SURGHICO: HIERARCHICAL COMPOSITION OF LONG-HORIZON SURGICAL VIDEO–LANGUAGE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Surgical video–language pretraining uses narrated videos and their hierarchical captions, from individual maneuvers to stretches of a procedure. However, clip-based baselines encode segments with a fixed frame budget or aggregate local features by mean pooling, providing limited modeling of how local events form longer sequences. Under matched controls, neither incorporating hierarchical captions nor enlarging the temporal receptive field substantially improves the baseline, suggesting that the bottleneck lies in how longer-range representations are formed rather than in supervision or input. Motivated by this finding, we propose SurgHiCo, a hierarchical video–language pretraining framework that organizes representation learning by temporal scale, following the structure of surgery itself: maneuvers make up steps, and steps make up phases. Each level is composed from the level below by a lightweight composer and aligned with text at its own span. SurgHiCo achieves state-of-the-art zero-shot surgical phase recognition across five datasets and the strongest linear-probing performance at every labeled-video fraction on all of them. Notably, it improves average zero-shot accuracy by 6.4 points over the strongest prior model and full-shot linear-probing macro-F1 by 12.2 points.

1 INTRODUCTION

Surgical phase recognition is, at its core, a problem of long-horizon workflow modeling (Liu et al., 2026). A procedure unfolds as an ordered sequence of phases, and adjacent phases often share the same instruments, anatomy, and camera views, so a short window of frames is frequently ambiguous about which phase it belongs to. Distinguishing them requires reasoning over a longer temporal context: what has been done before, and how the current activity fits into the procedure as a whole (Twinanda et al., 2017; Czempiel et al., 2020). Because frame-level phase annotation also demands expert time, it is attractive to learn such representations without task-specific labels.

Surgical vision–language pretraining (VLP) provides such supervision from narrated videos. In surgical lectures, surgeons describe operative actions, instruments, anatomy, and procedural decisions as the video unfolds. Recent datasets organize these videos into video–text pairs at multiple nested temporal granularities (Yuan et al., 2025; 2024a;b; Perez et al., 2026). Existing methods, however, flatten this hierarchy during representation learning: pairs from every granularity supervise the same shared clip encoder, which sees a fixed number of frames regardless of segment length. A longer segment is represented either by that encoder or by averaging the features of the clips it contains. The hierarchy is present in the data, but not in how the visual representations are built.

This creates an asymmetry between the two modalities. For a long segment, the text encoder reads a caption describing a sequence of steps, whereas the clip encoder represents the corresponding minutes of video from only sixteen sampled frames. The long-span caption therefore supervises a visual representation with only sparse access to the temporal process it describes. If so, longer captions should add little to a flat model, and under matched controls they do. Holding the encoder architecture, the evaluation window, and the pretraining videos fixed, adding intermediate- and coarse-level captions to a flat model raises five-dataset average zero-shot phase accuracy from 38.1 to 39.5. Widening the visual input sixteen-fold at test time and averaging clip features over the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

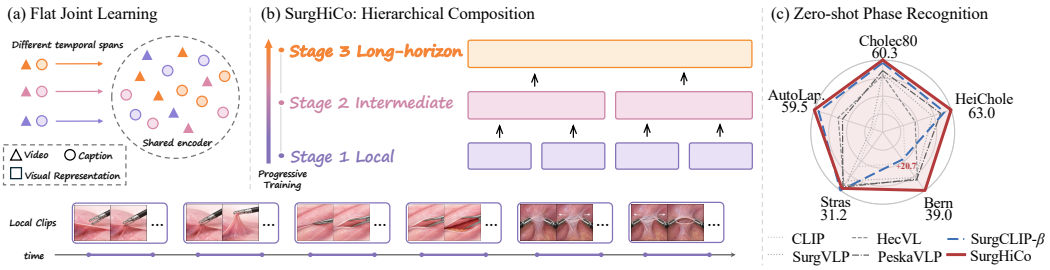


Figure 1: Flat versus hierarchical surgical video–language pretraining. (a) Flat joint learning encodes video segments of every temporal span with one shared clip encoder, without modeling how shorter segments compose into longer ones. (b) SurgHiCo composes local visual representations into intermediate and then long-horizon representations over three training stages, aligning each level with text of its own span. (c) Zero-shot phase recognition accuracy on five datasets, with the best value on each axis marked; prior results are quoted from SurgLaVi (Perez et al., 2026) (see Table 1).

longer window yields 40.0. Neither more text nor more frames moves the result by more than two points.

We therefore propose SurgHiCo, which treats temporal scales as successive levels of representation learning rather than parallel sources of supervision for one clip encoder. At each level, the representation of a longer segment is composed from the already learned representations of its constituent shorter segments and aligned with text at the corresponding span. Applying this principle successively yields local, intermediate, and long-horizon representations that follow the way surgical workflow is described, from individual maneuvers to the steps and phases they make up (Lalys & Jannin, 2014; Meireles et al., 2021) (Figure 1). In practice the clip encoder is trained once and then frozen, and each higher level is a lightweight composer over the representations of the level below (Section 3).

We evaluate on five surgical phase recognition datasets. With the same pretraining videos and backbone as SurgCLIP-β, the composed long-horizon representation raises average zero-shot accuracy from 38.1 at the clip level to 50.4, 6.4 points above SurgCLIP-β and 9.9 points above a flat model that receives the same captions and input but averages its clip features. Concatenating the three levels (SurgHiCo-Fusion) gives the highest macro-F1 on all five datasets at every labeled-data fraction in video few- and full-shot linear probing, exceeding SurgCLIP-β on average by 8.57, 6.75, and 12.15 points with 10%, 50%, and 100% of the labeled videos, and improving over each level alone in full-shot ablations.

