# CoStoDet-DDPM: Collaborative Training of Stochastic and Deterministic Models Improves Surgical Workflow Anticipation and Recognition

Kaixiang Yang[1], Xin Li[1], Qiang Li[*], Zhiwei Wang[*]

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

[1]: Co-first authors, [*]: Corresponding authods.

2025-10-23

# CONTENTS


1 Background

2 Methods

3 Experiments

4 Conclusion

# 1.1 Surgical Workflow Analysis

## Surgical Workflow Analysis (SWA)

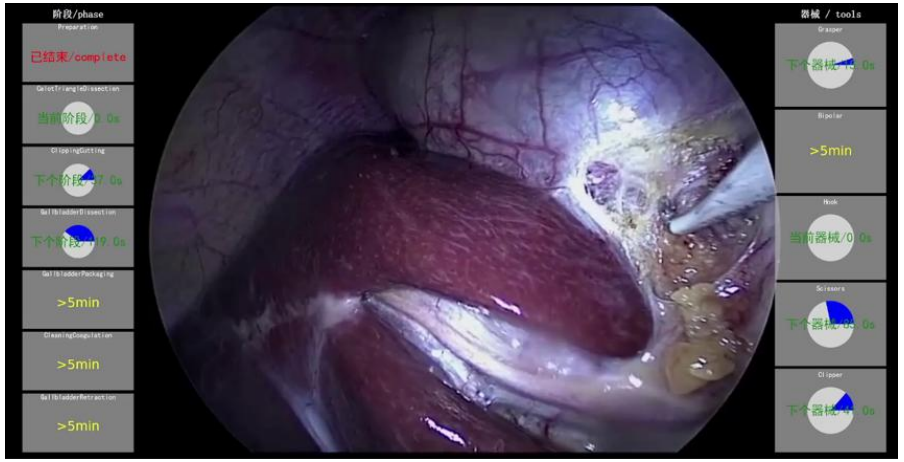
- Computational analysis of surgical videos to model phases, activities, and tool usage
- Enhances intraoperative coordination and supports surgical decision-making

## Anticipation

- Forecasting the **usage of surgical tools** several minutes ahead (*e.g.*, 5 minutes)
- Preparing tools in advance
- Ensure smoother surgical procedures

## Recognition

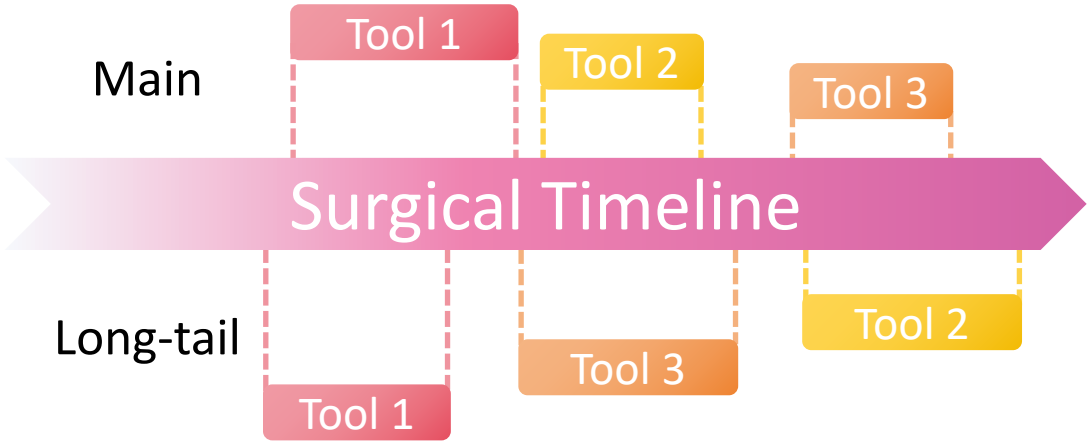
- Identifying the **current surgical phase** in real time
- Situational awareness in the operating room




