LATENT-ACTION-GUIDED VIDEO-LANGUAGE FEATURE LEARNING FOR SURGICAL INSTRUMENT–TISSUE INTERACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recognizing instrument–tissue interactions is essential for context-aware surgical AI. Vision-language models offer a natural way to inject semantic structure into surgical representations by aligning video features with textual action descriptions. However, pretrained encoders may lack spatial coherence, while global semantic alignment does not ensure precise spatial and temporal representations. By analyzing frame-to-frame feature changes, we find that semantic alignment increases their dimensionality, but larger increases do not necessarily improve recognition; encoders also differ in how strongly dominant changes localize to interaction regions. Motivated by these findings, we introduce `LAViF`, which compresses frame-to-frame changes into latent actions and predicts next-frame features during end-to-end video–language alignment. Without additional spatial or motion annotations, `LAViF` improves the interaction grounding of leading feature changes and temporal-direction sensitivity in our evaluated settings. We further characterize how action capacity and prediction strength affect recognition across encoders and triplet components. Using image encoders without large-scale video pretraining, `LAViF` achieves competitive recognition with faster inference and smaller INT4 accuracy drops than V-JEPA2/2.1, supporting its deployment potential.

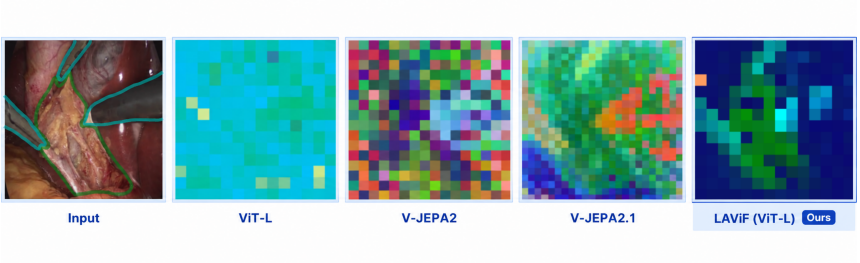

Figure 1: PCA-RGB maps encode the top three principal components of patch features as RGB channels. Our method adapts a standard ViT-L image backbone to produce spatially coherent maps visually comparable to those of V-JEPA2/2.1, while showing clearer emphasis on interacting instruments and target tissues and less background detail in the displayed example.

1 INTRODUCTION

Artificial intelligence is increasingly being developed to make surgery more context-aware, assistive, and autonomous (Yip, 2025; Carstens et al., 2026; Cella et al., 2026), requiring robust scene understanding to interpret instruments, anatomy, and their interactions in real time. Recognizing instrument–tissue interactions represents what is being done, supporting context-aware assistance and learning from surgical demonstrations. This problem is commonly formulated as surgical action-triplet recognition, which predicts $\langle \text{instrument}, \text{verb}, \text{target} \rangle$ triplet (Nwoye & Padoy, 2022). Existing methods typically treat these components as categorical prediction tasks (Nwoye et al., 2022; Liu et al., 2024; Gui & Wang, 2024). While effective, closed-set supervision provides

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.

limited semantic structure between related actions and can restrict generalization beyond predefined categories.

Vision-language models (VLMs) (Radford et al., 2021; Yuan et al., 2024a;b; 2025) offer a complementary paradigm by aligning visual representations with natural-language descriptions. Textual instrument-verb-target descriptions provide a structured semantic space in which related triplets share compositional structure (Sharma et al., 2025; Xi et al., 2023). Language-aligned visual representations also offer a natural-language interface for connecting perception with higher-level reasoning and planning in large surgical AI systems (Chen et al., 2023; Mao et al., 2026). More broadly, LaViLa (Zhao et al., 2023) and EgoVLP (Lin et al., 2022) demonstrate transfer of language-supervised video representations to action recognition and retrieval. In surgery, descriptions can be obtained from structured action labels or, at larger scale, narrated video through automatic transcription and language processing (Yuan et al., 2025; 2024a;b). For downstream tasks like surgical action, pretrained visual encoders can be kept frozen and their features matched to text-derived representations (Rao et al., 2025; Ikeido et al., 2026). Alternatively, language supervision can guide adaptation of the visual representations (Sharma et al., 2025; Li et al., 2025; Che et al., 2026a; Walimbe et al., 2025). These approaches face two representation-level limitations.

(1) Limitations inherited from frozen pretrained representations. Pretrained features can remain noisy or spatially imprecise (Khattak et al., 2023; Menghini et al., 2023; Wang et al., 2024; El Banani et al., 2024). Language-supervised pretraining provides strong semantics but does not necessarily yield spatially coherent patch representations (Wang et al., 2025a; Lan et al., 2024). Self-supervised visual objectives (Wu et al., 2026; Jamal & Mohareri, 2023; Batić et al., 2023) may recover stronger dense structure while retaining task-irrelevant features (Darcet et al., 2024; Vanyan et al., 2023). Freezing the encoder preserves these limitations.

(2) Global semantic alignment weakly constrains interaction-relevant visual structure. End-to-end semantic alignment encourages global representations to match target semantics but provides limited guidance on the local evidence supporting that match. Encoders may exploit background context, surgical phase, camera motion, or other correlated cues (Kumar et al., 2022; Ding et al., 2022; Choi et al., 2025; Lan et al., 2024; Zohra et al., 2026), without consistently organizing spatial and temporal feature changes around the instrument–tissue interaction.

These limitations matter for surgical action recognition, which requires localized interaction cues and temporal context (Nwoye et al., 2022; Lin et al., 2024), while complex and variable anatomy makes relevant structures difficult to distinguish (Madani et al., 2022; Honda et al., 2026; Cheng et al., 2026b). General-domain methods introduce spatial supervision through detector-derived regions, region grouping, or segmentation guidance (Chen et al., 2025; Zeng et al., 2022; Sharma et al., 2025; Zohra et al., 2026; Choi et al., 2025); surgical approaches similarly use instrument trajectories, explicit instrument–tissue interaction guidance, or interaction pseudo-labels (Cheng et al., 2026a; Pei et al., 2025b; Sharma et al., 2023b). However, automatically estimated spatial and motion cues can be unreliable under occlusion, tissue deformation, and blurred anatomical boundaries. We therefore aim to learn spatially grounded and temporally direction-sensitive representations through semantic alignment, without additional spatial or motion supervision.

The importance of localized action cues (Pei et al., 2025a; Wang et al., 2025b; Long et al., 2025) and temporal direction (Bagad & Zisserman, 2025) motivates analyzing *transition structure*: how frame-to-frame feature changes are distributed across feature-space directions and where those directions are spatially grounded. Let $s_t \in \mathbb{R}^{d_s}$ denote the visual feature vector (state) of frame t. We define the transition $\delta_t = s_{t+1} - s_t$ and analyze the spectrum of its covariance $\Sigma_\delta = \mathrm{Cov}(\delta)$ under semantic alignment training. Across the evaluated encoders, semantic alignment expands the effective transition space, but greater expansion does not consistently imply better recognition or stronger temporal-direction sensitivity. Instead, strong recognition can coexist with a compact transition space whose dominant changes are grounded in instrument–tissue interactions. This motivates investigating whether such transition structure can be encouraged during semantic alignment.