2 Related Work

Surgical workflow understanding. Surgical video understanding covers temporally coarse workflow recognition and fine-grained scene interpretation. EndoNet jointly modeled surgical phases and tool presence from laparoscopic video (Twinanda et al., 2017), while TeCNO used dilated temporal convolutions to refine phase predictions over a broad receptive field (Czempiel et al., 2020). Subsequent benchmarks have expanded the range of procedures and annotation levels: AutoLaparo and multi-center workflow datasets evaluate phase and step recognition across procedures and hospitals (Wang et al., 2022; Maier-Hein et al., 2023; Wagner et al., 2023); GraSP provides phase, step, action, and instrument labels (Ayobi et al., 2025); and CholecT50 supports instrument–verb–target triplet recognition (Nwoye et al., 2022). These supervised formulations establish the temporal diversity of surgical tasks, but typically learn from task-specific labels. SurgHiCo instead learns representations through language supervision and transfers them to phase recognition across procedures and clinical centers, either zero-shot or by linear probing with frozen visual representations.

Vision–language learning for surgery. Contrastive vision–language pretraining enables open-vocabulary transfer from paired visual and textual data (Radford et al., 2021; Miech et al., 2019). Surgical VLP adapts this paradigm to domain-specific procedural semantics. SurgVLP aligns short

clips with automatically transcribed narrations from surgical lectures (Yuan et al., 2025). HecVL extends these pairs with phase-level concepts and video-level abstracts, learning separate contrastive spaces while sharing the underlying visual and textual encoders (Yuan et al., 2024a). PeskaVLP further enriches the supervision with procedure-aware knowledge (Yuan et al., 2024b). More recently, SurgLaVi scales hierarchical surgical video–text data to fine-, mid-, and coarse-level segments; its SurgCLIP baseline mixes these levels when training a shared video–text model (Perez et al., 2026). These methods build long-span representations in one of two ways. SurgCLIP encodes a segment of any span with the same clip encoder from a fixed number of frames, while HecVL aggregates clip-level features over the longer span by parameter-free average pooling. In both cases the long-span objective ultimately updates the clip encoder, and no parameters are learned for how a longer representation is formed. SurgHiCo instead directs each span of text to a level whose input is a sequence of the next-lower level's representations, and leaves the clip encoder untouched after clip-level training.

Hierarchical and long-form video–language learning. Hierarchical supervision has also been explored outside surgery. HierVL pairs clip-level narrations with a whole-video summary and aggregates clip features, by average pooling or self-attention, for its video-level objective (Ashutosh et al., 2023). HiTeA constructs short and long temporal views and learns both fine-grained moment alignment and cross-view temporal relations (Ye et al., 2023). For long-form video, LF-VILA combines clip–sentence temporal alignment with hierarchical temporal-window attention to capture long-range dependencies efficiently (Sun et al., 2022). In each case the visual encoder is trained under the longer-span objective as well. SurgHiCo freezes each level before the next is trained, and aligns every level with text covering exactly its own span rather than with a single global summary.

3 METHOD

3.1 OVERVIEW

SurgHiCo follows one rule at every level above the clip: the representation of a segment is a function of the ordered representations of the shorter segments it contains, and it is aligned only with text at its own span. Applied twice, this rule yields the visual hierarchy

$$\boldsymbol{z}^{\mathrm{loc}} \longrightarrow \boldsymbol{z}^{\mathrm{int}} \longrightarrow \boldsymbol{z}^{\mathrm{long}}. \tag{1}$$

Stage 1 trains the Local Encoder and the text encoder through contrastive learning on short clips. The Local Encoder is then frozen before Stage 2, whose Intermediate Composer operates on cached local features. After Stage 2, the learned intermediate modules are frozen and the Long-horizon Composer operates on cached intermediate features. Composition realizes the rule above; freezing adds a second guarantee: supervision from a longer span updates only the modules introduced for that span and the text-side adapters, never the clip encoder. The local and intermediate levels are trained on the Fine and Mid pairs of SurgLaVi-β, respectively. Native Coarse annotations are not used; the long-horizon level is constructed from consecutive Mid segments (Section 3.3). Figure 2 details the three stages.

3.2 HIERARCHICAL VIDEO–LANGUAGE ORGANIZATION

Let a surgical video be $V = \{f_t\}_{t=1}^{T}$. Its narration is organized into segment–caption pairs

$$\mathcal{D}^s = \{(v_i^s, c_i^s)\}_{i=1}^{N_s}, \quad s \in \{\mathrm{loc}, \mathrm{int}\}. \tag{2}$$

The hierarchy is defined by temporal containment. The children of intermediate unit j are

$$\mathcal{C}(j) = \{i \mid v_i^{\mathrm{loc}} \subseteq v_j^{\mathrm{int}}\}. \tag{3}$$

Children are sorted by start time. When more than $K_{\max}$ local units fall inside one intermediate unit, we uniformly subsample them over time, retaining coverage of the complete interval. We set $K_{\max} = 6$.