Real-time surgical analysis

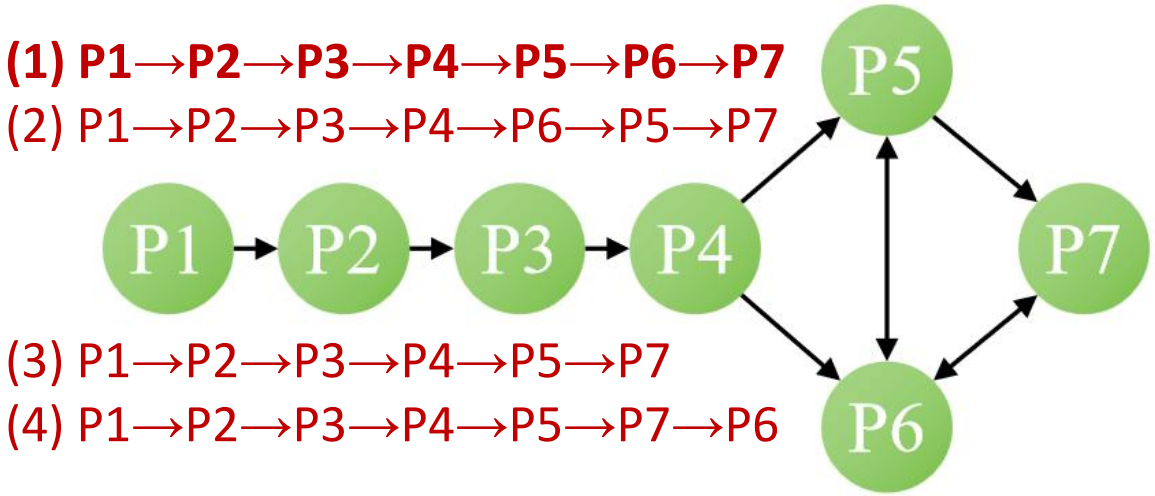
# 1.2 Difficulties and Challenges

➢ **Anticipation**: **long-tail** cases in tool types, usage frequencies, and timings




**(a) Long-tail Various Tools Usage**

➢ **Recognition**: **long-tail** variations in phase duration and order

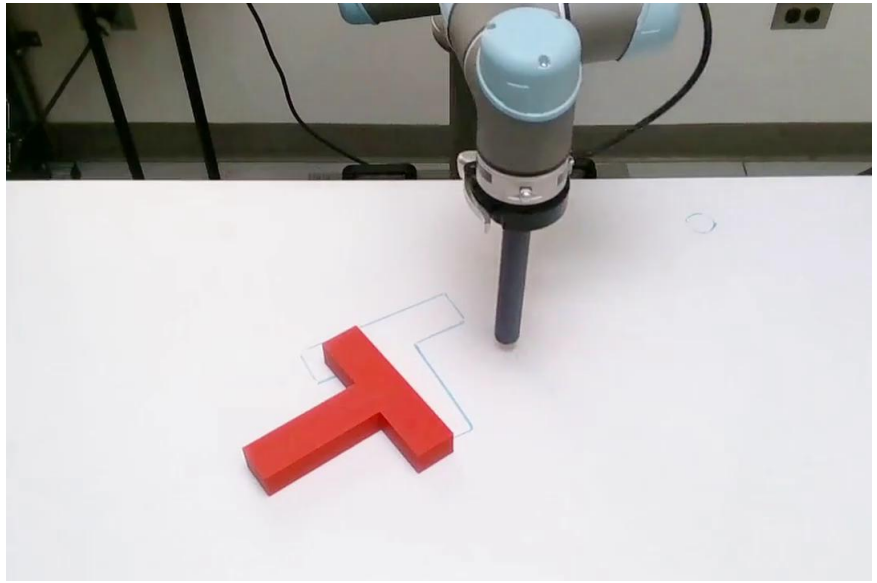



**(b) Long-tail Phase Order (2)-(4)**

# 1.3 Motivation

➢ **DiffusionPolicy[1] :** succeeded in robotic planning

- ✓ Handle **diverse scenarios**
- ✓ Robust under **human perturbations**



Could such ***stochastic*** paradigm benefit SWA?

- × **Uncontrollable process**: Surgery requires reliable outcomes, controllable and clinically interpretable process.
- × **Slow inference speed**: Impractical for real-time surgical assistance

Cannot be directly used in SWA!



