf{v}_i, \mathbf{t}_j)}}. \end{aligned} \quad (6)$$

**Hierarchy Entailment Loss.** The three-level hierarchy described in Sec. 3.2 forms a tree. Following Ganea *et al.* [7], we enforce this containment via entailment cones. For a parent text embedding  $\mathbf{p}$ , its entailment cone has half-aperture:

$$\psi(\mathbf{p}) = \operatorname{asin}\left(\frac{2r_{\min}}{\|\mathbf{p}\|_{\mathbb{R}} \sqrt{c}}\right), \quad (7)$$

where  $r_{\min}$  controls the minimum cone width. A child  $\mathbf{q}$  (e.g. a step or phase caption from the same video) must lie within the parent’s cone. Let  $\phi(\mathbf{p}, \mathbf{q})$  denote the exterior angle at  $\mathbf{p}$  in the hyperbolic triangle  $O\text{--}\mathbf{p}\text{--}\mathbf{q}$  (where  $O$  is the origin). The entailment loss penalizes violations with a slack factor  $\eta$ :

$$\mathcal{L}_{\text{entail}} = \frac{1}{|\mathcal{P}|} \sum_{(\mathbf{p}, \mathbf{q}) \in \mathcal{P}} \max(0, \phi(\mathbf{p}, \mathbf{q}) - \eta \psi(\mathbf{p})), \quad (8)$$

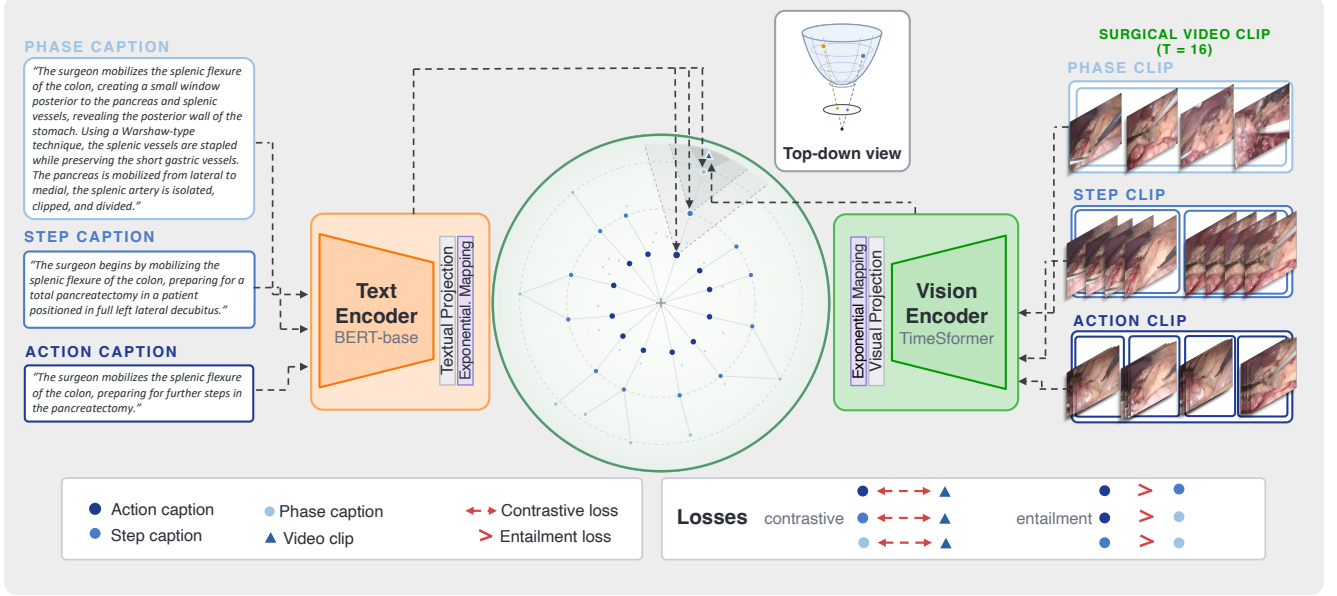


Figure 3. **Overview of the proposed hierarchy-aware HYPERSURG baseline model.** Surgical video clips at three hierarchy levels (action, step, phase) are encoded by a TimeSformer (ViT-B/16), while the corresponding narration-derived captions are encoded by BERT-base. Both branches project into a shared 256-D Lorentz hyperboloid  $\mathbb{L}_c^{256}$  via learned projections and the exponential map. The contrastive loss  $\mathcal{L}_{\text{contr}}$  aligns video-text pairs via Lorentz distance (Equation (5)); the entailment loss  $\mathcal{L}_{\text{entail}}$  enforces parent-child containment across hierarchy levels within the same cone (Equation (8)).

where  $\mathcal{P}$  is the set of parent-child pairs within each batch and  $\eta=1.2$  provides geometric slack.

**Hierarchical Triplet Sampling.** During training, each batch consists of video-caption pairs sampled from the dataset. Standard random sampling rarely produces within-video hierarchical pairs: with captions spread across thousands of videos, a random batch of a few hundred samples is unlikely to contain multiple hierarchy levels from the same video, leaving the entailment loss effectively without training signal. We introduce a novel hybrid sampler that composes each batch of  $B$  samples from  $B-3k$  random samples and  $k=3$  forced *triplets*, each containing one caption per hierarchy level from the same video. This guarantees  $\sim 9$  entailment pairs per batch while maintaining contrastive diversity.

**Total Loss.**

$$\mathcal{L} = \mathcal{L}_{\text{contr}} + \lambda \mathcal{L}_{\text{entail}}, \quad (9)$$

with  $\lambda=0.1$ , selected from a range of values (Tab. 10). An ablation study confirms that both the entailment cones and the hierarchy-aware triplet batch sampler contribute beyond Lorentz distance alone (Sec. B).

**Training details.** We train for 50 epochs on  $8 \times$  NVIDIA RTX A5000 GPUs with an effective batch size of 312, using AdamW ( $\text{lr}=10^{-4}$ , weight decay = 0.02) with OneCycleLR (5-epoch warmup, cosine annealing). Video frames are augmented with random resized crop (scale 0.5–1.0) and horizontal flip. Since no training split is provided

with the dataset, we hold out 15% of the videos (369 out of 2,464) for monitoring training. Both HYPERSURG and SurgCLIP $^*_{(\beta)}$  are trained with the same procedure, data, and hyperparameters, including masking same-video captions from the contrastive denominator to prevent conflicting gradients from clips that share a source video. Full details are provided in supplemental Sec. A.4.