3.3 LONG-HORIZON LEVEL RECONSTRUCTION

We retain the captions and temporal segmentation provided by SurgLaVi-β at the lower levels, but reconstruct its longest level. The native 15,493 Coarse samples contain only 2.04 Mid children on

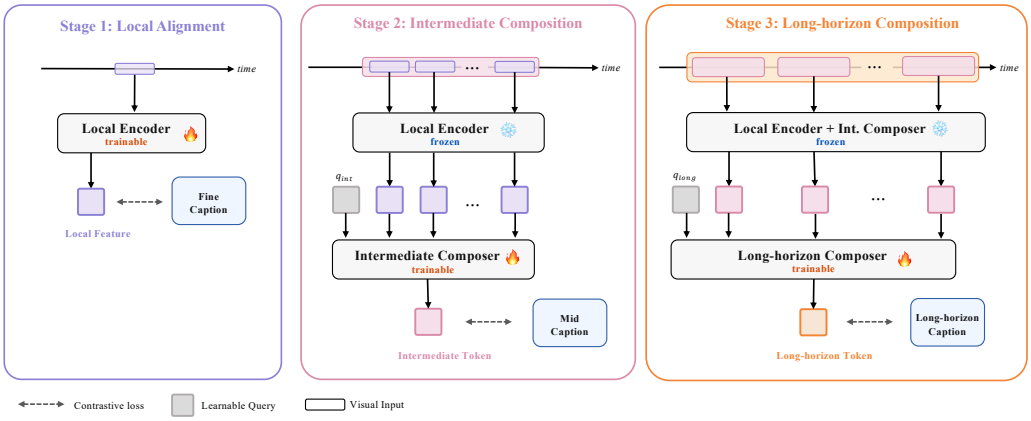


Figure 2: Staged training of SurgHiCo. Stage 1 trains the Local Encoder on Fine captions; it is then frozen. Stage 2 composes the ordered local features of a Mid segment into an Intermediate Token aligned with its Mid caption. Stage 3 freezes all preceding visual modules and composes consecutive Intermediate Tokens into a Long-horizon Token aligned with a synthetic long-horizon caption. In Stages 2 and 3, only the new composer (with its projection in Stage 2) and the text-side LoRA adapters are updated; Stage 2 additionally retains an auxiliary local contrastive loss.

average: 30.17% contain one Mid and 48.29% contain two, totaling 78.46%. Consequently, most Coarse samples provide little temporal expansion beyond a single Mid segment.

We instead build long-horizon windows from consecutive Mid segments. Within each video, adjacent Mid segments separated by at most 45 seconds form a run, and a window of up to four consecutive segments slides over each run with stride one, so adjacent windows share up to three segments. Tail windows with two or three segments are retained, and windows spanning more than 320 seconds are discarded. We denote the ordered set of Mid segments in long-horizon window k by $\mathcal{M}(k)$, the long-horizon analogue of $\mathcal{C}(j)$. For language supervision, we use Gemini 3.0 Pro Thinking to summarize each Mid caption into a concise one-sentence description and concatenate the resulting summaries in temporal order; windows without any summary are discarded. This reconstruction increases the average number of Mid children from 2.04 to 3.67 and the average temporal span from 91.34 to 147.93 seconds, yielding 24,299 candidate windows. Because adjacent windows overlap, training draws at most three temporally spread windows per video in each epoch.

3.4 Local Alignment

For a local pair $(v_i^{\mathrm{loc}}, c_i^{\mathrm{loc}})$, the Local Encoder E_v and the text encoder E_t produce

$$\begin{aligned} \boldsymbol{h}_i^{\mathrm{loc}} &= \mathrm{Pool}(E_v(v_i^{\mathrm{loc}})) \in \mathbb{R}^{d_v}, \\ \boldsymbol{h}_i^{t,\mathrm{loc}} &= E_t(c_i^{\mathrm{loc}})_{\mathrm{CLS}} \in \mathbb{R}^{d_t}. \end{aligned} \tag{4}$$

Following SurgCLIP (Perez et al., 2026), Pool drops the visual class token and averages the patch tokens over space and then time, and CLS takes the final hidden state of the text classification token, with $d_v = d_t = 768$. Linear projection heads map both features to a shared 256-dimensional space, where Norm denotes ℓ_2 normalization:

$$\begin{aligned} \boldsymbol{z}_i^{\mathrm{loc}} &= \mathrm{Norm}(P_v^{\mathrm{loc}} \boldsymbol{h}_i^{\mathrm{loc}}), \\ \boldsymbol{z}_i^{t,\mathrm{loc}} &= \mathrm{Norm}(P_t \boldsymbol{h}_i^{t,\mathrm{loc}}). \end{aligned} \tag{5}$$

Stage 1 optimizes all Transformer blocks in E_v and E_t, both projection heads, and a learnable temperature initialized to 0.07 using the symmetric contrastive objective defined below. After Stage 1, the visual encoder and its projection are frozen. We cache the projection-input vector $\boldsymbol{h}_i^{\mathrm{loc}}$, rather than the complete token sequence, so hierarchical training does not decode video again.

3.5 INTERMEDIATE COMPOSITION

For intermediate unit j, we collect its ordered child features:

$$\boldsymbol{X}_j^{\text{int}} = [\boldsymbol{h}_{i_1}^{\text{loc}}, \ldots, \boldsymbol{h}_{i_{K_j}}^{\text{loc}}], \quad i_r \in \mathcal{C}(j). \tag{6}$$

Variable-length sequences are padded and paired with a validity mask. We prepend a learned token $\boldsymbol{q}_{\text{int}}$ and add learned positional embeddings. A lightweight two-layer pre-norm Transformer A_{int}, the Intermediate Composer, produces

$$\widetilde{\boldsymbol{H}}_j^{\text{int}} = A_{\text{int}}([\boldsymbol{q}_{\text{int}}; \boldsymbol{X}_j^{\text{int}}], \boldsymbol{m}_j^{\text{int}}). \tag{7}$$

The Intermediate Token and embedding are

$$\boldsymbol{h}_j^{\text{int}} = \widetilde{\boldsymbol{H}}_j^{\text{int}}[0], \qquad \boldsymbol{z}_j^{\text{int}} = \text{Norm}(P_v^{\text{int}} \boldsymbol{h}_j^{\text{int}}). \tag{8}$$