[1] Chi C, Xu Z, Feng S, et al. Diffusion policy: Visuomotor policy learning via action diffusion[J]. The International Journal of Robotics Research, 2023.

CONTENTS

1 Background

2 Methods

3 Experiments

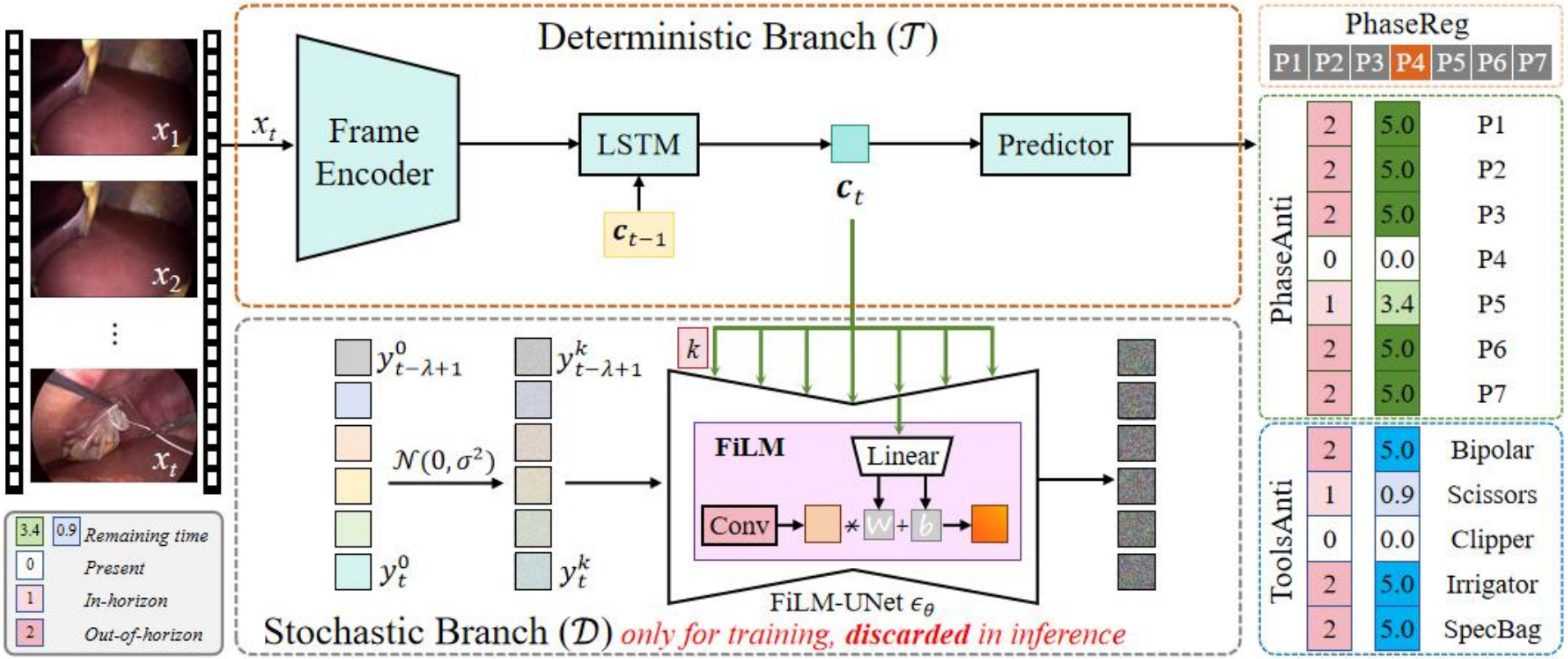
4 Conclusion

# 2 Framework

## CoStoDet-DDPM

CoStoDet-DDPM consists of a **Deterministic Task Branch** ($\mathcal{T}$) and a **Stochastic DDPM Branch** ($\mathcal{D}$).

