

---

# Multi-scale Temporal Prediction via Incremental Generation and Multi-agent Collaboration

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Integrated IG-MC Framework

The complete IG-MC pipeline operates on certain task ensembles  $Q$ , where each task  $q \in Q$  represents a distinct procedure. Our framework features a decoupled architecture where the DM module and SD module undergo separate training phases. This design enables flexible combination during inference while maintaining modularity. For each sampled task  $q$ , the DM module generates predicted state trajectories  $\{\mathcal{S}_k\}_{k=1}^N$  through iterative application of the decision-making function:

$$\mathcal{S}_{k+1} = \text{DM}(\mathcal{S}_k, \mathcal{I}_k; \theta_{\text{DM}}), \quad (1)$$

where  $\theta_{\text{DM}}$  denotes the trainable parameters of the DM module,  $\mathcal{S}_k$  represents the predicted state at time step  $t_k$ , and  $\mathcal{I}_k$  is the visual guidance synthesized up to  $t_k$ . Concurrently, the SD module produces the visual sequence  $\{\mathcal{I}_k\}_{k=1}^N$  through a conditioned diffusion process:

$$\mathcal{I}_{k+1} = \text{SD}(\mathcal{S}_{k+1}, \mathcal{I}_k; \theta_{\text{SD}}), \quad (2)$$

where  $\theta_{\text{SD}}$  parameterizes the SD module, and  $\mathcal{I}_{k+1}$  is synthesized by denoising a latent representation conditioned on both the predicted state  $\mathcal{S}_{k+1}$  and previous visual guidance  $\mathcal{I}_k$ . The temporal resolution is determined by  $N = \lceil T/\tau \rceil$  time steps, with  $T$  being the total procedure duration and  $\tau$  the incremental time interval.

The learning objective maximizes the temporal average accuracy of state predictions relative to ground-truth annotations:

$$\mathcal{L}_{\text{IG-MC}} = \max_{\theta_{\text{DM}}, \theta_{\text{SD}}} \mathbb{E}_{q \sim Q} \left[ \frac{1}{N} \sum_{k=1}^N P(\mathcal{S}_k = \hat{\mathcal{S}}_k) \mathbb{I}\left(\frac{k}{\hat{\tau}} \in \mathbb{Z}^+\right) \right], \quad (3)$$

where  $\hat{\mathcal{S}}_k$  is the ground-truth state at  $t_k$ ,  $\hat{\tau}$  represents the temporal resolution of ground-truth annotations, and  $\mathbb{I}(\cdot)$  is an indicator function enforcing temporal alignment. The expectation  $\mathbb{E}_{q \sim Q}$  is approximated via Monte Carlo sampling over the task distribution  $Q$ . The probability term  $P(\mathcal{S}_k = \hat{\mathcal{S}}_k)$  derives from a cross-entropy loss between predicted and true state distributions, ensuring differentiability throughout the optimization process.

The term  $P(\mathcal{S}_k = \hat{\mathcal{S}}_k)$  represents the probability that the predicted state  $\mathcal{S}_k$  matches the ground-truth state  $\hat{\mathcal{S}}_k$  at time step  $t_k$ . This probability serves as a direct measure of prediction accuracy, where higher values indicate better alignment between predicted and actual states.

However, not every incremental time step requires state prediction. The framework operates with two distinct temporal resolutions: the *incremental scale*  $\tau$  (e.g., 5s) for internal state updates and the *temporal scale*  $\hat{\tau}$  (e.g., 30s) for meaningful prediction outputs. The indicator function  $\mathbb{I}(\frac{k}{\hat{\tau}} \in \mathbb{Z}^+)$

enforces this distinction by evaluating to 1 only when the current step index  $k$  corresponds to an integer multiple of the prediction interval ratio  $\frac{\hat{\tau}}{\tau}$ .

Mathematically, this condition:

$$\frac{k}{\hat{\tau}} \in \mathbb{Z}^+ \quad (4)$$

ensures that state predictions are generated precisely at the coarser temporal time steps (every  $\hat{\tau}$  seconds), while allowing continuous internal updates at finer incremental intervals. For the example where  $\tau = 5s$  and  $\hat{\tau} = 30s$ , predictions would occur at every 6th incremental step (since  $30/5 = 6$ ), maintaining computational efficiency without sacrificing temporal granularity where needed.

## 2 More Experiments

As shown in Table 1, Table 2 and Table 3, we show comparisons of different VLMs with and without the plug-and-play DM module on the MSTP-Surgery Benchmark. The tables display results across various temporal and incremental scales, using metrics like Accuracy, Precision, etc. In each case, adding the DM module boosts model performance significantly, demonstrating its effectiveness in multi-scale temporal surgical prediction tasks.