To adapt the text side to longer captions without moving its output space, we freeze the base text encoder and shared text projection, and insert rank-4 LoRA adapters (α=8, dropout 0.1) into the query and value projections of its final two layers. The embedding of the paired Mid caption is

$$\boldsymbol{z}_j^{t,\text{int}} = \text{Norm}(P_t E_t^{\text{LoRA}}(c_j^{\text{int}})). \tag{9}$$

The same adapted text path encodes the Fine captions of the child local units in each batch. Stage 2 retains local alignment with

$$\mathcal{L}_2 = \mathcal{L}_{\text{int}} + \lambda_{\text{loc}} \mathcal{L}_{\text{loc}}, \tag{10}$$

where $\lambda_{\text{loc}} = 0.3$. $\mathcal{L}_{\text{loc}}$ is the same symmetric contrastive loss applied to the valid child local features and their original Fine captions. The local visual features are frozen, so this auxiliary objective updates only the text-side adapters.

3.6 LONG-HORIZON COMPOSITION

After Stage 2, A_{int} and P_v^{int} are frozen. For long-horizon window k, we gather projection-input Intermediate Tokens

$$\boldsymbol{X}_k^{\text{long}} = [\boldsymbol{h}_{j_1}^{\text{int}}, \ldots, \boldsymbol{h}_{j_{M_k}}^{\text{int}}], \quad j_r \in \mathcal{M}(k). \tag{11}$$

A separate two-layer Transformer A_{long}, the Long-horizon Composer, receives a learned Long-horizon Token:

$$\widetilde{\boldsymbol{H}}_k^{\text{long}} = A_{\text{long}}([\boldsymbol{q}_{\text{long}}; \boldsymbol{X}_k^{\text{long}}], \boldsymbol{m}_k^{\text{long}}). \tag{12}$$

We use its first output directly as the long-horizon representation:

$$\boldsymbol{h}_k^{\text{long}} = \widetilde{\boldsymbol{H}}_k^{\text{long}}[0]. \tag{13}$$

Rather than learning a new long-horizon projection, we apply the frozen intermediate projection:

$$\boldsymbol{z}_k^{\text{long}} = \text{Norm}(P_v^{\text{int}} \boldsymbol{h}_k^{\text{long}}). \tag{14}$$

This keeps the long-horizon representation in the intermediate visual space. Its paired text embedding uses the synthetic long-horizon caption c_k^{long} and the same text encoder, adapters, and frozen projection:

$$\boldsymbol{z}_k^{t,\text{long}} = \text{Norm}(P_t E_t^{\text{LoRA}}(c_k^{\text{long}})). \tag{15}$$

3.7 LANGUAGE-GUIDED OPTIMIZATION

At scale $s \in \{\text{loc}, \text{int}, \text{long}\}$, the batch similarity is

$$S_{ij}^s = \frac{(\boldsymbol{z}_i^s)^\top \boldsymbol{z}_j^{t,s}}{\tau}. \tag{16}$$

We use symmetric video-to-text and text-to-video contrast:

$$\mathcal{L}_s = \frac{1}{2} \left(\mathcal{L}_{v \to t}^s + \mathcal{L}_{t \to v}^s \right), \tag{17}$$

$$\mathcal{L}^s_{v\to t} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(S^s_{ii})}{\sum_{j=1}^{B}\exp(S^s_{ij})}, \tag{18}$$

with the reverse direction defined analogously. Stage 1 optimizes both encoders, both projection heads, and τ with $\mathcal{L}_{\mathrm{loc}}$; embeddings are gathered across GPUs before forming negatives. Stage 2 freezes the Stage-1 modules and τ, and optimizes the Intermediate Composer, intermediate projection, and text adapters using Equation (10). Its intermediate objective uses cross-GPU negatives, whereas the auxiliary local objective uses valid child pairs on each GPU. Stage 3 freezes all earlier visual modules and τ, and optimizes the Long-horizon Composer and text adapters with $\mathcal{L}_{\mathrm{long}}$ using in-batch negatives.

3.8 Multi-Scale Zero-Shot Interface

SurgHiCo exposes

$$\mathcal{Z} = \{\boldsymbol{z}^{\mathrm{loc}}, \boldsymbol{z}^{\mathrm{int}}, \boldsymbol{z}^{\mathrm{long}}\}. \tag{19}$$

Downstream videos carry no segment annotations, so the hierarchy is instantiated with fixed windows centered at each time point: four 16-second clips form an intermediate unit, and four intermediate units form a long-horizon window (Appendix B). For a downstream label set, we embed class prompts with the shared text tower and independently compare one selected visual level by cosine similarity. Each level is evaluated with the text path from the end of its own training stage: the Stage-1 text encoder for the local level, and the Stage-2 and Stage-3 adapters for the intermediate and long-horizon levels, respectively. For supervised downstream use, the projected embeddings can also be concatenated and passed to a linear classifier.

4 Experiments

4.1 Experimental Setup

Pretraining data and hierarchy. We use the captions and temporal segmentation of SurgLaVi-β (Perez et al., 2026), treating its Fine and Mid pairs as the local and intermediate levels and reconstructing the long-horizon level as described in Section 3.

The final local-level set contains 66,083 samples from 2,433 videos, with an average duration of 18.48 seconds and caption length of 24.60 words. The intermediate-level set contains 29,668 samples from 2,422 videos, averaging 46.08 seconds, 39.81 caption words, and 2.42 local children per intermediate segment. The long-horizon set contains 24,299 candidate windows from 2,371 videos; candidates average 3.67 intermediate children, span 147.93 seconds, and contain 60.14 caption words.