## 5. Results

### 5.1. Task 1: Single- and Multi-Level Recognition

We present results for Task 1 in Table 2, where we compare recognition performance across granularity levels. The CLIP baseline, which has no surgical pretraining, provides a cautionary note about interpreting accuracy in isolation. On MIPO, CLIP achieves higher accuracy than the Euclidean surgical model at both levels (16.2% vs 13.6% phase, 17.0% vs 8.3% step), yet its F1 is lower in both cases (6.6 vs 7.9, 4.9 vs 5.6), indicating that CLIP collapses predictions into the most frequent classes, inflating accuracy while failing to distinguish rarer categories. The pattern holds at the aggregate level, where CLIP’s overall F1 remains near chance (4.1 phase, 1.8 step).

Surgical pretraining in Euclidean space substantially improves phase recognition (23.0% accuracy, F1 16.4), but step accuracy lags behind (14.2%). These results indicate that coarse-grained surgical categories are already separa-

Table 2. **Task 1: Single- and Multi-Level Recognition.** Zero-shot video-averaged accuracy (%), **bold** = best per dataset. Overall phase and step recognition both improve with hierarchy-awareness, though per-dataset performance varies. The multi-level score, which requires both levels to be strong, tends to benefit from hierarchy-awareness, with gains scaling with tree-likeness.

|                                           | Phase ( $\uparrow$ ) |             | Step ( $\uparrow$ ) |             | Multi-level ( $\uparrow$ ) |
|-------------------------------------------|----------------------|-------------|---------------------|-------------|----------------------------|
|                                           | Accuracy             | F1          | Accuracy            | F1          |                            |
| <i>MIPO</i>                               |                      |             |                     |             |                            |
| CLIP                                      | 16.2                 | 6.6         | 17.0                | 4.9         | 16.6                       |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 13.6                 | 7.9         | 8.3                 | 5.6         | 10.6                       |
| HYPER SURG                                | <b>20.9</b>          | <b>16.2</b> | <b>17.6</b>         | <b>8.9</b>  | <b>19.2</b>                |
| <i>StrasbyPass70</i>                      |                      |             |                     |             |                            |
| CLIP                                      | 18.5                 | 4.2         | 3.3                 | 0.7         | 7.8                        |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 29.5                 | 22.7        | 17.2                | 7.4         | 22.5                       |
| HYPER SURG                                | <b>30.2</b>          | <b>25.2</b> | <b>32.0</b>         | <b>12.0</b> | <b>31.1</b>                |
| <i>BernBypass70</i>                       |                      |             |                     |             |                            |
| CLIP                                      | 20.9                 | 4.2         | 1.8                 | 0.3         | 6.1                        |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | <b>23.9</b>          | <b>17.4</b> | 13.5                | 5.6         | 18.0                       |
| HYPER SURG                                | 17.4                 | 12.5        | <b>29.7</b>         | <b>8.5</b>  | <b>22.7</b>                |
| <i>GraSP</i>                              |                      |             |                     |             |                            |
| CLIP                                      | 4.3                  | 1.5         | 6.4                 | 1.2         | 5.2                        |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 25.1                 | <b>17.7</b> | <b>17.9</b>         | <b>9.0</b>  | <b>21.2</b>                |
| HYPER SURG                                | <b>26.6</b>          | 17.0        | 14.7                | 7.8         | 19.8                       |
| <i>Overall</i>                            |                      |             |                     |             |                            |
| CLIP                                      | 15.0                 | 4.1         | 7.1                 | 1.8         | 8.9                        |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 23.0                 | 16.4        | 14.2                | 6.9         | 18.1                       |
| HYPER SURG                                | <b>23.8</b>          | <b>17.7</b> | <b>23.5</b>         | <b>9.3</b>  | <b>23.2</b>                |

ble in flat Euclidean space, while steps require finer distinctions among classes that share a common parent phase, and this is where hierarchical structure in the embedding geometry makes the difference. With HYPERSURG, step accuracy rises to 23.5% while phase accuracy remains comparable (23.8%), narrowing the gap between the two levels. The multi-level score, which requires both levels to be strong, confirms that the benefit comes from making phase and step recognition work together rather than improving either in isolation (overall 23.2% vs 18.1%), though per-dataset gains depend on the degree of tree-structure in each annotation hierarchy. The gains track this spectrum closely: the bypass datasets (54–74% tree-likeness) show strong multi-level improvements, and MIPO (100% tree-likeness) sees the largest. The exception is GraSP, where SurgCLIP\*<sub>( $\beta$ )</sub> leads on step accuracy: its mixed-granularity step vocabulary means that 67% of steps recur across multiple phases (Sec. 3.2), breaking the parent–child containment that entailment cones enforce. The more tree-like the hierarchy, the larger the gains, exactly the expected behaviour for a geometric inductive bias designed to exploit tree structure.

Worth noting, MIPO is an orthopaedic procedure outside the surgical domains covered by the pretraining data [18],

Table 3. **Task 2: Cross-Level Consistency (%).** **Bold** = best per dataset. As models become more hierarchy-aware, catastrophic failures where both levels are wrong decrease, and a correct phase prediction more often leads to a correct step, showing that the two granularity levels become coupled rather than independent.

|                                           | Both Wrong ( $\downarrow$ ) | Both Right ( $\uparrow$ ) | Step   Phase ( $\uparrow$ ) |
|-------------------------------------------|-----------------------------|---------------------------|-----------------------------|
| <i>MIPO</i>                               |                             |                           |                             |
| CLIP                                      | 82.2                        | <b>15.4</b>               | <b>94.3</b>                 |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 82.1                        | 4.0                       | 27.2                        |
| HYPERSURG                                 | <b>70.9</b>                 | 9.4                       | 44.3                        |
| <i>StrasbyPass70</i>                      |                             |                           |                             |
| CLIP                                      | 80.4                        | 2.2                       | 10.9                        |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 57.9                        | 4.5                       | 15.2                        |
| HYPERSURG                                 | <b>44.8</b>                 | <b>7.0</b>                | <b>22.8</b>                 |
| <i>BernBypass70</i>                       |                             |                           |                             |
| CLIP                                      | 78.1                        | 0.8                       | 3.5                         |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 66.1                        | 3.5                       | 14.5                        |
| HYPERSURG                                 | <b>57.2</b>                 | <b>4.3</b>                | <b>24.2</b>                 |
| <i>GraSP</i>                              |                             |                           |                             |
| CLIP                                      | 89.9                        | 0.7                       | 20.9                        |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | <b>67.3</b>                 | <b>10.2</b>               | <b>35.3</b>                 |
| HYPERSURG                                 | 67.8                        | 9.1                       | 30.9                        |
| <i>Overall</i>                            |                             |                           |                             |
| CLIP                                      | 82.6                        | 4.8                       | <b>32.4</b>                 |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 68.3                        | 5.6                       | 23.0                        |
| HYPERSURG                                 | <b>60.2</b>                 | <b>7.4</b>                | 30.6                        |

yet HYPERSURG achieves the best accuracy and F1 at both levels, suggesting that the hierarchical inductive bias generalises to out-of-distribution procedures where domain-specific pretraining alone falls short.