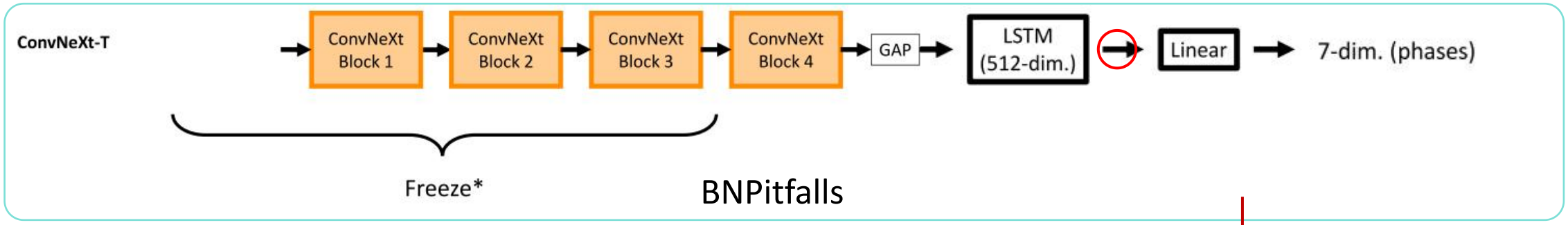
✓ **Co-training:** Short-long-term collaboration between $\mathcal{T}$ and $\mathcal{D}$ branches. $\mathcal{T}$ focuses on regular instantaneous mapping, $\mathcal{D}$ models a longer window to enhance temporal coherence.

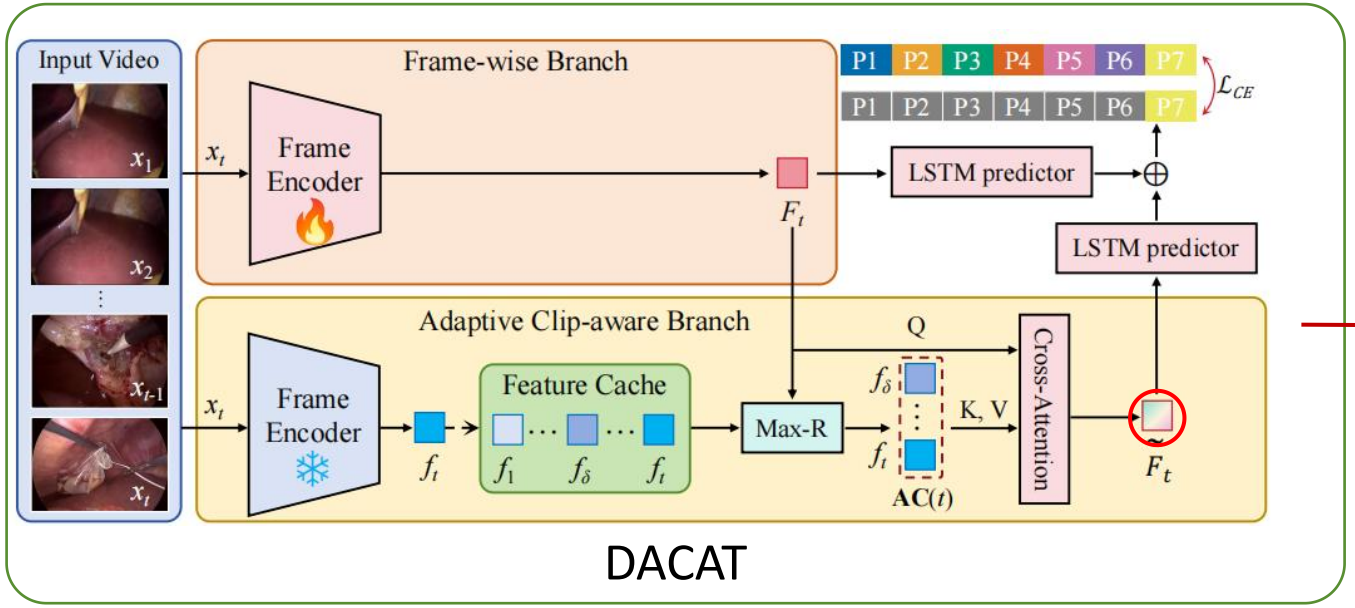


# 2 Framework

## CoStoDet-DDPM

- **Deterministic Task Branch ($\mathcal{T}$)**: BNPitfalls[1], DACAT[2]




Condition

Condition

**Stochastic DDPM Branch**

[1] Rivoir D, Funke I, Speidel S. On the pitfalls of batch normalization for end-to-end video learning: a study on surgical workflow analysis[J]. MIA, 2024.
[2] Yang K, Li Q, Wang Z. Dacat: Dual-stream adaptive clip-aware time modeling for robust online surgical phase recognition[C]. ICASSP, 2025.