Architectures and optimization. We use TimeSformer-B (Bertasius et al., 2021) as the visual backbone and BERT-base as the text encoder. Each local clip contains one randomly sampled frame from each of 16 temporal bins at 224×224 resolution. Stage 1 trains both encoders on local pairs, after which the Local Encoder is fixed. Stage 2 trains the Intermediate Composer, intermediate projection, and text LoRA adapters; Stage 3 trains the Long-horizon Composer and the same adapters, with all preceding visual modules frozen. Both composers are two-layer pre-norm Transformers with hidden dimension 768 and eight heads. Optimization settings for all stages are listed in Appendix D.

Downstream tasks and metrics. We evaluate phase recognition on Cholec80 (Twinanda et al., 2017), HeiChole (Maier-Hein et al., 2023), BernBypass and StrasBypass (Wagner et al., 2023), and AutoLaparo (Wang et al., 2022), reporting video-averaged accuracy: accuracy is computed within each test video and averaged over videos, following SurgLaVi. Linear probing additionally reports macro-F1 (Appendix C). All zero-shot tables use the official evaluation sets. At each annotated time point, we use fixed, centered (offline) evaluation contexts to standardize coverage across datasets: the local level uses a centered 16-frame window, the intermediate level uses four local children spanning 64 seconds, and the long-horizon level uses four intermediate units spanning 256 seconds. Training retains variable child counts as described in Section 3.

Table 1: Zero-shot phase recognition (video-averaged accuracy, %). † denotes results quoted from SurgLaVi (Perez et al., 2026). SurgHiCo uses its long-horizon representation from one fixed model across all datasets, without dataset-specific selection.

Method	Dataset	Backbone	Cholec80	HeiChole	Bern	Stras	AutoLap.	Avg.
CLIP† (2021)	CLIP400M	ViT-B/16	27.81	19.21	20.73	18.52	8.02	18.86
MedSigLIP† (2023)	Medical V–L	SigLIP-400M	40.94	45.36	30.73	20.15	26.54	32.74
SurgVLP† (2025)	SVL	ResNet-50	49.02	45.70	30.46	26.08	10.02	32.26
HecVL† (2024a)	SVL	ResNet-50	47.50	29.91	32.09	28.59	38.72	35.36
PeskaVLP† (2024b)	SVL	ResNet-50	51.43	53.33	31.48	29.42	34.94	40.12
SurgCLIP-β† (2026)	SurgLaVi-β	TimeSformer-B	57.98	56.95	18.30	**31.24**	55.72	44.04
SurgHiCo	SurgLaVi-β	TimeSformer-B	**60.28**	**63.04**	**39.02**	30.14	**59.55**	**50.41**

Table 2: Controlled zero-shot phase-recognition ablations (video-averaged accuracy, %). F/M/C denote Fine, Mid, and Coarse captions; Syn-W denotes synthetic long-horizon captions. Local evaluates one 16-frame clip, whereas Long uses the same 16 local clips (256 sampled frames) for all long-context variants. Mean denotes parameter-free averaging.

Variant	PT captions	Eval. context	Agg.	Cholec80	HeiChole	Bern	Stras	AutoLap.	Avg.
Flat supervision									
Fine-only	F	Local	None	53.71	56.82	17.98	20.60	41.34	38.09
Flat	F/M/C	Local	None	53.96	58.07	19.21	16.58	49.52	39.47
Flat	F/M/Syn-W	Local	None	54.10	59.23	19.37	16.95	50.04	39.94
Long-context modeling									
Flat + Mean	F/M/C	Long	Mean	54.27	55.71	21.42	18.31	50.37	40.02
Flat + Mean	F/M/Syn-W	Long	Mean	54.25	57.18	20.64	19.30	51.00	40.47
SurgHiCo	F/M/Syn-W	Long	Hier.	60.28	63.04	39.02	30.14	59.55	50.41

Table 3: Stage-wise zero-shot phase recognition (video-averaged accuracy, %). Local, Intermediate, and Long-horizon denote the outputs of Stages 1, 2, and 3, respectively. Each row uses the fixed evaluation context assigned to that level; the long-horizon representation is used for the main zero-shot comparison.

Representation	Cholec80	HeiChole	BernBypass	StrasBypass	AutoLaparo	Avg.
Local	53.71	56.82	17.98	20.60	41.34	38.09
Intermediate	56.46	55.54	30.32	23.47	34.32	40.02
Long-horizon	**60.28**	**63.04**	**39.02**	**30.14**	**59.55**	**50.41**

4.2 Comparison with Prior Work

The main zero-shot comparison focuses on phase recognition, for which broader procedural context is most relevant. Results in Table 1 marked with † are quoted from prior work and provide broader context across different pretraining data and backbones under the SurgLaVi protocol, which encodes a centered 16-frame window; the remaining row uses our evaluation protocol. SurgHiCo uses one fixed trained model across all datasets, without dataset-specific model selection. Across the reported results, SurgHiCo achieves the highest five-dataset average accuracy, exceeding SurgCLIP-β by 6.37 points. It also leads on four of the five datasets; on StrasBypass, SurgCLIP-β remains 1.10 points higher. Because the quoted rows use a shorter evaluation context, this table serves as an external reference point; the controlled comparisons in Section 4.3, which hold the pretraining videos, architecture, and evaluation input fixed and in which SurgHiCo exceeds a caption- and input-matched flat model by 9.94 points, carry the causal argument.

4.3 Ablation Studies

Tables 3 and 2 address three questions: what each level adds, what longer captions add to a flat model, and what wider coverage adds without learned composition.