## 5.2. Task 2: Cross-Level Consistency

We present results for Task 2 in Table 3, where we evaluate whether phase and step predictions are mutually coherent. The Both Wrong rates are high across the board, which shows how challenging zero-shot recognition at two granularity levels is. CLIP fails on the vast majority of frames (82.6% Both Wrong), and appears to lead on Both Right and Step|Phase on MIPO, but this is the same class-collapse artefact from Task 1. With a phase F1 of only 6.6, Step|Phase is computed on very few frames and the high score is misleading. These metrics should therefore be read jointly with Task 1, as they can overstate coherence when overall accuracy is low.

The consistent reduction in Both Wrong from CLIP (82.6%) to SurgCLIP\*<sub>( $\beta$ )</sub> (68.3%) to HYPERSURG (60.2%) shows that surgical pretraining helps and encoding hierarchical structure aids further, with the largest drop on StrasbyPass70 (from 57.9% for SurgCLIP\*<sub>( $\beta$ )</sub> to 44.8% for HYPERSURG). The exception is again GraSP, where SurgCLIP\*<sub>( $\beta$ )</sub> leads for the same structural reasons discussed

Table 4. **Task 3: Error severity (%)**. **Bold**=best per dataset. The hierarchy-aware model’s predictions more often fall within the correct phase and preserve the surgical action, indicating that its step predictions are more context-aware.

|                                           | Sibling ( $\uparrow$ ) | Cousin ( $\downarrow$ ) | Semantic Step ( $\uparrow$ ) |
|-------------------------------------------|------------------------|-------------------------|------------------------------|
| <i>MIPO</i>                               |                        |                         |                              |
| CLIP                                      | 20.6                   | 79.4                    | 20.4                         |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 21.9                   | 78.1                    | 9.1                          |
| HYPER SURG                                | <b>27.9</b>            | <b>72.1</b>             | <b>24.2</b>                  |
| <i>StrasbyPass70</i>                      |                        |                         |                              |
| CLIP                                      | 20.2                   | 79.8                    | 8.8                          |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 43.4                   | 56.6                    | 30.9                         |
| HYPER SURG                                | <b>51.3</b>            | <b>48.7</b>             | <b>44.6</b>                  |
| <i>BernBypass70</i>                       |                        |                         |                              |
| CLIP                                      | 21.3                   | 78.7                    | 9.9                          |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 36.8                   | 63.2                    | 20.6                         |
| HYPER SURG                                | <b>46.3</b>            | <b>53.7</b>             | <b>40.9</b>                  |
| <i>GraSP</i>                              |                        |                         |                              |
| CLIP                                      | 14.0                   | 86.0                    | 10.7                         |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | <b>31.6</b>            | <b>68.4</b>             | <b>29.1</b>                  |
| HYPER SURG                                | 29.2                   | 70.8                    | 24.4                         |
| <b>Overall</b>                            |                        |                         |                              |
| CLIP                                      | 19.0                   | 81.0                    | 12.4                         |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 33.4                   | 66.6                    | 22.4                         |
| HYPER SURG                                | <b>38.7</b>            | <b>61.3</b>             | <b>33.5</b>                  |

in Task 1. Step|Phase shows that when the phase is correct, the hierarchy-aware model picks the right step more often (30.6% vs 23.0% overall), suggesting that hierarchy-awareness couples the two levels through the embedding geometry.

### 5.3. Task 3: Hierarchy-Aware Error Severity

The previous tasks measured *whether* the model is correct; this task examines *where* predictions fall in the procedural hierarchy and whether they preserve the surgical action. We present results in Table 4. CLIP shows the lowest sibling accuracy across all datasets (19.0% overall), indicating that its predictions fall in unrelated phases almost uniformly. Surgical pretraining improves this substantially (33.4% sibling accuracy), meaning a larger share of step predictions fall within the correct phase, and hierarchy-awareness pushes it further (38.7%). Semantic step accuracy follows the same trend, rising from 12.4% (CLIP) to 22.4% (SurgCLIP\*<sub>( $\beta$ )</sub>) to 33.5% (HYPER SURG), meaning the model more often recognises the surgical action being performed even when it confuses the anatomical site. The top confusions illustrate what these numbers mean in practice. On BernBypass70, HYPER SURG’s dominant confusion for *Gastrojejunal Anastomosis Reinforcement* is *Gastrojejunal Defect Closure*, a sibling step within the same phase that involves suturing the same anatomical structure.

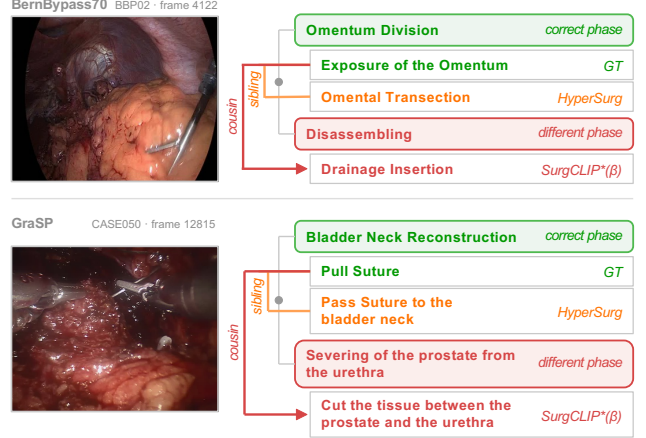


Figure 4. **Per-level metrics cannot distinguish between errors.** On both frames, each model predicts the correct phase but the wrong step, so accuracy and F1 record the same outcome. The errors differ in where they land in the hierarchy: the **orange** prediction is a sibling of the ground truth, a different step within the correct phase, while the **red** one is a cousin, from an unrelated phase. SURGHIBENCH measures this distinction.