Model	Temp. Scale	State Scale	Accuracy	Precision	Recall	F1	Jaccard
InternVL3-8B	1	Phase	18.30	12.09	10.84	10.44	5.79
+ DM	1	Phase	57.20 (+38.90)	68.55 (+56.46)	63.56 (+52.72)	65.52 (+55.08)	50.65 (+44.86)
InternVL3-8B	1	Step	17.50	8.83	7.06	6.84	3.74
+ DM	1	Step	56.90 (+39.40)	40.45 (+31.62)	40.46 (+33.40)	39.86 (+33.02)	28.35 (+24.61)
InternVL3-8B	1	Phase&Step	13.60	3.61	3.42	2.91	1.58
+ DM	1	Phase&Step	36.20 (+22.60)	23.22 (+19.61)	17.18 (+13.76)	18.80 (+15.89)	13.22 (+11.64)
InternVL3-8B	5	Phase	20.80	14.25	14.13	12.98	7.47
+ DM	5	Phase	57.20 (+36.40)	68.44 (+54.19)	66.42 (+52.29)	66.85 (+53.87)	52.25 (+44.78)
InternVL3-8B	5	Step	18.60	8.20	9.17	8.08	4.47
+ DM	5	Step	60.60 (+42.00)	49.90 (+41.70)	48.47 (+39.30)	48.45 (+40.37)	36.77 (+32.30)
InternVL3-8B	5	Phase&Step	13.80	3.48	3.93	3.25	1.78
+ DM	5	Phase&Step	37.30 (+23.50)	25.60 (+22.12)	19.62 (+15.69)	21.25 (+17.99)	15.75 (+13.97)
InternVL3-8B	30	Phase	21.10	13.33	14.23	12.89	7.31
+ DM	30	Phase	56.60 (+35.50)	65.31 (+51.98)	63.23 (+49.00)	63.85 (+50.96)	49.32 (+42.01)
InternVL3-8B	30	Step	18.40	4.57	6.18	5.04	2.84
+ DM	30	Step	60.30 (+41.90)	41.20 (+36.63)	41.64 (+35.46)	40.76 (+35.72)	29.93 (+27.09)
InternVL3-8B	30	Phase&Step	14.40	2.11	3.69	2.34	1.30
+ DM	30	Phase&Step	42.30 (+27.90)	20.99 (+18.88)	18.04 (+14.35)	18.83 (+16.49)	13.83 (+12.53)
InternVL3-8B	60	Phase	26.50	20.40	18.87	18.54	10.80
+ DM	60	Phase	50.80 (+24.30)	58.56 (+38.16)	56.96 (+38.09)	57.10 (+38.56)	42.24 (+31.44)
InternVL3-8B	60	Step	22.30	10.93	10.65	9.41	5.29
+ DM	60	Step	63.00 (+40.70)	37.56 (+26.63)	38.68 (+28.03)	37.24 (+27.83)	28.03 (+22.74)
InternVL3-8B	60	Phase&Step	16.70	6.01	5.26	4.73	2.61
+ DM	60	Phase&Step	36.30 (+19.60)	19.29 (+13.28)	14.92 (+9.66)	15.99 (+11.26)	11.28 (+8.67)

Table 1: Comparison of InternVL3-8B with and without plug-and-play DM module on MSTP-Surgery Benchmark.

Model	Temp. Scale	State Scale	Accuracy	Precision	Recall	F1	Jaccard
Gemma-3-4B-IT	1	Phase	15.50	6.31	3.73	3.26	1.78
+ DM	1	Phase	57.20 (+41.70)	68.55 (+62.24)	63.56 (+59.83)	65.52 (+62.26)	50.65 (+48.87)
Gemma-3-4B-IT	1	Step	17.50	8.83	7.06	6.84	3.74
+ DM	1	Step	56.90 (+39.40)	40.45 (+31.62)	40.46 (+33.40)	39.86 (+33.02)	28.35 (+24.61)
Gemma-3-4B-IT	1	Phase&Step	13.60	3.61	3.42	2.91	1.58
+ DM	1	Phase&Step	36.20 (+22.60)	23.22 (+19.61)	17.18 (+13.76)	18.80 (+15.89)	13.22 (+11.64)
Gemma-3-4B-IT	5	Phase	20.80	14.25	14.13	12.98	7.47
+ DM	5	Phase	57.20 (+36.40)	68.44 (+54.19)	66.42 (+52.29)	66.85 (+53.87)	52.25 (+44.78)
Gemma-3-4B-IT	5	Step	18.60	8.20	9.17	8.08	4.47
+ DM	5	Step	60.60 (+42.00)	49.90 (+41.70)	48.47 (+39.30)	48.45 (+40.37)	36.77 (+32.30)
Gemma-3-4B-IT	5	Phase&Step	13.80	3.48	3.93	3.25	1.78
+ DM	5	Phase&Step	37.30 (+23.50)	25.60 (+22.12)	19.62 (+15.69)	21.25 (+17.99)	15.75 (+13.97)
Gemma-3-4B-IT	30	Phase	21.10	13.33	14.23	12.89	7.31
+ DM	30	Phase	56.60 (+35.50)	65.31 (+51.98)	63.23 (+49.00)	63.85 (+50.96)	49.32 (+42.01)
Gemma-3-4B-IT	30	Step	18.40	4.57	6.18	5.04	2.84
+ DM	30	Step	60.30 (+41.90)	41.20 (+36.63)	41.64 (+35.46)	40.76 (+35.72)	29.93 (+27.09)
Gemma-3-4B-IT	30	Phase&Step	14.40	2.11	3.69	2.34	1.30
+ DM	30	Phase&Step	42.30 (+27.90)	20.99 (+18.88)	18.04 (+14.35)	18.83 (+16.49)	13.83 (+12.53)
Gemma-3-4B-IT	60	Phase	26.50	20.40	18.87	18.54	10.80
+ DM	60	Phase	50.80 (+24.30)	58.56 (+38.16)	56.96 (+38.09)	57.10 (+38.56)	42.24 (+31.44)
Gemma-3-4B-IT	60	Step	22.30	10.93	10.65	9.41	5.29
+ DM	60	Step	63.00 (+40.70)	37.56 (+26.63)	38.68 (+28.03)	37.24 (+27.83)	28.03 (+22.74)
Gemma-3-4B-IT	60	Phase&Step	16.70	6.01	5.26	4.73	2.61
+ DM	60	Phase&Step	36.30 (+19.60)	19.29 (+13.28)	14.92 (+9.66)	15.99 (+11.26)	11.28 (+8.67)