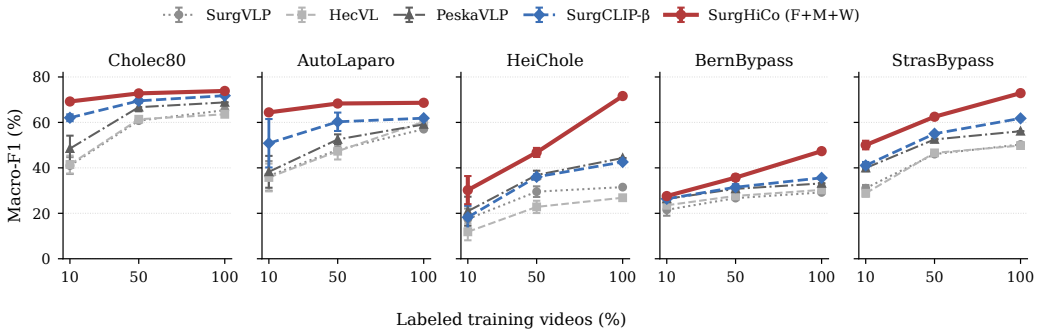


Figure 3: Video few-/full-shot linear probing across five phase-recognition datasets. Curves report video-averaged macro-F1 using 10%, 50%, and 100% of the labeled training videos. Error bars at 10% and 50% denote standard deviation over five video-subset seeds; 100% uses one run with seed 0. SurgHiCo-Fusion concatenates the frozen local, intermediate, and long-horizon representations.

Table 4: Full-shot linear probing with frozen single-scale and concatenated multi-scale representations. Entries report accuracy and macro-F1 (%). The best result is bolded and the strongest individual scale is underlined.

	AutoLaparo		BernBypass		StrasBypass	
Representation	Acc.	F1	Acc.	F1	Acc.	F1
Local	74.05	60.10	73.63	37.21	75.59	62.81
Intermediate	76.52	63.70	76.98	<u>40.37</u>	78.58	<u>66.46</u>
Long-horizon	<u>79.24</u>	<u>65.13</u>	<u>80.99</u>	39.89	<u>81.39</u>	65.97
SurgHiCo-Fusion	**80.88**	**68.64**	**82.98**	**47.36**	**84.12**	**72.89**

What does each level add? Table 3 tracks zero-shot phase recognition across the three levels. The Local representation, which is also the Fine-only flat model of Table 2, averages 38.09%. The Intermediate representation averages 40.02%, improving on Cholec80, BernBypass, and StrasBypass but falling on HeiChole and, by 7 points, on AutoLaparo. The Long-horizon representation is strongest on all five datasets and averages 50.41%, 12.32 points above the clip-level representation it is composed from, although the clip encoder is never updated after Stage 1. The intermediate level thus does not consistently improve on the local one, suggesting that the utility of a fixed 64-second context varies across procedures; the long-horizon level, whose captions concatenate several step summaries, instead matches phase prompts more closely on all five datasets.

Do longer captions help a flat model? The flat F/M/C model follows the mixed-level training of SurgCLIP-β (Perez et al., 2026), reproduced with our pipeline and protocol. Under the local evaluation context (Table 2, first block), adding Mid and Coarse captions to Fine-only training raises the five-dataset average from 38.09% to 39.47%, a gain of 1.38 points. Replacing the Coarse captions with the synthetic long-horizon captions used by SurgHiCo gives 39.94%, only 0.47 points above native Coarse captions.

Does wider coverage help without composition? Under the long evaluation context (second block), each flat model receives the same 16 local clips, or 256 sampled frames, that SurgHiCo receives, and its clip features are averaged. This adds 0.55 points with F/M/C captions (39.47% to 40.02%) and 0.53 points with F/M/Syn-W captions (39.94% to 40.47%). With the same captions and the same input, SurgHiCo reaches 50.41%, 9.94 points above mean pooling. The gain is therefore attributable to hierarchical composition rather than to the captions or the wider input.

4.4 MULTI-SCALE LINEAR PROBING

We evaluate frozen representations using video few-/full-shot linear probing. For each fraction, complete procedures are sampled from the training split and all annotated frames from each selected video are available to the probe. Figure 3 compares the concatenated representation with SurgVLP, HecVL, PeskaVLP, and SurgCLIP-β using 10%, 50%, and 100% of the training videos on five phase-recognition datasets. The 10% and 50% results report mean and standard deviation over seeds 0–4; 100% uses seed 0.

Each probe is a single linear layer trained on frozen features (Appendix C). For SurgHiCo, we L2-normalize and concatenate the three 256-dimensional local, intermediate, and long-horizon embeddings into a 768-dimensional input, which we denote SurgHiCo-Fusion. The fused representation achieves the highest macro-F1 at every displayed fraction on all five datasets. At 100% supervision, its gains over the strongest comparison method are especially large on HeiChole, BernBypass, and StrasBypass.

Table 4 isolates the contribution of multi-scale fusion under full supervision. The strongest individual scale depends on the metric: the long-horizon level gives the highest single-scale accuracy on all three datasets, whereas the intermediate level gives the highest single-scale F1 on BernBypass and StrasBypass. Concatenating all three scales is consistently best. Relative to the strongest individual scale, fusion improves accuracy/F1 by 1.64/3.51 points on AutoLaparo, 1.99/6.99 on BernBypass, and 2.73/6.43 on StrasBypass. This supports complementarity rather than uniform superiority of one temporal scale.

5 CONCLUSION

We presented SurgHiCo, which progressively composes local, intermediate, and long-horizon surgical video-language representations while keeping previously trained visual modules frozen. Across five datasets, the long-horizon representation improves average zero-shot phase accuracy by 12.32 points over the clip-level representation it is composed from. With matched caption supervision and evaluation inputs, the full model exceeds the flat mean-pooling baseline by 9.94 points. The fused representation achieves the highest performance in both few-shot and full-shot scenarios, which demonstrates its robustness and generalizability in the hierarchical surgical video understanding task.