By contrast, SurgCLIP\*<sub>( $\beta$ )</sub> confuses it with *Horizontal Stapling*, a step from a different phase (Gastric Pouch Creation) at a different anatomical site with a different surgical goal. Figure 4 shows the same pattern on GraSP and BernBypass70: the hierarchy-aware prediction remains within the correct phase, the Euclidean one does not. Flat accuracy treats both confusions identically, but these results provide insights that only hierarchy-aware evaluation can report and that directly address the needs of the surgical domain, where models with similar accuracy can produce errors of very different severity that should not be overlooked.

We provide qualitative examples of successful and failed predictions for both SurgCLIP\*<sub>( $\beta$ )</sub> and HYPER SURG, across all four datasets, in supplementary Figure 5.

## 6. Conclusion

In this work, we posit that phases and steps in surgical video should not be treated as independent recognition tasks. We introduce a hierarchy-aware evaluation suite that measures cross-level consistency and error severity alongside per-level accuracy, showing across four datasets and three procedure types that video-language models with comparable accuracy can result in very different error patterns. We propose HYPER SURG, a hierarchy-aware baseline that produces more accurate and consistent cross-level predictions by aligning the embedding geometry with the hierarchical structure of procedures. The resulting gains scale with the tree-likeness of each dataset’s annotation hierarchy, providing a principled indicator of when hierarchy-aware representations will help. As richer annotations become avail-

able, we expect both the evaluation framework and the geometric approach to extend naturally to deeper hierarchies, providing the community with tools to build surgical workflow systems that are not only accurate, but evaluated in ways that reflect what matters in the operating room.

## Acknowledgements

This work was supported by the University of Amsterdam’s Data Science Centre, as part of the HAVA Lab.

## References

- [1] Nicolas Ayobi, Santiago Rodriguez, Alejandra Perez, Isabela Hernandez, Nicolas Aparicio, Eugenie Dessevres, Sebastian Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolas Fernandez, and Pablo Arbelaez. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, page 103726, 2025. 3, 4, 11
- [2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, 2021. 4, 5
- [3] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12503–12512, 2020. 1, 2
- [4] Tobias Czempel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 343–352, 2020. 2
- [5] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning (ICML)*, pages 7694–7731, 2023. 2
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT)*, pages 4171–4186, 2019. 4, 5
- [7] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning (ICML)*, pages 1646–1655, 2018. 2, 5
- [8] Camille Graëff, Nicolas Padoy, Philippe Liverneaux, and Thomas Lampert. Introducing surgical workflow recognition in orthopaedic surgery with timestamp supervision. *Computers in Biology and Medicine*, 197:110995, 2025. 3, 4
- [9] Yaojun Hu, Kun Yuan, Nassir Navab, Haochao Ying, Jian Wu, and Nicolas Padoy. HyperVLP: Enhancing hierarchical surgical video-language pre-training in hyperbolic space. *arXiv preprint arXiv:2606.31245*, 2026. 3
- [10] Ruobing Jiang, Mengzhe Liu, Haobing Liu, and Yanwei Yu. Hierarchy-consistent learning and adaptive loss balancing for hierarchical multi-label classification. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2025. 3
- [11] Joel L. Lavanchy, Sanat Ramesh, Diego Dall’Alba, Cristians Gonzalez, Paolo Fiorini, Beat P. Müller-Stich, Philipp C. Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. Challenges in multi-centric generalization: Phase and step recognition in Roux-en-Y gastric bypass surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19:2249–2257, 2024. 2, 3, 4, 11
- [12] Teng Long, Pascal Mettes, Heng Tao Shen, and Cees G. M. Snoek. Searching for actions on the hyperbole. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1138–1147, 2020. 2, 3
- [13] Lena Maier-Hein, Swaroop S. Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Gianarou, Makoto Hashizume, Darko Katic, Hannes Kenngott, Michael Kranzfelder, Anand Malpani, Keno März, Thomas Neumuth, Nicolas Padoy, Carla Pugh, Nicolai Schoch, Danail Stoyanov, Russell Taylor, Martin Wagner, Gregory D. Hager, and Pierre Jannin. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017. 1
- [14] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132(9):3484–3508, 2024. 2
- [15] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6338–6347, 2017. 2
- [16] Avik Pal, Max van Spengler, Guido Maria D’Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. 2
- [17] Seulki Park, Youren Zhang, Stella X. Yu, Sara Beery, and Jonathan Huang. Visually consistent hierarchical image classification. In *International Conference on Learning Representations (ICLR)*, 2025. 3
- [18] Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. SurgLaVi: Large-scale hierarchical dataset for surgical vision-language representation learning. *Medical Image Analysis*, 110:103982, 2026. 1, 2, 4, 7, 11, 13
- [19] Zhi Qiao, Linbin Han, Xiantong Zhen, Jiahong Gao, and Zhen Qian. HYDEN: Hyperbolic density representations for medical images and reports. In *International Conference on Computational Linguistics (COLING)*, 2025. 2
- [20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Inter-*

- national Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. [2](#), [4](#)
- [21] Sanat Ramesh, Diego Dall’Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, and Nicolas Padoy. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *International Journal of Computer Assisted Radiology and Surgery*, 16:1111–1119, 2021. [3](#)
- [22] Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, Georgios Exarchakis, Alexandros Karargyris, and Nicolas Padoy. Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis*, 88:102844, 2023. [2](#)
- [23] Ana Manzano Rodriguez, Cees G. M. Snoek, and Marlies P. Schijven. Bridging the gap: Exposing the hidden challenges towards adoption of artificial intelligence in surgery. *British Journal of Surgery*, 112(11), 2025. [2](#)
- [24] Satwik Srivastava and Deepak Mishra. Severity of error in hierarchical datasets. *Scientific Reports*, 13:21903, 2023. [1](#), [2](#)
- [25] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017. [11](#)
- [26] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M. Lubotsky, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. *Medical Image Analysis*, 86:102770, 2023. [11](#)
- [27] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. 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 (MICCAI)*, pages 486–496, 2022. [11](#)
- [28] Lawrence W. Way, Lygia Stewart, Walter Gantert, Kingsway Liu, Crystine M. Lee, Karen Whang, and John G. Hunter. Causes and prevention of laparoscopic bile duct injuries: Analysis of 252 cases from a human factors and cognitive psychology perspective. *Annals of Surgery*, 237(4):460–469, 2003. [3](#)
- [29] Daiki Yoshikawa and Takashi Matsubara. PHyCLIP:  $\ell_1$ -product of hyperbolic factors unifies hierarchy and compositionality in vision-language representation learning. In *International Conference on Learning Representations (ICLR)*, 2026. [2](#)
- [30] Zhen Yu, Toan D. Nguyen, Lie Ju, Yaniv Gal, Maithili Sashindranath, Paul Bonnington, Lei Zhang, Victoria Mar, and Zongyuan Ge. Hierarchical skin lesion image classification with prototypical decision tree. *npj Digital Medicine*, 8: 26, 2025. [2](#)
- [31] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. HecVL: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024. [2](#)
- [32] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [2](#), [11](#)
- [33] Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L. Lavanchy, Jacques Marescaux, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, 105:103644, 2025. [1](#), [2](#)