We therefore investigate shaping transition structure through joint predictive and semantic learning. Prior analysis shows that next-state prediction through compact latent actions can preserve dominant transition directions in fixed representations (Zhang et al., 2025). Extending this setting to a jointly trained encoder, `LAViF` uses a *Causal Transition IDM* to compress current and preceding transitions into action tokens. These tokens condition next-state prediction through a forward dynamics

model (FDM) and join frame states for video–language alignment, supporting both dynamics and interaction semantics. A patch-level SIG regularizer further stabilizes local features during joint training. Related inverse/forward dynamics and latent-action work is discussed in Section 5.

Our contributions are threefold: **(1) Empirical finding.** We show that semantic alignment reshapes visual transition structure: strong recognition can coincide with compact spectra and interaction-grounded leading directions, whereas greater dimensional expansion alone does not imply better recognition (Section 2). **(2) Finding-driven method.** We introduce `LAViF` to investigate and explicitly shape this structure through a Causal Transition IDM, a compact predictive action bottleneck, video–language alignment, and patch-level SIG regularization (SIGReg) (Section 3). **(3) Empirical validation.** Across encoders and datasets, we show that `LAViF` can improve interaction grounding, temporal-direction sensitivity, and recognition without additional spatial or motion supervision, and characterize the effects of bottleneck capacity and prediction strength (Section 4). We also demonstrate faster inference and smaller performance drops under INT4 quantization than the evaluated pretrained video encoders.

2 Transition Structure under Semantic Alignment Training

We first characterize how semantic alignment training reshapes frame-to-frame visual transitions, and organize our analysis around the following empirical observations.

Setup. We use CholecT50 (Nwoye & Padoy, 2022), a laparoscopic cholecystectomy action-triplet benchmark. For each annotated frame x_t, we form an eight-frame clip $\{x_{t-7}, \ldots, x_t\}$ from the 1 fps sequence and predict its instrument–verb–target labels at x_t. All encoders use the same temporal context, video–text objective, and EmbeddingGemma text encoder (Vera et al., 2025). Image encoders (ViT-L (Dosovitskiy et al., 2021), DINOv2-L (Oquab et al., 2024), SurgeNet-L (Jaspers et al., 2026)) process frames independently followed by a CLIP4Clip-style temporal Transformer (Luo et al., 2021), whereas native video encoders (TimeSformer, V-JEPA2 (Assran et al., 2025), and V-JEPA2.1 (Mur-Labadia et al., 2026)) process clips jointly. For patch features ϕ_p^t, we define $s_t = \frac{1}{N}\sum_{p=1}^{N} \phi_p^t$ and $\delta_t = s_{t+1} - s_t$. We use *transition structure* to denote **(1)** how transition variance is distributed across feature-space directions and **(2)** where those directional changes occur spatially. We estimate $\Sigma_\delta = \mathrm{Cov}(\delta)$ from 20,000 consecutive-frame pairs and summarize its spectrum by d_{90}, the minimum number of leading directions explaining 90% of transition variance. Spatial grounding is measured by projecting patch transitions $\phi_p^{t+1} - \phi_p^t$ onto covariance-eigenvector bands, while temporal-direction sensitivity is measured by the relative recognition drop under clip reversal.

Observation 1: A larger transition space does not consistently imply better recognition or stronger temporal-direction sensitivity. Semantic alignment increases d_{90} for every evaluated encoder (Figure 2(a)). After training, ViT-L and TimeSformer have larger d_{90} (537 and 431) than V-JEPA2.1 (203), yet lower mAP_{IVT} (31.29 and 27.32 versus 34.82) and smaller reversal drops (21.8% and 5.1% versus 30.0%). Thus, transition dimensionality alone does not explain recognition or temporal-direction sensitivity.

Observation 2: Interaction grounding of transition structure varies across encoders and transition ranks. We rank the eigenvectors of Σ_δ by decreasing variance and group them into bands of 16. Each patch is scored by the fraction of its frame-to-frame feature-change energy captured by a band; the highest-scoring 25% are foreground and the remainder background, forming a binary saliency map. We compute its IoU with the instrument–target region formed by combining SASVi-estimated segmentation masks from both frames (Sivakumar et al., 2025), averaging over 5,000 evaluation frames. Projection and mask-construction details are provided in Appendix A.3. Figure 2 (a, b) shows that V-JEPA2.1, with the highest mAP_{IVT} (34.82), has higher IoU in its leading bands than its tail; TimeSformer shows the opposite trend with lower recognition (27.32). DINOv2-L exceeds ViT-L in several leading bands, while SurgeNet-L exhibits a pronounced first-band peak; both outperform ViT-L in recognition. Notably, these self-distillation-based encoders (DINOv2, SurgeNet, V-JEPA2.1) exhibit stronger localization in their dominant directions of transition structure. Together, these observations suggest that a larger effective transition space does not necessarily imply that more transition directions capture action-relevant information, as some may instead reflect task-irrelevant variation. This motivates the hypothesis that transition directions of

(a) Transition dimensionality, recognition, and reversal

Encoder	d_{90} Before → after	mAP (%) IVT	I	V	T	Reversal drop (%) Δ_{IVT}	Δ_{I}	Δ_{V}	Δ_{T}
Image encoders									
DINOv2-L	278 → 391	33.52	94.26	67.26	48.09	19.6	13.1	13.8	16.1
ViT-L	406 → 537	31.29	91.66	66.41	45.56	21.8	13.9	15.1	13.0
SurgeNet-L	261 → 404	33.09	94.29	69.62	48.58	28.2	15.5	**19.0**	14.7
Video encoders									
TimeSformer	333 → 431	27.32	88.67	64.69	44.43	5.1	3.8	3.4	4.2
V-JEPA2	80 → 231	29.68	92.16	68.55	47.97	28.5	14.0	16.5	17.2
V-JEPA2.1	**45 → 203**	**34.82**	**93.99**	**69.76**	**50.23**	**30.0**	**17.7**	17.0	14.4

(b) Spatial grounding of transitions

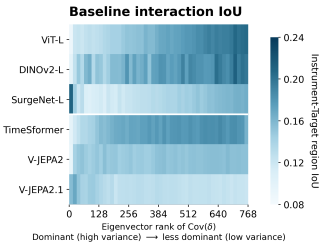

Figure 2: **Transition structure on CholecT50. (a)** d_{90} before and after semantic alignment, mAP after training, and relative mAP drop under clip reversal. **(b)** Per-band IoU with the instrument–target region; higher-variance transition directions appear on the left.

different visual encoders may differ in their task relevance, and that stronger localization of dominant directions to interaction regions may be associated with better surgical action recognition.

These observations on the transition structure raise a question: *Can transition structure be explicitly shaped during semantic alignment so that dominant visual changes better reflect instrument–tissue interactions while mitigating irrelevant variation?* Recent work (Zhang et al., 2025) shows that compressing transitions into a low-dimensional latent action for next-state prediction can preserve dominant transition directions when visual features are fixed. We investigate whether jointly training the encoder with this training objective can reshape the transition structure and improve where they are grounded, without additional spatial or motion annotations.