The design has limits we make explicit. The intermediate representation does not consistently outperform the local representation, which suggests that the temporal context may vary in its utility across different surgical procedures. Evaluation uses centered offline windows, and long-horizon supervision relies on language-model paraphrases of existing captions. Our findings are therefore limited to the evaluated phase-recognition settings. Future work will examine whether the same composition framework transfers to other surgical tasks requiring long-range temporal context and to procedural reasoning with large language models.

AI USE STATEMENT

We used Gemini 3.0 Pro Thinking to rewrite the publicly available Mid-level SurgLaVi-β captions into concise one-sentence descriptions used in our long-horizon language supervision. We also used generative-AI-assisted tools for manuscript editing, formatting, and code assistance. All AI-assisted outputs were reviewed by the authors, and we take responsibility for the final content, claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

The model architecture, training objectives, data hierarchy, temporal sampling, and principal optimization settings are specified in Sections 3 and 4. The caption-generation prompt, training and evaluation windows, linear-probing protocol, computing infrastructure, and final hyperparameters are provided in Appendices A–D. Code will be released upon acceptance.

REFERENCES

Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. HierVL: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23066–23078, 2023.

Nicolas Ayobi, Santiago Rodriguez, Alejandra Perez, Isabela Hernandez, Nicolas Aparicio, Eugenie Dessevres, Sebastian Pena, Jessica Santander, Juan Ignacio Caicedo, Nicolas Fernandez, and Pablo Arbelaez. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, 2025. Advance online publication.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, pp. 813–824, 2021.

Tobias Czempiel, 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 *Medical Image Computing and Computer Assisted Intervention*, pp. 343–352, 2020.

Florent Lalys and Pierre Jannin. Surgical process modelling: A review. *International Journal of Computer Assisted Radiology and Surgery*, 9(3):495–511, 2014.

Yang Liu, Ning Zhu, Jingjing Peng, Xiwu Chen, Alejandro Granados, Guotai Wang, and Sébastien Ourselin. Stabilizing Temporal Inference Dynamics for Online Surgical Phase Recognition . In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume LNCS 16892. Springer Nature Switzerland, September 2026.

Lena Maier-Hein et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. *Medical Image Analysis*, 86:102770, 2023.

Ozanan R. Meireles, Guy Rosman, Maria S. Altieri, Lawrence Carin, Gregory Hager, Amin Madani, Nicolas Padoy, Carla M. Pugh, Patricia Sylla, Thomas M. Ward, Daniel A. Hashimoto, and the SAGES Video Annotation for AI Working Groups. SAGES consensus recommendations on an annotation framework for surgical video. *Surgical Endoscopy*, 35(9):4918–4929, 2021.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 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.

Chinedu I. Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. CholecT50: A dataset for surgical action triplet recognition. *Medical Image Analysis*, 78:102433, 2022.

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*, pp. 103982, 2026.

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 *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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.

Martin Wagner et al. Challenges in multi-centric generalization: Phase and step recognition in roux-en-y gastric bypass surgery, 2023.

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 *Medical Image Computing and Computer Assisted Intervention*, pp. 486–496, 2022.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. HiTeA: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15405–15416, 2023.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. HecVL: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *Medical Image Computing and Computer Assisted Intervention*, pp. 306–316, 2024a.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37:122952–122983, 2024b.

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*, pp. 103644, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

A Synthetic Long-Horizon Captions

Motivation. The long-horizon level is reconstructed from temporally consecutive Mid segments (Section 3) rather than supervised by the native Coarse captions. Its text target should therefore summarize a sequence of procedural steps: more abstract than the Fine captions, yet more specific than phase-level Coarse captions. Raw Mid captions are often verbose, repeat phrasing, and mix fine-grained instrument details into step-level intervals. Directly concatenating them would produce long and noisy targets. We therefore rewrite each Mid caption into one concise sentence before concatenation.

Generation pipeline. The rewriting is performed offline before Stage 2 and Stage 3 training:

1. We start from the public SurgLaVi-β Mid annotations, each consisting of `start_time`, `end_time`, and `text`.
2. Batches of 40 annotations are sent to Gemini 3.0 Pro Thinking with the fixed system prompt in Figure 4.
3. The model must return exactly one single-sentence rewrite per input row, stored as `rewritten_mid_step`; timestamps are never modified.
4. For each long-horizon window, the rewritten sentences of its Mid children are concatenated in temporal order. The resulting string, Syn-W, is the text target of the long-horizon contrastive objective.

Each rewrite paraphrases one existing Mid caption; it does not merge multiple captions or create new annotations. The prompt forbids introducing surgical events, anatomy, instruments, intentions, or causal claims absent from the original text. The Fine captions used as local supervision are the original SurgLaVi-β annotations. All long-horizon statistics in Section 4 are computed after rewriting and window construction.

Prompt and message format. The user message begins with "Below is the input JSON. Follow the system instructions and return ONLY the JSON array." It is followed by a JSON object whose `rows` field lists the Mid annotations. The model returns a JSON array aligned with the input rows, each record containing `start_time`, `end_time`, and `rewritten_mid_step`.