# A Hierarchy-Aware Video-Language Model Evaluation and Hyperbolic Baseline for Surgery

## Supplementary Material

### A. SURGHIBENCH Evaluation Suite

#### A.1. Semantic Step Groupings

These groupings are used to compute the semantic step accuracy reported in Task 3 (Sec. 5.3). Two steps are “semantically related” if they perform the same type of action at different anatomical sites or are closely related surgical variants. The groupings were defined manually by inspecting each dataset’s step vocabulary and identifying steps that share the same surgical action. Table 5 lists all groupings per dataset.

Table 5. **Semantic step groupings** used for Task 3. Steps within each group share the same surgical action at different anatomical sites.

| Group                                          | Steps                                                                                     |
|------------------------------------------------|-------------------------------------------------------------------------------------------|
| <i>MIPO (8 steps)</i>                          |                                                                                           |
| Fixation                                       | Distal fixation, Proximal fixation                                                        |
| Plate handling                                 | Introduction of the plate, Modification of plate positioning                              |
| <i>StrasbyPass70 / BernBypass70 (46 steps)</i> |                                                                                           |
| Defect closure                                 | Gastrojejunal Defect Closure, Jejunojejunal Defect Closure                                |
| Reinforcement                                  | Gastrojejunal Anast. Reinforcement, Jejunojejunal Anast. Reinforcement                    |
| Stapling                                       | Horizontal, Vertical, Gastrojejunal, Jejunojejunal Stapling                               |
| Limb measurement                               | Biliary Limb Measurement, Alimentary Limb Measurement                                     |
| Limb opening                                   | Jejunum Opening, Biliary Limb Opening, Alimentary Limb Opening                            |
| Dissection                                     | Fat Pad, Lesser Curvature, Retrogastric Dissection                                        |
| Space closure/exposure                         | Petersen Space Exposure/Closure, Mesenteric Defect Exposure/Closure                       |
| <i>GraSP (21 steps)</i>                        |                                                                                           |
| Lymph node dissection                          | Dissection Iliac Lymph Nodes, Dissection Obturator Lymph Nodes                            |
| Anatomical dissection                          | Prevesical Dissection, Prostate Dissection, Seminal Vessicle Dissection, Denon Dissection |
| Packing                                        | Pack Lymph Nodes, Pack Prostate                                                           |
| Cutting                                        | Cut Prostate, Cut, Cut Bladder                                                            |
| Pass suture                                    | Pass Suture Urethra, Pass Suture Neck                                                     |

#### A.2. Zero-Shot Evaluation and MIPO Prompts

All zero-shot evaluations use text descriptions as prompts, with no prompt engineering, ensembling, or templates (e.g. “a photo of { }”). The model computes cosine similarity (or Lorentz distance for HYPERSURG) between the video embedding and each text prompt to produce per-class scores. For GraSP, BernBypass70, and StrasbyPass70, we use the text descriptions provided by SurgLaVi [18], not from the original dataset annotation files. BernBypass70 and StrasbyPass70 share the same annotation vocabulary and prompts; the dataset defines 12 phases [11], but following SurgLaVi [18] and existing zero-shot studies [32], we exclude the “other” class for fair comparison. MIPO is not part of SurgLaVi and its annotations do not include text descriptions, so we wrote prompts following the same narration style, describing the surgical actions visible in each phase and step. Tables 6 and 7 list the MIPO prompts.

#### A.3. Replication of SurgCLIP<sub>(β)</sub>

To validate our reimplementation of SurgCLIP<sub>(β)</sub><sup>\*</sup>, we reproduce the zero-shot evaluation protocol of Perez *et al.* [18] (their Table 2) on the six benchmarks they use for phase and step recognition: Cholec80 [25] and HeiChole [26] (cholecystectomy), AutoLaparo [27] (hysterec-tomy), StrasbyPass70 and BernBypass70 [11] (gastric by-pass), and GraSP [1] (prostatectomy). Cholec80, AutoLaparo, and HeiChole provide only phase annotations and therefore cannot be evaluated on the cross-level tasks of SURGHIBENCH, which is why they are reported in this supplementary rather than in the main paper. Step recognition is reported only for GraSP, the only dataset for which SurgLaVi reports step-level results. Table 8 includes the original SurgCLIP<sub>(β)</sub> results reported by Perez *et al.* alongside our reimplementation SurgCLIP<sub>(β)</sub><sup>\*</sup>, allowing readers to assess replication fidelity.

#### A.4. Additional Training Details

**Per-level training dynamics.** The contrastive loss converges at different rates per hierarchy level: phase loss (0.032) < step loss (0.050) < action loss (0.109) at the end of training. Phases (Level 3), whose rich captions map to few unique labels, are easiest to discriminate contrastively. Actions (Level 1), being brief action-level descriptions, are hardest. HYPERSURG preserves this natural difficulty gradient across hierarchy levels rather than collapsing it.

**Optimizer.** AdamW with lr=10<sup>−4</sup>, weight decay 0.02, and gradient clipping at max norm 1.0. LayerNorm parameters,

Table 6. **Zero-shot prompts: MIPO phases** (5 classes). The MIPO repository documentation reports 4 phases and 6 steps; the released annotation files contain 5 and 8 distinct labels, respectively. We use the annotation file counts throughout.

| Phase        | Prompt                                                                                                                         |
|--------------|--------------------------------------------------------------------------------------------------------------------------------|
| Approach     | I incise the skin over the flexor carpi radialis tendon and elevate the pronator quadratus to expose the distal radius.        |
| Closure      | I close the incision with intradermal sutures and apply a sterile compress bandage over the wound.                             |
| Fixation     | I place the volar locking plate and secure the distal radius fracture with K-wires, distal locking screws and proximal screws. |
| Installation | I prepare the volar locking plate with the aiming guide, apply a Velpeau bandage and inflate the pneumatic tourniquet.         |
| Verification | I perform fluoroscopic controls with anteroposterior and lateral views and check flexor tendon freedom and wrist mobility.     |