3 Latent-Action-Guided Design

To investigate whether transition structure can be explicitly shaped during semantic alignment, and whether doing so benefits surgical action recognition, we propose `LAViF`. Building on latent-action modeling for predictive representation learning and control (Zhang et al., 2025; Ye et al., 2025; Schmidt & Jiang, 2024; Cui et al., 2024), `LAViF` introduces compact latent actions that jointly support next-state prediction and semantic interaction recognition. The framework consists of a vision encoder E_ϕ, an inverse dynamics model (IDM) I_θ, a Forward Dynamics Model (FDM) F_ψ, a clip aggregator A_ω, and a text encoder G_χ (Figure 3).

Vision Encoder. For each frame, E_ϕ produces patch tokens $\mathbf{p}_t$ and a pooled visual state $s_t = \mathrm{pool}(\mathbf{p}_t) \in \mathbb{R}^{d_s}$. All visual representations are jointly adapted during training.

Causal Transition Inverse Dynamics Model. We represent each state transition by $\delta_t = s_{t+1} - s_t$, following displacement-based action decoding (Zhang et al., 2026). The IDM processes the transition sequence with a causally masked Transformer and outputs

$$a_t = I_\theta(\delta_{\le t}) \in \mathbb{R}^{d_a}, \qquad d_a \ll d_s. \tag{1}$$

The causal mask incorporates transition history while preserving temporal asymmetry, and the low-dimensional a_t forms a compact bottleneck for transition-specific information.

Forward Dynamics Model. The FDM predicts the next visual state from the current state history conditioned on the latent action, $\hat{s}_{t+1} = F_\psi(s_{\le t}, a_t)$. We train it with a stop-gradient prediction objective,

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{T}\sum_{t=0}^{T-1} \|\hat{s}_{t+1} - \mathrm{sg}(s_{t+1})\|_2^2 . \tag{2}$$

Because transition-specific information is routed through the compact a_t, this objective encourages the latent action to retain predictive state changes. The FDM uses a causal DiT-style architecture with adaLN-zero conditioning (Peebles & Xie, 2023; Maes et al., 2026).

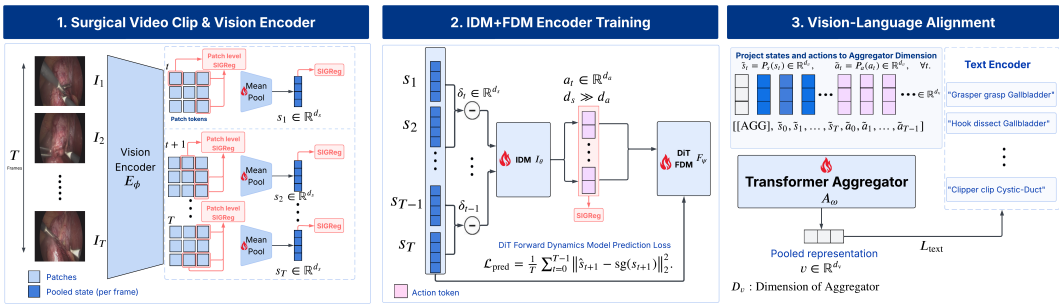

Figure 3: **Overview of LAViF.** From left to right: **(1)** the surgical video is encoded into frame-level visual states; **(2)** the causal IDM maps state transitions to latent actions, which condition FDM next-state prediction; and **(3)** state–action tokens are aggregated and aligned with text embeddings for surgical interaction recognition. SIGReg regularizes state, action, and patch features.

SIG Regularization. We apply SIGReg (Maes et al., 2026; Balestriero & LeCun, 2025) to state, action, and patch representations. Specifically,

$$\mathcal{L}_{\text{SIG}} = \text{SIGReg}(\{s_t\}) + \text{SIGReg}(\{a_t\}), \qquad \mathcal{L}_{\text{SIG}}^{\text{patch}} = \text{SIGReg}\big(\{p_{t,k} : k \in \mathcal{K}_t\}\big), \quad |\mathcal{K}_t| = 32,$$

where $\mathcal{K}_t$ contains 32 patch tokens sampled uniformly from frame t, resampled independently for each clip and frame at every training step. State regularization prevents representation collapse, action regularization discourages the FDM from ignoring the latent action, and patch-level SIGReg preserves local feature diversity during adaptation. The SIGReg statistic and projection settings are given in Appendix A.4.

Aggregation and Semantic Alignment. The aggregator A_ω linearly projects state and action tokens, prepends a learnable `[AGG]` token, and processes the full sequence with a bidirectional Transformer: $v = A_\omega\big([\texttt{[AGG]}, s_0, \ldots, s_T, a_0, \ldots, a_{T-1}]\big)$. Combining states (*what is present*) and latent actions (*what is changing*) allows the IDM to receive semantic supervision through the recognition pathway together with predictive supervision from the FDM, while jointly adapting the visual encoder. EmbeddingGemma (Vera et al., 2025) encodes complete instrument–verb–target descriptions and their individual components. We additionally supervise instrument, verb, and target semantics, following prior work that uses component-level supervision for fine-grained action and video–language learning (Wray et al., 2019; Momeni et al., 2023; Vo et al., 2024; Sharma et al., 2025). Component labels are obtained directly from the fixed triplet-to-component mapping. Temperature-scaled cosine similarity between clip and text embeddings produces triplet and component logits, optimized with binary cross-entropy:

$$\mathcal{L}_{\text{text}} = \lambda_{\text{IVT}}\mathcal{L}_{\text{IVT}} + \sum_{q\in\{\text{I,V,T}\}} \lambda_q \mathcal{L}_q.$$

The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}} + \lambda_s\mathcal{L}_{\text{SIG}} + \lambda_p\mathcal{L}_{\text{SIG}}^{\text{patch}}.$$

The FDM is used only during training; at inference, E_ϕ, I_θ, and A_ω produce the clip representation matched against instrument–verb–target text embeddings. We use $\lambda_{\text{pred}} = 1$ for the full model, while $\lambda_{\text{pred}} = 0$ retains the IDM and state–action recognition pathway. Unless explicitly stated otherwise, all controlled baselines and ablations are trained with the same BCE-based semantic alignment objective. Additional architectural and optimization details are provided in Appendix A.4.

4 Experiments

We first describe the datasets, baselines, ablations, and metrics. Section 4.1 evaluates recognition, component ablations, temporal-direction sensitivity, and deployment efficiency. The default full model uses $d_a = 128$ and $\lambda_{\text{pred}} = 1$. Section 4.2 varies d_a and λ_{pred} to study transition structure and recognition, using the transition-band IoU from Section 2 and PCA-RGB maps to examine spatial grounding and local feature organization.

Table 1: Main comparison on CholecT50 and ProstaTD with $d_a = 128$. † denotes frame-wise image encoders and ‡ denotes native video encoders. △ marks methods using a classification head without video–language alignment. ◇ marks methods evaluated with different metrics and therefore requiring cautious comparison. All remaining baselines and LAViF variants are trained under our controlled training protocol. Best and second-best results are shown in **bold** and underline, respectively.