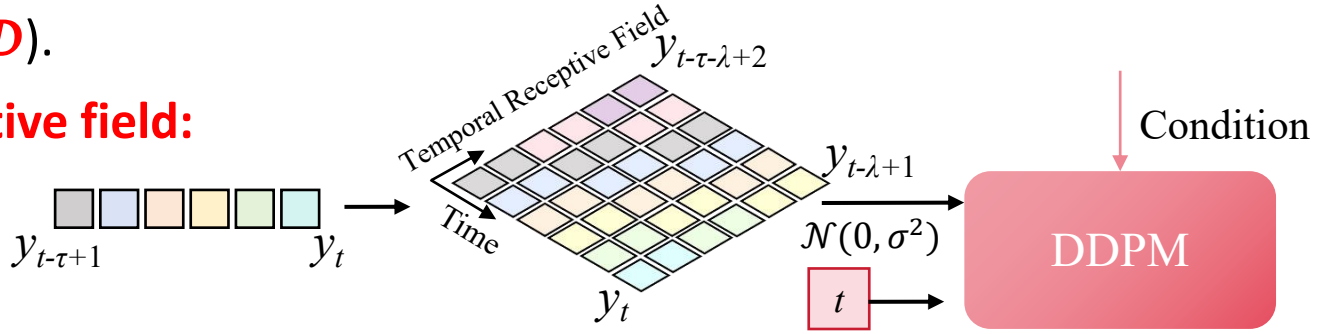
# 2 Framework

## CoStoDet-DDPM

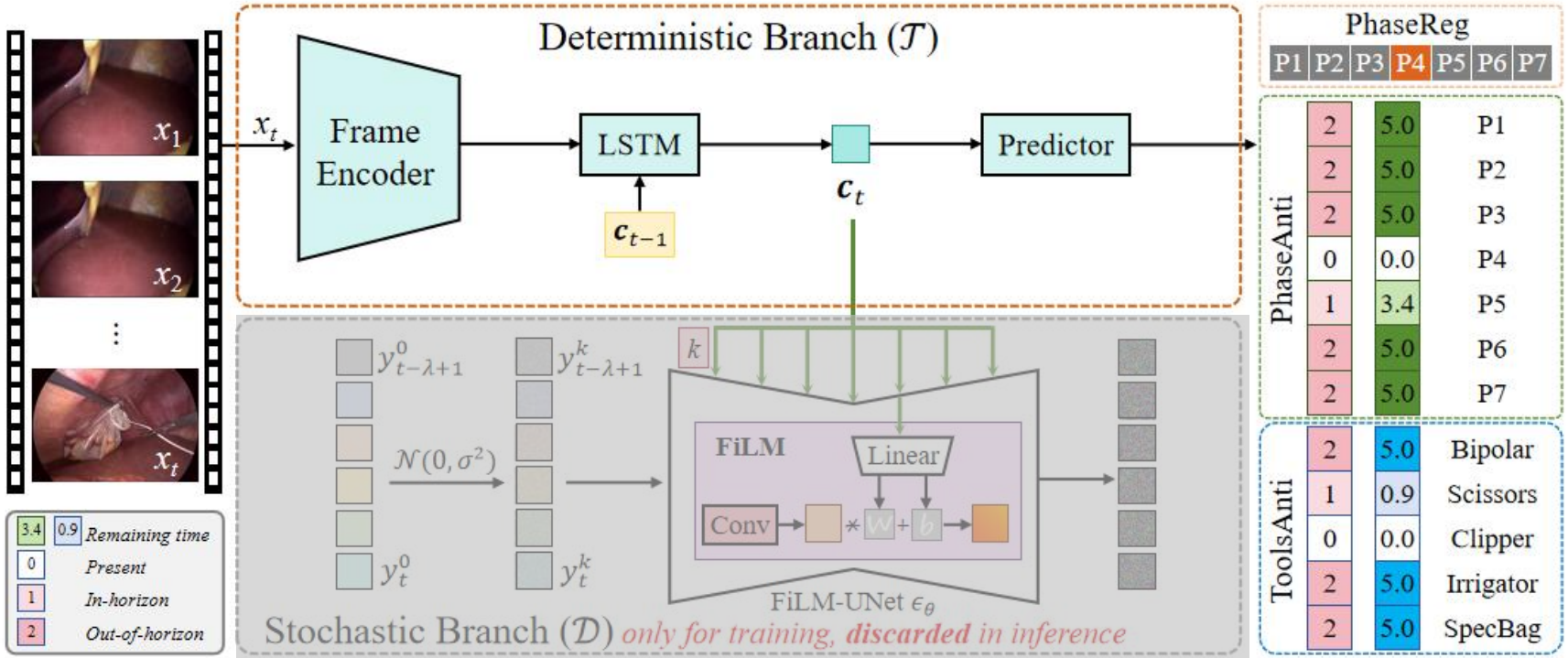
### ➢ Stochastic DDPM Branch ($\mathcal{D}$).