Table 7. **Zero-shot prompts: MIPO steps** (8 classes; see Tab. 6 caption).

| Step                              | Prompt                                                                                                                    |
|-----------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| Approach                          | I incise the skin over the flexor carpi radialis tendon and elevate the pronator quadratus muscle to expose the fracture. |
| Closure                           | I close the wound with intradermal absorbable sutures and apply a sterile compress dressing.                              |
| Distal Fixation                   | I place K-wires and insert distal locking screws near the wrist joint using a drill guide.                                |
| Installation                      | I prepare the volar locking plate with the aiming guide and inflate the pneumatic tourniquet.                             |
| Introduction of the Plate         | I introduce the volar locking plate along the distal radius volar cortex without screws.                                  |
| Modification of Plate Positioning | I remove and reposition the volar plate to optimize alignment with the volar cortex.                                      |
| Proximal Fixation                 | I insert cortical screws through the proximal plate holes into the intact diaphyseal shaft.                               |
| Verification                      | I perform fluoroscopic control with anteroposterior and lateral views to confirm reduction and screw placement.           |

bias terms, positional embeddings, and learnable scalar parameters (temperature, curvature, projection scales) are excluded from weight decay.

**Scheduler.** OneCycleLR with 5-epoch linear warmup (initial  $\text{lr} = 10^{-6}$ ), cosine annealing, and a final learning rate of  $10^{-6}$ .

**Batch Composition.** Each GPU processes 39 samples: 30 randomly sampled and 9 from 3 forced triplets (one caption per hierarchy level from the same video). Total effective batch size across 8 GPUs: 312.

**Numerical Stability.** All Lorentz distance computations (including  $\text{acosh}$ ) are performed in float32, even when the rest of the forward pass uses bfloat16 mixed precision. The argument to  $\text{acosh}$  is clamped to  $[1 + \epsilon, \infty)$  to prevent NaN gradients.

**Learnable Parameters.** Temperature  $\tau$  and curvature  $c$  are stored in log-space and clamped to  $[0.01, 1.0]$  and  $[0.1, 10.0]$  respectively. Projection scaling factors  $\alpha_v, \alpha_t$  are stored in log-space, initialized to  $\log(1/\sqrt{256})$ , and clamped to  $(-\infty, 0]$  so that the effective scales  $e^{\alpha_v}, e^{\alpha_t} \in (0, 1]$ . At convergence:  $\tau=0.035$ ,  $c=0.815$ ,  $e^{\alpha_v}=0.054$ ,  $e^{\alpha_t}=0.060$ .

## B. HYPERSURG Ablation

We ablate the three components of HYPERSURG — Lorentz distance, hierarchy-aware triplet batch sampler, and entailment cones — by training variants with different combinations. Table 9 compares four configurations: (i) Lorentz distance only (no batch sampler, no cones), (ii) Lorentz distance with entailment cones but random batches, (iii) Lorentz distance with the triplet batch sampler but no cones, and (iv) the full HYPERSURG with all three components.

Both components improve over the Lorentz-only baseline, but they play different roles. The triplet batch sampler is more reliable, improving both phase and step on every multi-level dataset. Where the full model pulls ahead is on the Bypass datasets, where cones and the sampler reinforce each other and produce gains of up to +6.7,pp on phase and +8.6,pp on step over the sampler alone. The exceptions follow the expected pattern. On GraSP, whose DAG-like hierarchy violates the tree assumption behind entailment cones, adding cones provides no further benefit at the phase level. On MIPO, the sampler alone edges out the full model on step accuracy, though cones still improve phase accuracy (+4.1,pp). Overall, the full model improves over the Lorentz-only baseline by +6.0,pp on phase and +6.9,pp on step, raising the multi-level score from 16.9% to 23.2%.

Table 8. **Reimplementation validation.** Zero-shot downstream performance (video-averaged, %) comparing the original SurgCLIP<sub>( $\beta$ )</sub> results from Perez *et al.* [18] (their Table 2) with our reimplementation SurgCLIP\*<sub>( $\beta$ )</sub>, trained on the same data with the same procedure.

|                                           | Phase ( $\uparrow$ ) |      | Step ( $\uparrow$ ) |      |
|-------------------------------------------|----------------------|------|---------------------|------|
|                                           | Accuracy             | F1   | Accuracy            | F1   |
| <i>Cholec80</i>                           |                      |      |                     |      |
| SurgCLIP <sub>(<math>\beta</math>)</sub>  | 58.0                 | 39.4 | —                   | —    |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 57.8                 | 34.5 | —                   | —    |
| <i>AutoLaparo</i>                         |                      |      |                     |      |
| SurgCLIP <sub>(<math>\beta</math>)</sub>  | 55.7                 | 46.0 | —                   | —    |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 58.0                 | 44.3 | —                   | —    |
| <i>HeiChole</i>                           |                      |      |                     |      |
| SurgCLIP <sub>(<math>\beta</math>)</sub>  | 57.0                 | 44.0 | —                   | —    |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 63.6                 | 35.5 | —                   | —    |
| <i>BernBypass70</i>                       |                      |      |                     |      |
| SurgCLIP <sub>(<math>\beta</math>)</sub>  | 18.3                 | 15.1 | —                   | —    |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 23.9                 | 17.4 | —                   | —    |
| <i>StrasbyPass70</i>                      |                      |      |                     |      |
| SurgCLIP <sub>(<math>\beta</math>)</sub>  | 31.2                 | 26.1 | —                   | —    |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 29.5                 | 22.7 | —                   | —    |
| <i>GraSP</i>                              |                      |      |                     |      |
| SurgCLIP <sub>(<math>\beta</math>)</sub>  | 34.8                 | 28.0 | 14.2                | 11.1 |
| SurgCLIP* <sub>(<math>\beta</math>)</sub> | 25.1                 | 17.7 | 17.9                | 9.0  |

Table 9. **Ablation of HYPERSURG components** (video-averaged accuracy, %). All variants use Lorentz distance. The full model achieves the best overall performance, with both components contributing complementary gains. **Bold** = best per dataset.