Method	CholecT50 (RDV)				ProstaTD (5-fold)			
	$\mathbf{mAP_{IVT}}$	$\mathbf{mAP_{I}}$	$\mathbf{mAP_{V}}$	$\mathbf{mAP_{T}}$	$\mathbf{mAP_{IVT}}$	$\mathbf{mAP_{I}}$	$\mathbf{mAP_{V}}$	$\mathbf{mAP_{T}}$
Published task-specific methods								
Rendezvous (Nwoye et al., 2022)†△	29.90	92.20	60.07	38.30	–	–	–	–
Chain-of-Look (Xi et al., 2023)†△	**38.00**	94.10	62.50	41.90	–	–	–	–
MML-SurgAdapt (Walimbe et al., 2025)†	30.30	87.30	61.10	43.20	–	–	–	–
TDNet (Chen et al., 2026)†◇△	–	–	–	–	36.10 ± 3.40	89.90 ± 1.30	61.70 ± 2.90	55.70 ± 2.40
Controlled baselines								
General-domain pretraining								
CLIP ViT-L/14†	26.17	86.26	62.54	42.93	26.80 ± 0.96	90.97 ± 0.76	65.00 ± 3.28	53.51 ± 2.71
DINOv2-L†	27.41	90.08	65.28	45.44	36.01 ± 1.95	93.13 ± 0.72	70.75 ± 2.78	63.84 ± 2.32
V-JEPA2‡	29.68	92.16	68.55	47.97	32.66 ± 1.85	91.30 ± 1.27	71.89 ± 3.29	59.95 ± 1.78
V-JEPA2.1‡	34.82	93.99	69.76	50.23	35.87 ± 2.50	94.36 ± 1.80	**74.35** ± 2.47	62.64 ± 2.47
TimeSformer‡	27.32	88.67	64.69	44.43	29.43 ± 2.11	86.19 ± 1.71	64.69 ± 3.19	56.14 ± 3.25
Surgical-domain pretraining								
EndoViT†	24.71	85.19	61.41	41.23	26.94 ± 0.89	88.35 ± 0.82	62.02 ± 3.97	53.63 ± 2.44
LEMON†	28.71	87.34	65.04	44.70	32.26 ± 1.22	91.72 ± 1.33	66.51 ± 3.29	59.71 ± 1.41
SurgVLP†	20.28	73.96	52.33	37.76	20.64 ± 0.30	78.07 ± 1.73	52.45 ± 2.29	45.88 ± 2.36
HecVLP†	21.41	77.59	57.23	39.29	22.42 ± 0.42	81.16 ± 2.68	55.89 ± 2.70	48.76 ± 2.65
PeskaVLP†	22.02	79.20	57.22	39.26	23.09 ± 0.64	82.45 ± 2.01	56.99 ± 3.44	50.75 ± 3.34
LAViF – ViT-L	32.15	93.24	68.29	47.05	32.82 ± 1.29	92.82 ± 1.00	68.28 ± 2.80	60.63 ± 1.73
LAViF – DINOv2-L	33.13	94.60	68.91	48.81	38.14 ± 2.61	96.03 ± 2.19	73.01 ± 2.98	65.21 ± 2.92
LAViF – SurgeNet-L	32.78	**95.02**	**71.64**	**50.89**	**38.85** ± 2.07	**96.74** ± 1.24	72.82 ± 1.32	**65.73** ± 2.38

Datasets. We evaluate on two surgical action-triplet benchmarks. **CholecT50** (Nwoye & Padoy, 2022) contains 50 laparoscopic cholecystectomy videos (90,489 frames at 1 fps) and 100 triplet classes over 6 instruments, 10 verbs, and 15 targets; we use the standard Rendezvous (RDV) split. **ProstaTD** (Chen et al., 2026) contains 21 robot-assisted radical prostatectomy videos (∼72K frames) and 89 triplet classes over 7 instruments, 10 verbs, and 10 targets; we use its five-fold protocol. Together they span general surgery and urology, including hepatobiliary dissection and crowded pelvic scenes with difficult tissue-layer identification (Madani et al., 2022; Honda et al., 2026; Chen et al., 2026). For each labeled frame x_t, we form $\mathbf{X}_t = \{x_{t-7}, \ldots, x_t\}$ and assign it the label at x_t. Image baselines use only x_t, native video encoders use the full clip, and LAViF constructs an eight-frame state–action representation from pretrained image backbones.

Baselines. We compare general-domain encoders CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), V-JEPA2/2.1 (Assran et al., 2025; Mur-Labadia et al., 2026), and TimeSformer (Bertasius et al., 2021); surgical-domain encoders EndoViT (Batić et al., 2023), LEMON (Che et al., 2026b), SurgVLP (Yuan et al., 2025), HecVLP (Yuan et al., 2024a), and PeskaVLP (Yuan et al., 2024b); and task-specific methods Rendezvous (Nwoye et al., 2022), Chain-of-Look (Xi et al., 2023), and MML-SurgAdapt (Walimbe et al., 2025). We include V-JEPA2 as a strong pretrained video encoder and V-JEPA2.1 as its enhanced variant, which further incorporates self-distillation to promote more spatially coherent, object-level representations. The task-specific methods use their original frameworks and serve as reference comparisons; Chain-of-Look also uses task-specific prior information. We instantiate LAViF with ViT-L/16, DINOv2-L/16, and SurgeNet-L/14 (Jaspers et al., 2026), representing vanilla, general-domain self-supervised, and surgical-domain self-supervised pretraining. These models and controlled baselines are fully trainable with the same EmbeddingGemma text encoder (Vera et al., 2025) and BCE-based semantic alignment objective from Section 3. Baseline and training details are in Appendix A.2.

Ablations and metrics. Within each backbone and dataset, all variants share data splits, learning rates, training schedules, epoch budgets, and semantic objectives. We compare the no-IDM+FDM baseline, IDM only, IDM+FDM without patch-level SIGReg, and the full model. IDM only disables both forward prediction and patch-level SIGReg; the no-patch-SIGReg variant isolates the regularizer. Note that without IDM+FDM, image encoders use frame-wise features followed by a

Table 2: Component ablations at $d_a = 128$. Values are recognition mAP (%); ProstaTD reports mean$_{\pm\text{std}}$ over five folds.