✓ **Longer temporal receptive field:**

$\tau$: Training clip length, set to 64
$\lambda$: Temporal receptive field, set to 32




✓ **Real-time inference: Test-time DDPM-discardable**



CONTENTS

1 Background

2 Methods

3 Experiments

4 Conclusion

# 3.1 Datasets

## Dataset

1. 2 surgical video dataset;  2 laparoscopic surgical procedures
2. lasts from **more than ten *mins*** to **several *hours***

| Dataset | Type | Phase | Train (Anticipation) | Validation | Test (Anticipation) |
|---|---|---|---|---|---|
| Cholec80[1] | Cholecystectomy | 7 | 40 (60) | / | 40 (20) |
| AutoLaparo[2] | Hysterectomy | 7 | 10 | 4 | 7 |

## Metrics

- Anticipation

  eMAE, outMAE, wMAE

- Recognition

  *Video*-level: Accuracy

  *Phase*-level: Precision, Recall, Jaccard

[1] Twinanda *et al*. Endonet: a deep architecture for recognition tasks on laparoscopic videos. TMI, 2016.
[2] Wang *et al*. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. MICCAI, 2022.

# 3.2 Comparison with SOTA

➢ 1. Comparison for anticipation on Cholec80

| Methods | $wMAE \downarrow / outMAE \downarrow / eMAE \downarrow$ | | | | | |
|---|---|---|---|---|---|---|
| | Tool | | | Phase | | |
| | $h = 2$ min | $h = 3$ min | $h = 5$ min | $h = 2$ min | $h = 3$ min | $h = 5$ min |
| TimeLSTM [2] | 0.51/0.16/1.56 | 0.76/0.29/2.26 | 1.32/0.67/3.62 | 0.47/0.22/1.29 | 0.64/0.32/1.73 | 1.07/0.60/2.60 |
| RSDNet [55] | 0.48/0.23/1.25 | 0.73/0.42/1.80 | 1.26/0.90/2.61 | 0.43/0.20/1.15 | 0.63/0.33/1.55 | 1.10/0.65/2.40 |
| Bayesian [47] | 0.43/0.08/1.12 | 0.66/0.15/1.65 | 1.09/0.44/2.68 | 0.39/0.15/1.02 | 0.59/0.32/1.47 | 0.85/0.52/1.54 |
| IIA-Net [66] | 0.38/0.10/1.01 | 0.58/0.19/1.46 | 0.92/0.44/2.14 | 0.36/0.10/1.18 | 0.49/0.18/1.42 | 0.68/0.28/1.09 |
| GCN-MSTCN [68] | 0.48/0.33/0.87 | 0.72/0.52/1.26 | 1.21/0.99/1.90 | 0.45/0.29/0.77 | 0.67/0.47/1.06 | 1.06/0.85/1.46 |
| ADANet [69] | 0.39/0.07/0.99 | 0.57/0.19/1.21 | 1.03/0.73/1.49 | 0.35/0.07/0.95 | 0.47/0.17/1.03 | 0.80/0.55/1.06 |
| BNPitfalls [48] | 0.40/0.06/1.06 | 0.58/0.13/1.42 | 0.96/0.41/2.11 | 0.29/0.06/0.63 | **0.39**/0.10/0.78 | 0.61/0.24/0.99 |
| Ours | **0.39/0.05/0.93** | **0.56/0.09/1.28** | **0.91/0.29/1.77** | **0.27/0.05/0.52** | **0.39/0.09/0.67** | **0.59/0.18/0.88** |

h: anticipation horizon

✦ The results show a significant improvement, particularly at $h$ = 5min.