|                      | Phase ( $\uparrow$ ) | Step ( $\uparrow$ ) | Multi-level ( $\uparrow$ ) |
|----------------------|----------------------|---------------------|----------------------------|
| <i>MIPO</i>          |                      |                     |                            |
| Lorentz only         | 13.8                 | 11.9                | 12.8                       |
| + Entailment cones   | 9.3                  | 20.6                | 13.8                       |
| + Triplet sampler    | 16.8                 | <b>22.4</b>         | <b>19.4</b>                |
| + Both (HYPERSURG)   | <b>20.9</b>          | 17.6                | 19.2                       |
| <i>StrasbyPass70</i> |                      |                     |                            |
| Lorentz only         | 20.4                 | 24.1                | 22.2                       |
| + Entailment cones   | 24.3                 | 20.2                | 22.2                       |
| + Triplet sampler    | 23.5                 | 25.4                | 24.4                       |
| + Both (HYPERSURG)   | <b>30.2</b>          | <b>32.0</b>         | <b>31.1</b>                |
| <i>BernBypass70</i>  |                      |                     |                            |
| Lorentz only         | 12.6                 | 17.7                | 14.9                       |
| + Entailment cones   | 13.8                 | 17.4                | 15.5                       |
| + Triplet sampler    | 13.0                 | 21.1                | 16.6                       |
| + Both (HYPERSURG)   | <b>17.4</b>          | <b>29.7</b>         | <b>22.7</b>                |
| <i>GraSP</i>         |                      |                     |                            |
| Lorentz only         | 24.2                 | 12.9                | 17.7                       |
| + Entailment cones   | 26.7                 | <b>22.0</b>         | <b>24.2</b>                |
| + Triplet sampler    | <b>28.1</b>          | 13.8                | 19.7                       |
| + Both (HYPERSURG)   | 26.6                 | 14.7                | 19.8                       |
| <i>Overall</i>       |                      |                     |                            |
| Lorentz only         | 17.8                 | 16.7                | 16.9                       |
| + Entailment cones   | 18.5                 | 20.1                | 18.9                       |
| + Triplet sampler    | 20.4                 | 20.7                | 20.0                       |
| + Both (HYPERSURG)   | <b>23.8</b>          | <b>23.5</b>         | <b>23.2</b>                |

Table 10. **Sensitivity of HYPERSURG to the entailment loss weight  $\lambda$**  (video-averaged accuracy, %). All variants use the full configuration (Lorentz + cones + triplet sampler), only  $\lambda$  varies. We select  $\lambda=0.1$  as it yields the most balanced performance across both levels. **Bold** = best per dataset.

|                      | Phase ( $\uparrow$ ) | Step ( $\uparrow$ ) | Multi-level ( $\uparrow$ ) |
|----------------------|----------------------|---------------------|----------------------------|
| <i>MIPO</i>          |                      |                     |                            |
| $\lambda=0.01$       | 18.6                 | 17.4                | 18.0                       |
| $\lambda=0.05$       | <b>22.8</b>          | 20.0                | <b>21.4</b>                |
| $\lambda=0.1$        | 20.9                 | 17.6                | 19.2                       |
| $\lambda=0.2$        | 17.5                 | <b>23.6</b>         | 20.3                       |
| $\lambda=0.3$        | 13.7                 | 19.9                | 16.5                       |
| <i>StrasbyPass70</i> |                      |                     |                            |
| $\lambda=0.01$       | 22.1                 | 26.5                | 24.2                       |
| $\lambda=0.05$       | 27.2                 | 31.1                | 29.1                       |
| $\lambda=0.1$        | <b>30.2</b>          | 32.0                | <b>31.1</b>                |
| $\lambda=0.2$        | 26.1                 | 25.4                | 25.7                       |
| $\lambda=0.3$        | 28.8                 | <b>32.3</b>         | 30.5                       |
| <i>BernBypass70</i>  |                      |                     |                            |
| $\lambda=0.01$       | 10.2                 | 22.8                | 15.2                       |
| $\lambda=0.05$       | <b>18.2</b>          | 26.3                | 21.9                       |
| $\lambda=0.1$        | 17.4                 | <b>29.7</b>         | <b>22.7</b>                |
| $\lambda=0.2$        | 17.7                 | 23.4                | 20.4                       |
| $\lambda=0.3$        | 17.6                 | 28.6                | 22.4                       |
| <i>GraSP</i>         |                      |                     |                            |
| $\lambda=0.01$       | 31.1                 | <b>18.3</b>         | 23.9                       |
| $\lambda=0.05$       | 33.8                 | 10.0                | 18.4                       |
| $\lambda=0.1$        | 26.6                 | 14.7                | 19.8                       |
| $\lambda=0.2$        | <b>35.7</b>          | 17.7                | <b>25.1</b>                |
| $\lambda=0.3$        | 28.6                 | 16.1                | 21.5                       |
| <i>Overall</i>       |                      |                     |                            |
| $\lambda=0.01$       | 20.5                 | 21.3                | 20.3                       |
| $\lambda=0.05$       | <b>25.5</b>          | 21.9                | 22.7                       |
| $\lambda=0.1$        | 23.8                 | 23.5                | <b>23.2</b>                |
| $\lambda=0.2$        | 24.3                 | 22.5                | 22.9                       |
| $\lambda=0.3$        | 22.7                 | <b>24.2</b>         | 22.2                       |

tions from SurgCLIP<sub>( $\beta$ )</sub>\* and HYPERSURG across all four hierarchical datasets.

**Sensitivity to  $\lambda$ .** The entailment loss weight  $\lambda$  (Equation (9)) controls how strongly the model enforces parent-child containment via entailment cones. Table 10 reports phase and step accuracy of HYPERSURG across the four multi-level datasets for five values of  $\lambda$ , with all other hyperparameters held fixed.

No single value of  $\lambda$  dominates across all datasets. While a lower value ( $\lambda=0.05$ ) favor phase accuracy, a higher value ( $\lambda=0.3$ ) favor step accuracy,  $\lambda=0.1$  strikes the best balance between both levels, achieving the highest overall multi-level score (23.2%) and the best or near-best results on three of four datasets.