		CholecT50				ProstaTD			
Backbone	Variant	IVT	I	V	T	IVT	I	V	T
ViT-L	Full	32.15	93.24	68.29	47.05	$32.82_{\pm 1.29}$	$92.82_{\pm 1.00}$	$68.28_{\pm 2.80}$	$60.63_{\pm 1.73}$
	No Patch-Level SIGReg	31.34	91.45	67.09	46.03	$32.15_{\pm 1.66}$	$92.17_{\pm 0.63}$	$67.69_{\pm 2.84}$	$54.29_{\pm 2.17}$
	IDM only	30.74	91.71	65.48	46.41	$30.75_{\pm 1.73}$	$91.67_{\pm 1.04}$	$65.90_{\pm 2.93}$	$58.12_{\pm 2.38}$
	No IDM+FDM	31.29	91.66	66.41	45.56	$30.79_{\pm 1.21}$	$90.87_{\pm 0.82}$	$65.12_{\pm 2.61}$	$55.09_{\pm 1.84}$
DINOv2-L	Full	33.13	94.60	68.91	48.81	$38.14_{\pm 2.61}$	$96.03_{\pm 2.19}$	$73.01_{\pm 2.98}$	$65.21_{\pm 2.92}$
	No Patch-Level SIGReg	31.54	91.69	66.90	43.08	$37.10_{\pm 2.87}$	$95.66_{\pm 0.39}$	$71.05_{\pm 1.84}$	$64.19_{\pm 3.39}$
	IDM only	30.89	94.35	66.87	47.81	$37.21_{\pm 2.62}$	$95.62_{\pm 1.95}$	$72.02_{\pm 2.61}$	$64.22_{\pm 3.62}$
	No IDM+FDM	33.52	94.26	67.26	48.09	$36.27_{\pm 1.88}$	$95.03_{\pm 1.53}$	$71.21_{\pm 2.16}$	$63.77_{\pm 2.98}$
SurgeNet-L	Full	32.78	95.02	71.64	50.89	$38.85_{\pm 2.07}$	$96.74_{\pm 1.24}$	$72.82_{\pm 1.32}$	$65.73_{\pm 2.38}$
	No Patch-Level SIGReg	32.53	93.93	69.03	49.79	$38.17_{\pm 2.43}$	$96.19_{\pm 1.47}$	$72.74_{\pm 2.98}$	$64.35_{\pm 3.65}$
	IDM only	32.56	94.61	69.56	48.34	$37.60_{\pm 2.52}$	$96.11_{\pm 2.30}$	$72.21_{\pm 3.20}$	$64.87_{\pm 3.00}$
	No IDM+FDM	33.09	94.29	69.62	48.58	$34.81_{\pm 2.12}$	$96.42_{\pm 1.30}$	$70.69_{\pm 1.90}$	$64.41_{\pm 2.85}$

CLIP4Clip-style temporal Transformer (Luo et al., 2021), as in Section 2. We report mAP_{IVT} and component-level mAP_I, mAP_V, and mAP_T using *ivtmetrics* (Nwoye & Padoy, 2022); ProstaTD reports mean $\pm$ standard deviation over five folds. Deployment evaluation measures per-clip latency, throughput, peak GPU memory, and mAP_{IVT} changes from FP32 under INT8/INT4 post-training weight-only quantization without fine-tuning.

4.1 Main Recognition Performance and Ablations

Main recognition results. Table 1 reports recognition on CholecT50 and ProstaTD. With ViT-L, `LAViF` reaches $32.15/32.82$ mAP_{IVT} on CholecT50/ProstaTD, exceeding V-JEPA2 on CholecT50 (29.68) and remaining comparable on ProstaTD (32.66). It also achieves higher mAP_I on both datasets and comparable $\text{mAP}_V/\text{mAP}_T$ on CholecT50, showing that an image-pretrained ViT without temporal pretraining can compete with a pretrained video encoder. With DINOv2-L and SurgeNet-L, `LAViF` reaches 38.14 and 38.85 mAP_{IVT} on ProstaTD, versus 35.87 for V-JEPA2.1. SurgeNet-L attains the highest mean triplet, instrument, and target mAP among the listed methods, while V-JEPA2.1 retains the highest verb mAP. On CholecT50, SurgeNet-L exceeds V-JEPA2.1 in all three component metrics but not mAP_{IVT}. Patch-level SIGReg has a relatively modest effect for SurgeNet-L on ProstaTD. Section 4.2 studies the effects of d_a and prediction weight.

Table 3: Relative mAP drop (%) under clip reversal on CholecT50 ($d_a = 128$).

Backbone	Variant	Δ_{IVT}	Δ_{I}	Δ_{V}	Δ_{T}
ViT-L	Full	25.3	15.7	18.5	9.7
	IDM only	21.4	14.1	16.0	9.2
	No IDM+FDM	21.8	13.9	15.1	13.0
DINOv2-L	Full	27.9	16.6	17.8	<u>16.7</u>
	IDM only	23.1	15.7	16.4	14.4
	No IDM+FDM	19.6	13.1	13.8	16.1
SurgeNet-L	Full	**32.1**	**18.9**	**21.3**	14.7
	IDM only	28.7	<u>18.6</u>	**21.3**	14.2
	No IDM+FDM	28.2	15.5	<u>19.0</u>	14.7
V-JEPA2	Baseline	28.5	14.0	16.5	**17.2**
V-JEPA2.1	Baseline	<u>30.0</u>	17.7	17.0	14.4
TimeSformer	Baseline	5.1	3.8	3.4	4.2

Component ablations at $d_a = 128$. Table 2 reports ablations at $d_a = 128$. Full ViT-L improves all four metrics on both datasets, gaining $0.86/2.03$ mAP_{IVT} on CholecT50/ProstaTD. DINOv2-L and SurgeNet-L improve verb/target mAP on CholecT50 but remain slightly below their triplet baselines, while on ProstaTD all three full models improve triplet, verb, and target mAP; DINOv2-L and SurgeNet-L gain $1.87/4.04$ triplet points. Adding FDM prediction and patch-level SIGReg over IDM alone further improves triplet, verb, and target mAP across backbones and datasets. Section 4.2 studies capacity and prediction weight. For temporal-direction sensitivity, full `LAViF` increases verb reversal drops from $13.8 \rightarrow 17.8\%$ (DINOv2-L),

Table 4: Relative mAP drop (%) under INT4 quantization on CholecT50 ($d_a = 128$); lower is better.

Backbone	Variant	Δ_{IVT}	Δ_{I}	Δ_{V}	Δ_{T}
ViT-L	Full	**2.7**	0.9	0.8	0.0
	IDM only	10.5	1.9	3.5	0.6
	No IDM+FDM	4.2	0.1	0.5	2.9
DINOv2-L	Full	**2.8**	0.3	0.2	0.1
	IDM only	6.8	0.4	0.3	1.3
	No IDM+FDM	4.2	0.5	0.9	0.9
SurgeNet-L	Full	$-$**0.4**	0.1	0.6	-0.2
	IDM only	0.3	0.0	-1.4	-0.6
	No IDM+FDM	1.3	0.2	1.0	0.5
V-JEPA2	Baseline	7.7	-0.2	2.4	3.9
V-JEPA2.1	Baseline	5.6	2.4	4.8	4.7