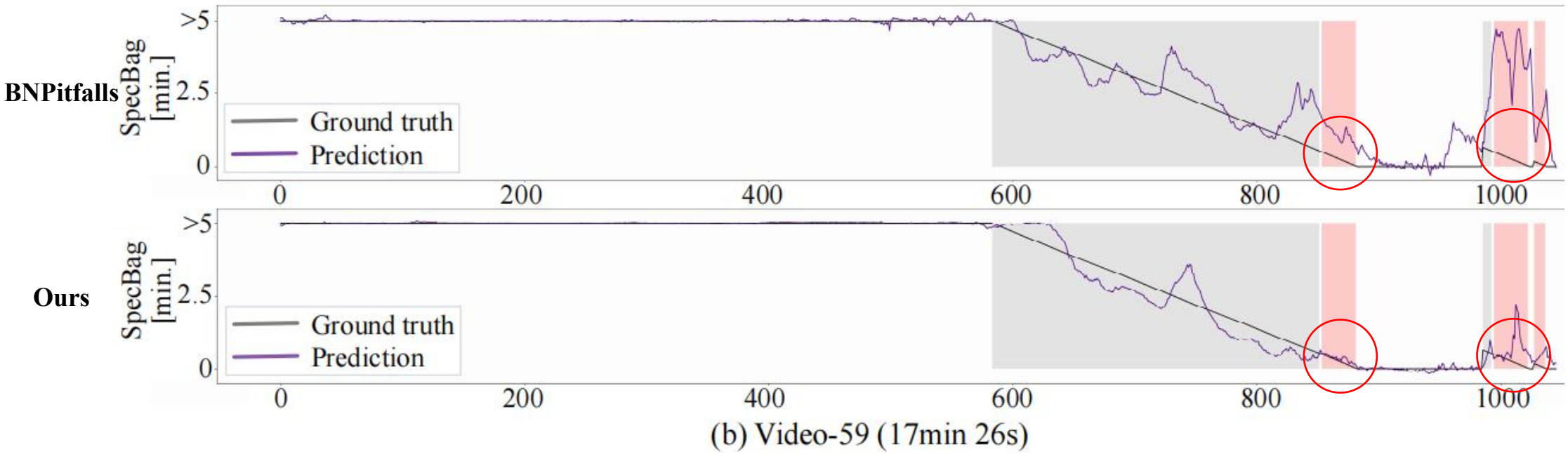
# 3.2 Comparison with SOTA

➢ 2. Comparison for recognition on Cholec80 and AutoLaparo

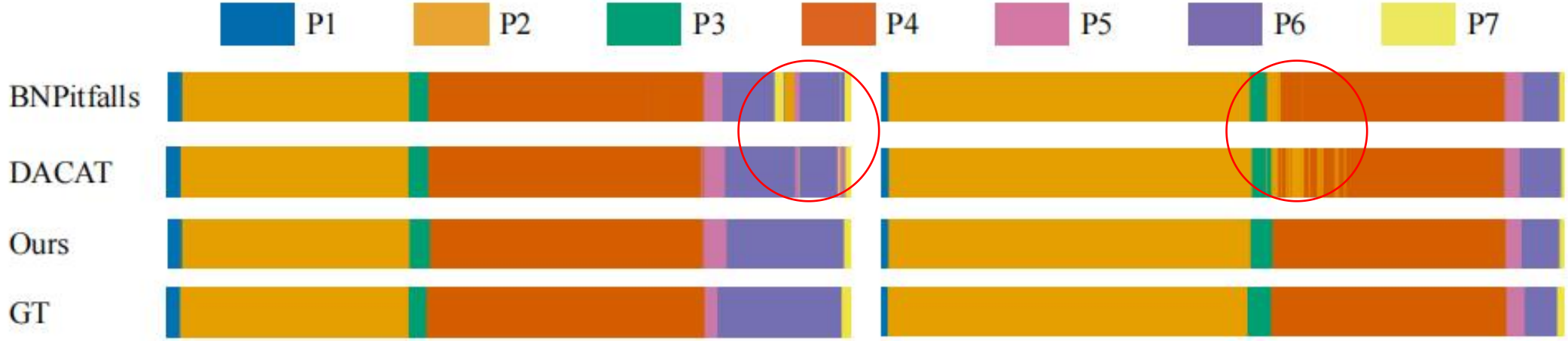
| Datasets | Methods | Accuracy ↑ | Precision ↑ | Recall ↑ | Jaccard ↑ |
|---|---|---|---|---|---|
| Cholec80 | Trans-SVNet [14] | 89.1 ± 7.0 | 84.7 | 83.6 | 72.5 |
| | AVT [16] | 86.7 ± 7.6 | 77.3 | 82.1 | 66.4 |
| | TeSTra [70] | 90.1 ± 6.6 | 82.8 | 83.8 | 71.6 |
| | LoViT [35] | 91.5 ± 6.1 | 83.1 | 86.5 | 74.2 |
| | SKiT [34] | 92.5 ± 5.1 | 84.6 | 88.5 | 76.7 |
| | Surgformer [62] | 92.4 ± 6.4 | 87.9 | 89.3 | 79.9 |
| | BNPitfalls [48] | 93.7 ± 4.1 | 88.7 | 87.7 | 79.2 |
| | DACAT [61] | 94.1 ± 4.3 | **89.2** | 88.5 | 80.6 |
| | Ours | **94.5 ± 3.6** | 88.4 | **90.2** | **81.4** |
| AutoLaparo | SV-RCNet [24] | 75.6 | 64.0 | 59.7 | 47.2 |
| | TMRNet [26] | 78.2 | 66.0 | 61.5 | 49.6 |
| | Trans-SVNet [14] | 78.3 | 64.2 | 62.1 | 50.7 |
| | LoViT [35] | 81.4 ± 7.6 | **85.1** | 65.9 | 55.9 |
| | SKiT [34] | 82.9 ± 6.8 | 81.8 | 70.1 | 59.9 |
| | SPHASE [37] | 83.8 | 75.7 | 71.3 | 57.8 |
| | Surgformer [62] | 85.7 ± 6.9 | 82.3 | 75.7 | 66.7 |
| | BNPitfalls [48] | 86.8 ± 1.5 | 78.2 | 72.0 | 64.2 |
| | DACAT [61] | **87.8 ± 7.6** | 78.5 | 75.0 | 66.9 |
| | Ours | 87.6 ± 7.8 | 84.5 | **76.4** | **67.9** |

✦ Our method achieves relative improvements of 1.0% and 1.5% in Jaccard index on the two datasets, respectively.

# 3.3 Visualization for anticipation and recognition




(b) Video-59 (17min 26s)




(a) Video-53 (54min 43s)

(b) Video-62 (33min 52s)

CONTENTS

1 Background

2 Methods

3 Experiments

4 Conclusion

# 4 Conclusion

- We propose a novel dual-branch collaborative training framework that explores the simultaneous training of stochastic and deterministic models to enhance their individual performance.

- We introduce a new functionality of DDPM, i.e., feature enhancement, and applie DDPM for the first time in surgical workflow analysis. By effectively reducing prediction errors in uncertain scenarios and adapting to patient-specific anatomical and procedural variations, our approach addresses key challenge in clinical reliability.

- Our method achieves state-of-the-art (SOTA) results in both anticipation task (Cholec80) and recognition task (Cholec80, AutoLaparo). The inference speed is 91 FPS, fully meeting real-time requirements.

# Thank you

CoStoDet-DDPM: Collaborative Training of Stochastic and Deterministic Models Improves Surgical Workflow Anticipation and Recognition

Kaixiang Yang[1], Xin Li[1], Qiang Li*, Zhiwei Wang*

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

2025-10-23