### C. Qualitative Results

To complement the quantitative results, Figure 5 presents qualitative examples of both successful and failed predic-

| Frame                                                                                                                                                                                                                                                              | Ground Truth                                                                                                                                          | SurgCLIP* <sub>(β)</sub>                                                                                                                                                                                                                                                                                                                                                   | HyperSurg                                                                                                                                                                                                                                                                                                                                                                                      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  <p><b>BernBypass</b><br/>(BBP02, frame 910)</p>                                                                                                                                  | <p>PHASE</p> <p><b>Gastric Pouch Creation</b></p> <p>STEP</p> <p><b>Lesser Curvature Dissection</b></p>                                               | <p>PHASE ✓ STEP <b>cousin</b> GT @ #8</p> <ol style="list-style-type: none"> <li>1. <b>Drainage Insertion</b> +0.696</li> <li>2. Gastric Opening +0.663</li> <li>3. Fat Pad Dissection +0.645</li> <li>4. Retrogastric Dissection +0.645</li> <li>5. NullStep +0.639</li> </ol>                                                                                            | <p>PHASE ✓ STEP <b>exact</b></p> <ol style="list-style-type: none"> <li>1. <b>Lesser Curvature Dissection</b> -0.685</li> <li>2. Retrogastric Dissection -0.687</li> <li>3. Gastric Opening -0.740</li> <li>4. Omental Transection -0.761</li> <li>5. Vertical Stapling -0.778</li> </ol>                                                                                                      |
|  <p><b>MIPO</b><br/>(video24, frame 21)</p>                                                                                                                                       | <p>PHASE</p> <p><b>Installation</b></p> <p>STEP</p> <p><b>Installation</b></p>                                                                        | <p>PHASE ✗ STEP <b>cousin</b> GT @ #4</p> <ol style="list-style-type: none"> <li>1. <b>Introduction of the Plate</b> +0.580</li> <li>2. Distal Fixation +0.545</li> <li>3. Approach +0.531</li> <li>4. <b>Installation</b> +0.494</li> <li>5. Proximal Fixation +0.425</li> </ol>                                                                                          | <p>PHASE ✗ STEP <b>exact</b></p> <ol style="list-style-type: none"> <li>1. <b>Installation</b> -0.765</li> <li>2. Closure -0.785</li> <li>3. Introduction of the Plate -0.795</li> <li>4. Distal Fixation -0.804</li> <li>5. Approach -0.812</li> </ol>                                                                                                                                        |
|  <p><b>StrasbyPass</b><br/>(SBP02, frame 90)</p>                                                                                                                                  | <p>PHASE</p> <p><b>Gastric Pouch Creation</b></p> <p>STEP</p> <p><b>Retrogastric Dissection</b></p>                                                   | <p>PHASE ✓ STEP <b>exact</b></p> <ol style="list-style-type: none"> <li>1. <b>Retrogastric Dissection</b> +0.683</li> <li>2. Omental Transection +0.670</li> <li>3. Drainage Insertion +0.669</li> <li>4. Fat Pad Dissection +0.666</li> <li>5. Lesser Curvature Dissection +0.666</li> </ol>                                                                              | <p>PHASE ✓ STEP <b>sibling</b> GT @ #2</p> <ol style="list-style-type: none"> <li>1. <b>Gastric Opening</b> -0.633</li> <li>2. <b>Retrogastric Dissection</b> -0.637</li> <li>3. Gastric Pouch Reinforcement -0.655</li> <li>4. Vertical Stapling -0.656</li> <li>5. Gastric Remnant Reinforcement -0.660</li> </ol>                                                                           |
|  <p><b>GraSP</b><br/>(CASE041, frame 60)</p>                                                                                                                                    | <p>PHASE</p> <p><b>Right pelvic isolated lymphadenectomy</b></p> <p>STEP</p> <p><b>Identification and dissection of the iliac vein and artery</b></p> | <p>PHASE ✗ STEP <b>sibling</b> GT @ #2</p> <ol style="list-style-type: none"> <li>1. <b>Cutting and dissection of the external iliac vein's lymph node</b> +0.599</li> <li>2. <b>Identification and dissection of the iliac vein and artery</b> +0.578</li> <li>3. Idle +0.509</li> <li>4. Cut suture or tissue +0.501</li> <li>5. Prevesical dissection +0.500</li> </ol> | <p>PHASE ✗ STEP <b>cousin</b> GT @ #4</p> <ol style="list-style-type: none"> <li>1. <b>Idle</b> -0.842</li> <li>2. Cutting and dissection of the external iliac vein's lymph node -0.863</li> <li>3. Prevesical dissection -0.889</li> <li>4. <b>Identification and dissection of the iliac vein and artery</b> -0.892</li> <li>5. Cut between the prostate and bladder neck -0.904</li> </ol> |
|  <p><b>BernBypass</b><br/>(BBP29, frame 720)</p>                                                                                                                                | <p>PHASE</p> <p><b>Gastrojejun Anastomosis</b></p> <p>STEP</p> <p><b>Gastrojejun Stapling</b></p>                                                     | <p>PHASE ✗ STEP <b>exact</b></p> <ol style="list-style-type: none"> <li>1. <b>Gastrojejun Stapling</b> +0.332</li> <li>2. Jejunojejunal Stapling +0.326</li> <li>3. Suture of Small Bowel Lesion +0.325</li> <li>4. Horizontal Stapling +0.323</li> <li>5. Fat Pad Dissection +0.323</li> </ol>                                                                            | <p>PHASE ✗ STEP <b>cousin</b> GT @ #8</p> <ol style="list-style-type: none"> <li>1. <b>Jejunojejunal Stapling</b> -0.585</li> <li>2. Jejunal Transection -0.606</li> <li>3. Biliary Limb Opening -0.656</li> <li>4. Alimentary Limb Opening -0.668</li> <li>5. Biliary Limb Measurement -0.696</li> </ol>                                                                                      |
| <div> <div>Correct (rank #1)</div> <div>Ground truth (not #1)</div> <div>Wrong prediction (#1)</div> </div> <div> <div>PHASE ✓ / ✗ — phase verdict</div> <div>STEP exact / sibling / cousin — step severity</div> <div>GT @ #n — rank of ground truth</div> </div> |                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                                                                                                                                                                                                                                                                                                |

Figure 5. **Where does each model get it right, and when wrong, how bad is the error?** Each row shows a test frame with ground-truth labels and top-5 step predictions. *Top two rows:* HYPERSURG predicts the exact step while SurgCLIP\*<sub>(β)</sub> predicts a cousin from a different phase. *Row 3:* HYPERSURG predicts the wrong step but stays within the correct phase (sibling, 0.004 margin), while SurgCLIP\*<sub>(β)</sub> gets the exact step. Flat metrics treat both errors identically. *Row 4:* HYPERSURG predicts Idle, a cross-phase error on GraSP, consistent with its DAG-like hierarchy. *Row 5 (subtree lock-in):* HYPERSURG commits to the wrong phase subtree, with all five predictions from Jejunojejunal Anastomosis instead of the ground-truth Gastrojejun Anastomosis, while SurgCLIP\*<sub>(β)</sub> ranks the correct step first. Both the strengths and this failure mode are consistent with the same geometric property: hierarchical containment helps when the correct subtree is identified, but can trap the model when sibling phases are visually similar.