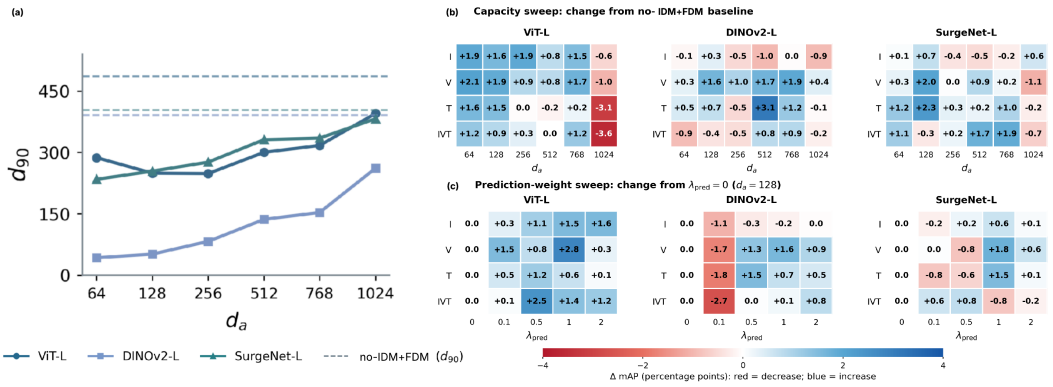

Figure 4: **Transition structure and recognition under action-capacity and prediction-weight sweeps on CholecT50. (a)** d_{90} versus d_a; dashed lines denote no-IDM+FDM baselines. **(b)** Recognition change across d_a relative to the no-IDM+FDM baseline. **(c)** Recognition change across λ_{pred} at $d_a = 128$, relative to $\lambda_{\text{pred}} = 0$. Heatmaps report I/V/T/IVT mAP differences (percentage points); blue/red indicate gains/drops.

$15.1 \rightarrow 18.5\%$ (ViT-L), and $19.0 \rightarrow 21.3\%$ (SurgeNet-L), exceeding V-JEPA2/2.1 without large-scale video pretraining (Table 3). ProstaTD results are in Appendix A.6.

Inference efficiency and deployment potential. For T frames with N patches, LAViF uses $O(TN^2 + T^2)$ attention versus $O(T^2N^2)$ for joint spatiotemporal attention, and removes the FDM at inference. ViT-L/DINOv2-L reach 72.8/50.8 clips/s, respectively. Under INT4 weight-only quantization of all linear layers, including the vision encoder, IDM, aggregator, and text encoder, the full LAViF models lose at most 2.8% in mAP_{IVT} under INT4 quantization, compared with 1.3–4.2% for the corresponding no-IDM+FDM baselines and 5.6–7.7% for V-JEPA2 and V-JEPA2.1 (Table 4). We interpret this result as being consistent with greater low-bit tolerance from decoupled frame-wise spatial encoding and compact temporal aggregation. More comprehensive results tables are provided in Appendix A.10.

4.2 Transition Structure and Recognition

Action capacity and predictive pressure. Under a simplified linear model (Zhang et al., 2025), the optimal prediction loss is $\mathcal{L}^*_{\text{pred}} = \sum_{i>d_a} \lambda_i$, where λ_i are transition-covariance eigenvalues in descending order. Thus, d_a controls how much dominant transition variance can be represented by the latent action; smaller d_a leaves more variance outside the bottleneck and contributes to prediction error. In joint training, this pressure may discourage task-irrelevant changes, while semantic supervision preserves directions useful for recognition. We summarize the resulting transition space using d_{90}, the number of directions explaining 90% of its variance. Figure 4(a,b) shows that ViT-L improves all four recognition metrics at $d_a \in \{64, 128\}$ even though $d_{90} > d_a$, indicating that the full transition space need not fit inside the action bottleneck. DINOv2-L reaches its highest triplet mAP at $d_a = 768$, while removing the bottleneck at $d_a = d_s = 1024$ lowers triplet mAP relative to $d_a = 768$ for all three encoders. SurgeNet-L similarly favors $d_a = 128$ for verb mAP but $d_a = 768$ for triplet mAP. At fixed $d_a = 128$, increasing λ_{pred} from 0 to 1 improves ViT-L verb/triplet mAP by $2.81/1.40$ points and SurgeNet-L verb/target mAP by $1.75/1.51$ points, while reducing its triplet mAP by 0.79 points (Figure 4(c)). Overall, predictive pressure can help select useful transitions, but stronger pressure or greater action capacity does not uniformly benefit all components.

Spatial feature reorganization and transition grounding. PCA-RGB qualitatively reflects patch-level spatial coherence, while transition-band IoU measures where dominant feature changes are grounded. Figure 5a shows that `LAViF` often makes target tissues more distinct than the no-IDM+FDM baselines and V-JEPA2/2.1, with the strongest changes for ViT-L and weaker effects for SurgeNet-L. Figure 5c further suggests an encoder-dependent effect of λ_{pred} on boundary clarity and local detail. Using the setup of Section 2, Figure 5b shows that `LAViF` increases IoU in several higher-variance transition bands while reducing it in many later bands. Thus, latent-action training

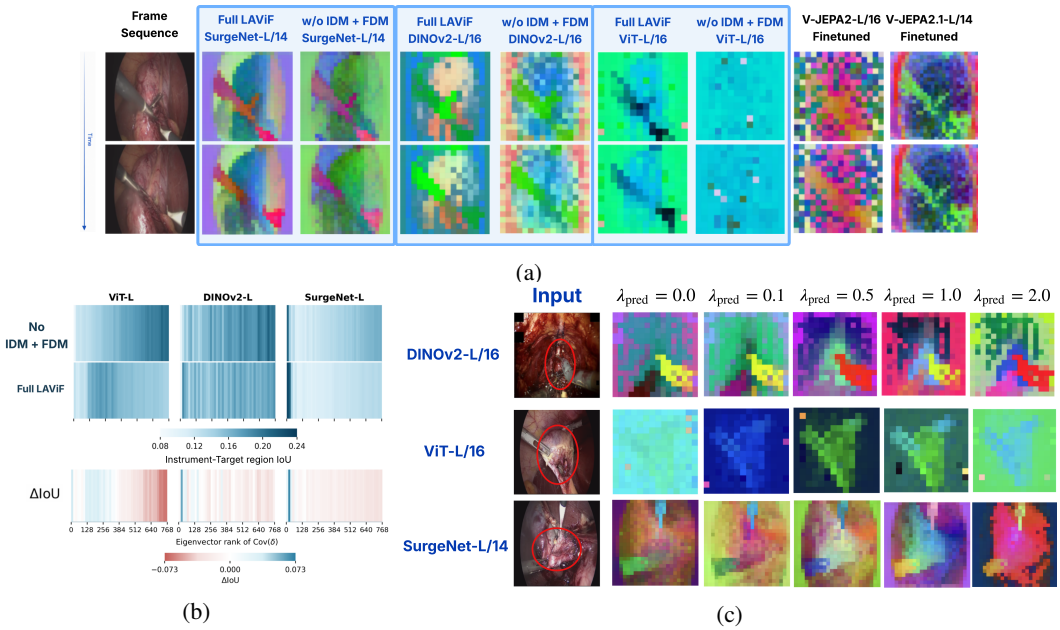

Figure 5: **Spatial feature organization and interaction grounding.** **(a)** PCA-RGB maps comparing LAViF, no-IDM+FDM baselines, and V-JEPA2/2.1. **(b)** Transition-band interaction-region IoU at $d_a = 128$; ΔIoU is relative to the no-IDM+FDM baseline. **(c)** PCA-RGB maps across λ_{pred}.

reorganizes both local patch structure and the rank-dependent grounding of transition features rather than uniformly improving localization. Additional examples and sweeps are provided in Appendices A.7 and A.8.