CapSum Mid System Prompt
Model: Gemini 3.0 Pro Thinking | Role: system

```
You are given mid-level surgical video annotations.

Hierarchy:
- fine = task level: local and detailed surgical tasks
- mid = step level: an intermediate procedural step composed of several fine-level tasks
- coarse = phase level: a higher-level surgical phase

Your task is to rewrite each mid-level annotation as a concise step-level description, while
keeping the original time span unchanged.

General rules:
- Do not change the start time or end time.
- Do not add new surgical events, anatomical structures, instruments, intentions, risks, or
causal explanations that are not supported by the original annotation.
- Do not over-infer clinical meaning. Only summarize and abstract information explicitly
present in the original annotation.
- Correct only obvious spelling or terminology errors when the correction is medically
unambiguous (e.g., "calbotomy" -> "colpotomy", "adenexial" -> "adnexal"). If a term is
uncertain, use a more general phrase rather than guessing.
- Remove redundant wording, repeated phrases, and generic phrases such as "during the
procedure" unless necessary.
- Keep the rewritten text concise and suitable for video-language alignment.
- Return exactly one sentence for each annotation.

Mid/step-level rewriting rules:
- Summarize the annotation into one procedural step.
- Preserve the main surgical action or local goal, but remove unnecessary fine-level
details.
- Remove instrument-level details unless the instrument defines the step, such as trocar
placement, stapling, clipping, suturing, or specimen retrieval.
- If the original annotation contains many actions, compress them into one or two step-level
goals in a single sentence.
- If the original annotation contains only one task-like action, keep it concise instead of
forcing a broader interpretation.
- The output should be more abstract than fine/task-level descriptions, but still more
specific than coarse/phase-level descriptions.

=== I/O FORMAT (STRICT) ===

You will receive a JSON object from the user with this shape:
{
  "rows": [
    { "start_time": <number>, "end_time": <number>, "text": "<original mid-level
    annotation>" },
    ...
  ]
}

You MUST return ONLY a JSON array (no markdown fences, no commentary). Each element
corresponds to one input row in the same order as "rows". Schema per element:
{
  "start_time": <number; MUST equal input start_time>,
  "end_time": <number; MUST equal input end_time>,
  "original": "<verbatim copy of input text>",
  "rewritten_mid_step": "<single concise English sentence; step-level rewrite>"
}

Length of the array MUST equal len(rows). Do not merge or split rows.
```

Figure 4: System prompt used to rewrite each SurgLaVi-β Mid caption into one concise sentence. Timestamps are unchanged, and the model is instructed not to introduce unsupported clinical content.

B Training and Evaluation Windows

Training uses the variable temporal extents of the data hierarchy, whereas zero-shot evaluation uses fixed contexts to standardize coverage across datasets. Table 5 summarizes both settings.

C Linear-Probing Protocol

Feature extraction. At each annotated time point, we extract three centered windows offline and encode them with the frozen model. The local branch encodes a 16-second clip (16 frames at 1 fps) with the Local Encoder, averages its temporal outputs, and applies the local projection. The intermediate branch encodes four local children (64 seconds) with the Intermediate Composer and intermediate projection. The long-horizon branch encodes four intermediate units (256 seconds) with the

Table 5: Temporal extents used for training and zero-shot evaluation. Training values are dataset averages; evaluation contexts are fixed and centered at each annotated time point.

	Training		Evaluation	
Level	Children	Span (s)	Children	Span (s)
Local	–	18.48	–	16
Intermediate	2.42	46.08	4 local	64
Long-horizon	3.67	147.93	4 intermediate	256

Long-horizon Composer followed by the frozen intermediate projection. Each branch yields a post-projection 256-dimensional embedding. For SurgHiCo-Fusion, each embedding is L2-normalized and the three are concatenated into a 768-dimensional feature. Every backbone parameter remains frozen during probing.

Probe training. Each probe is a single `nn.Linear`(d_{in}, C) layer with bias, without dropout or normalization, where C is the number of phases and $d_{\text{in}} \in \{256, 768\}$. We train for 100 epochs with cross-entropy, batch size 1024, SGD with momentum 0.9, learning rate 0.1, no weight decay, and cosine annealing. The epoch is selected by validation macro-F1.

Data fractions and splits. Following SurgLaVi-β, the 10% and 50% settings randomly sample the corresponding fraction of complete procedure videos from the training split without replacement, and all annotated frames from each selected video are used. Results are averaged over seeds 0–4. The 100% setting uses all training videos with seed 0. Cholec80 uses the official split: videos 1–40 for training, 41–48 for validation, and 49–80 for testing. For all other benchmarks, we use their official dataset splits.

Metrics. Each annotated frame is one sample. We compute accuracy and macro-F1 within each test video and average over videos.

D Implementation Details

Computing infrastructure. Pretraining and feature extraction were run on two NVIDIA RTX 6000D GPUs (85 GB each), an AMD EPYC 7402 CPU (24 cores), and 125 GB RAM under Ubuntu 20.04.6 LTS, using Python 3.10.20, PyTorch 2.12.0 with CUDA 13.0, `transformers` 4.30.2, and `timm` 1.0.27.

Hyperparameters. Table 6 lists the final settings. All stages use AdamW, cosine decay, gradient clipping at 1.0, BF16 precision, and seed 42. Captions are tokenized with BERT-base-uncased at a maximum length of 256 tokens, and the Stage-1 temperature is initialized to 0.07. For Stage 3, the batch size was selected from $\{128, 256\}$. Stage 2 freezes the Stage-1 visual modules and temperature while training the Intermediate Composer, intermediate projection, and text LoRA adapters. Stage 3 additionally freezes the Intermediate Composer and intermediate projection, trains the Long-horizon Composer, and continues training the same adapters.

Table 6: Final training hyperparameters for the TimeSformer-B model. Both composers are pre-norm Transformers with 8 heads and FFN dimension 3072.

Hyperparameter	Stage 1	Stage 2	Stage 3
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Weight decay	0.02	0.02	0.02
Epochs	50	10	10
Batch size	140×2 GPUs	375×2 GPUs	256×1 GPU
Trainable visual modules	full encoder	Intermediate Composer	Long-horizon Composer
Composer depth / width	–	2 / 768	2 / 768
Text adaptation	full encoder	LoRA (r=4, α=8)	same LoRA
λ_{loc}	–	0.3	–