Table 2: Comparison of Gemma-3-4B-IT with and without plug-and-play DM module on MSTP-Surgery Benchmark.

Model	Temp. Scale	State Scale	Accuracy	Precision	Recall	F1	Jaccard
Qwen2.5-VL-7B	1	Phase	20.50	13.17	14.41	13.04	7.42
+ DM	1	Phase	73.10 (+52.60)	47.74 (+34.57)	44.98 (+30.57)	46.27 (+33.23)	37.99 (+30.57)
Qwen2.5-VL-7B	1	Step	20.30	8.27	9.23	8.35	4.71
+ DM	1	Step	50.60 (+30.30)	42.87 (+34.60)	39.90 (+30.67)	41.08 (+32.73)	28.68 (+23.97)
Qwen2.5-VL-7B	1	Phase&Step	14.00	3.61	4.29	3.58	1.97
+ DM	1	Phase&Step	49.90 (+35.90)	30.45 (+26.84)	26.79 (+22.50)	28.24 (+24.66)	19.67 (+17.70)
Qwen2.5-VL-7B	5	Phase	19.90	14.41	15.14	13.48	7.64
+ DM	5	Phase	81.50 (+61.60)	58.21 (+43.80)	54.80 (+39.66)	56.11 (+42.63)	48.77 (+41.13)
Qwen2.5-VL-7B	5	Step	16.40	7.90	7.94	7.20	3.96
+ DM	5	Step	52.00 (+35.60)	45.08 (+37.18)	39.72 (+31.78)	40.85 (+33.65)	28.54 (+24.58)
Qwen2.5-VL-7B	5	Phase&Step	12.00	3.32	3.69	2.90	1.57
+ DM	5	Phase&Step	50.80 (+38.80)	32.51 (+29.19)	28.83 (+25.14)	29.03 (+26.13)	20.06 (+18.49)
Qwen2.5-VL-7B	30	Phase	17.80	14.88	12.65	12.29	6.86
+ DM	30	Phase	64.40 (+46.60)	39.54 (+24.66)	31.66 (+19.01)	34.65 (+22.36)	27.03 (+20.17)
Qwen2.5-VL-7B	30	Step	18.20	7.62	7.99	7.09	3.93
+ DM	30	Step	50.90 (+32.70)	38.91 (+31.29)	36.09 (+28.10)	36.24 (+29.15)	25.27 (+21.34)
Qwen2.5-VL-7B	30	Phase&Step	13.50	3.92	4.02	3.51	1.92
+ DM	30	Phase&Step	45.90 (+32.40)	17.49 (+13.57)	13.81 (+9.79)	14.48 (+10.97)	9.94 (+8.02)
Qwen2.5-VL-7B	60	Phase	20.30	16.05	15.22	14.47	8.04
+ DM	60	Phase	70.80 (+50.50)	40.67 (+24.62)	35.45 (+20.23)	37.67 (+23.20)	30.47 (+22.43)
Qwen2.5-VL-7B	60	Step	18.40	7.64	8.25	7.23	4.00
+ DM	60	Step	54.50 (+36.10)	36.10 (+28.46)	34.61 (+26.36)	35.06 (+27.83)	24.19 (+20.19)
Qwen2.5-VL-7B	60	Phase&Step	13.20	4.48	4.71	3.83	2.07
+ DM	60	Phase&Step	51.90 (+38.70)	17.81 (+13.33)	16.09 (+11.38)	16.58 (+12.75)	11.44 (+9.37)

Table 3: Comparison of Qwen2.5-VL-7B-Instruct with and without plug-and-play DM module on MSTP-Surgery Benchmark.