5 Related Work

Surgical action-triplet recognition models ⟨instrument, verb, target⟩ interactions using attention, temporal modeling, multi-task learning, and vision–language supervision (Nwoye et al., 2020; 2022; Sharma et al., 2023a; Li et al., 2023; Sharma et al., 2025; Walimbe et al., 2025; Xi et al., 2023). Recent methods also use instrument trajectories, explicit interaction guidance, or pseudo-labels (Cheng et al., 2026a; Pei et al., 2025b; Sharma et al., 2023b). We instead study how semantic alignment reorganizes visual representations without additional spatial or motion supervision. **Latent-action models** infer hidden actions from state transitions using inverse and forward dynamics, primarily for action discovery, control, and VLA learning (Edwards et al., 2019; Cui et al., 2024; Ye et al., 2025; Schmidt & Jiang, 2024; Bu et al., 2025; Sun et al., 2026). Prior methods often operate on fixed visual features, while predictive work studies how such objectives shape representations (Maes et al., 2026; Zhang et al., 2025). In contrast, LAViF jointly optimizes the visual encoder, IDM, FDM, temporal aggregator, and semantic objective to reshape visual transition structure during video–language alignment. Extended discussion is in Appendix A.1.

6 Conclusion and Future Work

We analyze how semantic alignment reshapes visual transition structure and find that greater transition-space expansion does not necessarily imply stronger recognition. Guided by this observation, we introduce LAViF to study whether latent-action prediction can reshape transition structure for surgical interaction recognition without additional spatial or motion supervision. Across multiple image backbones and two surgical datasets, we evaluate recognition, interaction grounding, temporal-direction sensitivity, and their dependence on d_a and λ_{pred}. Results show encoder- and dataset-dependent reorganization of dominant transition directions and local feature structure. Future work will study encoder-aware parameter selection, broader surgical and non-surgical settings, and larger multi-procedure pretraining.

AI USE STATEMENT

Generative AI tools assisted with software debugging, code editing, and improving the manuscript's grammar and clarity. The authors identified relevant prior work independently and used AI as a secondary aid to check citation details and related-work summaries against the original publications. AI also assisted in refining descriptions of the analyses. All methodological and experimental decisions, interpretations, and scientific conclusions were made and verified by the authors, who take full responsibility for the final content.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Piyush Nitin Bagad and Andrew Zisserman. Chirality in action: Time-aware video representation learning by latent straightening. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Dominik Batić, Felix Holm, Ege Özsoy, Tobias Czempiel, and Nassir Navab. Whether and when does endoscopy domain pretraining make sense? *arXiv preprint arXiv:2303.17636*, 2023.

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.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Matthias Carstens, Shubha Vasisht, Zheyuan Zhang, Iulia Barbur, Annika Reinke, Lena Maier-Hein, Daniel A. Hashimoto, and Fiona R. Kolbinger. Artificial intelligence for surgical scene understanding: a systematic review and reporting quality meta-analysis. *npj Digital Medicine*, 9:59, 2026. doi: 10.1038/s41746-025-02227-4.

Ludovica Cella, Jasmine Lin, Mitchell G. Goldenberg, and Andrew J. Hung. The future of robotic surgery in the age of artificial intelligence. *Nature Reviews Urology*, 2026. doi: 10.1038/s41585-026-01153-8.

Chengan Che, Chao Wang, Jiayuan Huang, Xinyue Chen, and Luis C. Garcia-Peraza-Herrera. Can LLM-generated text empower surgical vision-language pre-training? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 9099–9106, June 2026a.

Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. LEMON: A large endoscopic monocular dataset and foundation model for perception in surgical settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42659–42669, 2026b.

Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11509–11522, 2023. doi: 10.1109/ICRA48891.2023.10161534. URL https://arxiv.org/abs/2209.09874.

Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8386–8402. PMLR, 2025.

Yiliang Chen, Zhixi Li, Cheng Xu, Alex Qinyang Liu, Ruize Cui, Xuemiao Xu, Jeremy Yuen-Chun Teoh, Shengfeng He, and Jing Qin. ProstaTD: Bridging surgical triplet from classification to fully supervised detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0NkXZ98BjJ.

Jiajun Cheng, Xiaofan Yu, Subarna Tripathi, Sainan Liu, and Shan Lin. TrajPred: Trajectory-conditioned joint embedding prediction for surgical instrument-tissue interaction recognition in vision-language models. *arXiv preprint arXiv:2603.06999*, 2026a.

Jiajun Cheng, Xianwu Zhao, Sainan Liu, Xiaofan Yu, Ravi Prakash, Patrick J Codd, Jonathan Elliott Katz, and Shan Lin. SurgXBench: Explainable vision-language model benchmark for surgery. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8188–8198, 2026b.

Hyungyu Choi, Young Kyun Jang, and Chanho Eom. GOAL: Global-local object alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4070–4079, 2025.

Zichen J Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. DynaMo: In-domain dynamics pretraining for visuo-motor control. *Advances in Neural Information Processing Systems*, 37:33933–33961, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations (ICLR)*, 2024.

Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9716–9726, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Ashley D. Edwards, Himanshu Sahni, Yannick Schroecker, and Charles L. Isbell. Imitating latent policies from observation. In *International Conference on Machine Learning (ICML)*, 2019.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3D awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21795–21806, 2024.

Shuangchun Gui and Zhenkun Wang. Tail-enhanced representation learning for surgical triplet recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 689–699. Springer, 2024.

Kazuki Honda, Suguru Oka, Kazuhide Makiyama, Tomoyuki Tatenuma, Munenori Fukunishi, Nao Kobayashi, Makoto Tanaka, Eri Fukagawa, Michikata Hayashida, Shinji Ito, Kazushige Sakaguchi, and Shinji Urakami. Artificial intelligence-based recognition of the prostatic capsule during nerve-sparing robot-assisted radical prostatectomy. *International Journal of Urology*, 33 (4):e70454, 2026. doi: 10.1111/iju.70454.

Taiyo Ikeido, Ren Togo, Takahiro Ogawa, Taku Sugiyama, Saseem Poudel, Hiroyuki Sugimori, Minghui Tang, Feng Han, Hidenori Koyano, Kenji Hirata, Kohsuke Kudo, and Miki Haseyama. Surgical video understanding with alignment-preserving temporal adaptation and action triplet text alignment. *Bioengineering*, 13(6):640, 2026. doi: 10.3390/bioengineering13060640.

Muhammad Abdullah Jamal and Omid Mohareri. SurgMAE: Masked autoencoders for long surgical video analysis. *arXiv preprint arXiv:2305.11451*, 2023.

Tim J.M. Jaspers, Ronald L.P.D. de Jong, Yiping Li, Carolus H.J. Kusters, Franciscus H.A. Bakker, Romy C. van Jaarsveld, Gino M. Kuiper, Richard van Hillegersberg, Jelle P. Ruurda, Willem M. Brinkman, Josien P.W. Pluim, Peter H.N. de With, Marcel Breeuwer, Yasmina Al Khalil, and Fons van der Sommen. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, 108:103873, 2026. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103873. URL https://www.sciencedirect.com/science/article/pii/S1361841525004190.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024.

Jiajie Li, Brian Quaranto, Chenhui Xu, Ishan Mishra, Ruiyang Qin, Dancheng Liu, Peter Kim, and Jinjun Xiong. Recognize any surgical object: unleashing the power of weakly-supervised data. In *International Conference on Learning Representations*, 2025.

Yuchong Li, Tong Xia, Huoling Luo, Baochun He, and Fucang Jia. Mt-fist: A multi-task fine-grained spatial-temporal framework for surgical action triplet recognition. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4983–4994, 2023. doi: 10.1109/JBHI.2023.3299321.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7575–7586, 2022.

Wenjun Lin, Yan Hu, Huazhu Fu, Mingming Yang, Chin-Boon Chng, Ryo Kawasaki, Cheekong Chui, and Jiang Liu. Instrument-tissue interaction detection framework for surgical video understanding. *IEEE Transactions on Medical Imaging*, 43(8):2803–2813, 2024. doi: 10.1109/TMI.2024.3381209.

Daochang Liu, Axel Hu, Mubarak Shah, and Chang Xu. Surgical triplet recognition via diffusion model. *arXiv preprint arXiv:2406.13210*, 2024.

Fei Long, Xiaoou Li, Jiaming Lv, Haoyuan Yang, Xianjun Cheng, and Peihua Li. BDC-CLIP: Brownian distance covariance for adapting CLIP to action recognition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40253–40269. PMLR, 2025.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021.

Amin Madani, Babak Namazi, Maria S. Altieri, Daniel A. Hashimoto, Angela Maria Rivera, Philip H. Pucher, Allison Navarrete-Welton, Ganesh Sankaranarayanan, L. Michael Brunt, Allan Okrainec, and Adnan Alseidi. Artificial intelligence for intraoperative guidance: Using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of Surgery*, 276(2):363–369, 2022. doi: 10.1097/SLA.0000000000004594.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Jecia Z. Y. Mao, Francis X. Creighton, Russell H. Taylor, and Manish Sahu. Speak, segment, track, navigate: An interactive system for video-guided skull-base surgery, 2026. URL https://arxiv.org/abs/2603.16024.

Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing CLIP with CLIP: Exploring pseudolabeling for limited-label prompt tuning. *Advances in Neural Information Processing Systems*, 36:60984–61007, 2023.

Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15579–15591, 2023.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Chinedu Innocent Nwoye and Nicolas Padoy. Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235*, 2022.

Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 364–374, 2020.

Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Baoqi Pei, Yifei Huang, Jilan Xu, Guo Chen, Yuping He, Lijin Yang, Yali Wang, Weidi Xie, Yu Qiao, Fei Wu, and Limin Wang. Modeling fine-grained hand-object dynamics for egocentric video representation learning. In *International Conference on Learning Representations*, 2025a.

Jialun Pei, Jiaan Zhang, Guanyi Qin, Kai Wang, Yueming Jin, and Pheng-Ann Heng. Instrument-tissue-guided surgical action triplet detection via textual-temporal trail exploration. *IEEE Transactions on Medical Imaging*, 44(12):5278–5289, 2025b. doi: 10.1109/TMI.2025.3590457.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Mingxing Rao, Yinhong Qin, Soheil Kolouri, Jie Ying Wu, and Daniel Moyer. Zero-shot prompt-based video encoder for surgical gesture recognition. *International Journal of Computer Assisted Radiology and Surgery*, 20(2):311–321, 2025.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations (ICLR)*, 2024.

Saurav Sharma, Chinedu Innocent Nwoye, Didier Mutter, and Nicolas Padoy. Rendezvous in time: An attention-based temporal fusion approach for surgical triplet recognition. *International Journal of Computer Assisted Radiology and Surgery*, 18(6):1043–1051, 2023a.

Saurav Sharma, Chinedu Innocent Nwoye, Didier Mutter, and Nicolas Padoy. Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 505–514. Springer, 2023b.

Saurav Sharma, Didier Mutter, and Nicolas Padoy. fine-CLIP: Enhancing zero-shot fine-grained surgical action recognition with vision-language models. *arXiv preprint arXiv:2503.19670*, 2025.

Ssharvien Kumar Sivakumar, Yannik Frisch, Amin Ranem, and Anirban Mukhopadhyay. SASVi: segment any surgical video. *International Journal of Computer Assisted Radiology and Surgery*, 20(7):1409–1419, 2025.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

Ani Vanyan, Alvard Barseghyan, Hakob Tamazyan, Vahan Huroyan, Hrant Khachatrian, and Martin Danelljan. Analyzing local representations of self-supervised vision transformers. *arXiv preprint arXiv:2401.00463*, 2023.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. EmbeddingGemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Khoa Vo, Thinh Phan, Kashu Yamazaki, Minh Tran, and Ngan Le. HENASY: Learning to assemble scene-entities for interpretable egocentric video-language model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 86483–86499, 2024.

Soham Walimbe, Britty Baby, Vinkle Srivastav, and Nicolas Padoy. Adaptation of multi-modal representation models for multi-task surgical computer vision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 24–33. Springer, 2025.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking self-attention for dense vision-language inference. In *European conference on computer vision*, pp. 315–332. Springer, 2024.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. DeCLIP: Decoupled learning for open-vocabulary dense perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14824–14834. IEEE, 2025a.

Mengmeng Wang, Zeyi Huang, Xiangjie Kong, Guojiang Shen, Guang Dai, Jingdong Wang, and Yong Liu. Action detail matters: Refining video recognition with local action queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19132–19142, 2025b.

Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 450–459, 2019.

Jinlin Wu, Felix Holm, Chuxi Chen, An Wang, Yaxin Hu, Xiaofan Ye, Zelin Zang, Miao Xu, Lihua Zhou, Huai Liao, et al. UniSurg: A video-native foundation model for universal understanding of surgical videos. *arXiv preprint arXiv:2602.05638*, 2026.

Nan Xi, Jingjing Meng, and Junsong Yuan. Chain-of-look prompting for verb-centric surgical triplet recognition in endoscopic videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5007–5016, 2023. doi: 10.1145/3581783.3611898.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Michael Yip. The robot will see you now: Foundation models are the path forward for autonomous robotic surgery. *Science Robotics*, 10(104):eadt0684, 2025.

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*, pp. 306–316. Springer, 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.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Chuheng Zhang, Tim Pearce, Pushi Zhang, Kaixin Wang, Xiaoyu Chen, Wei Shen, Li Zhao, and Jiang Bian. What do latent action models actually learn? *Advances in Neural Information Processing Systems*, 2025.

Zhenghao Zhang, Yuanxiang Wang, Zhenyu Guan, Yujia Yang, Bingkang Shi, Tianyu Zong, Hongzhu Yi, Guoqing Chao, Xingchen Chen, Tiankun Yang, et al. Delta-JEPA: Learning action-sensitive world models via latent difference decoding. *arXiv preprint arXiv:2606.31232*, 2026.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6586–6597, 2023.

Fatimah Zohra, Chen Zhao, Hani Itani, and Bernard Ghanem. b-CLIP: Text-conditioned contrastive learning for multi-granular vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 680–689, 2026.