

Latent Visual Diffusion Reasoning with Monte Carlo Tree Search

Xirui Teng¹, Nan Xi² , and Junsong Yuan³

¹ Beijing Jiaotong University, Beijing, China

² Virginia Commonwealth University, Richmond VA, USA

³ University at Buffalo, Buffalo NY, USA

 Corresponding author: xin@vcu.edu

Abstract. Analyzing fine-grained skill activities (e.g., sports, surgery) requires not only recognizing visual patterns but also performing step-by-step visual reasoning that leads to the final judgment. While recent advances in action quality assessment have achieved remarkable progress in evaluating performance, existing models remain black boxes, where they lack the ability to explicitly reveal the reasoning processes underlying their judgments. To address this limitation, we propose **Latent Visual Diffusion Reasoning (LVDR)**, a novel framework that integrates keypoint-guided Monte Carlo Tree Search (MCTS) to model and visualize the latent visual reasoning process. LVDR not only produces more accurate skill assessments but also uncovers the critical visual reasoning sequences that contribute to the final evaluation. Extensive experiments across four datasets spanning diverse sports and surgical domains demonstrate that LVDR achieves competitive quantitative performance while providing interpretable visual reasoning trajectories leading to the final predictions. Source codes and models can be found through the following link: https://github.com/XiruiTeng/LVDR_Official.git.

Keywords: Latent Visual Diffusion Reasoning · Monte Carlo Tree Search

1 Introduction

Assessing fine-grained skill activities such as athletic performances and surgical procedures poses a fundamental yet challenging problem in computer vision [1, 2, 10, 25, 30, 32]. Unlike conventional action recognition [21–24], which focuses solely on identifying what action is being performed, skill assessment requires understanding how well the action is executed. This subtle distinction introduces a higher level of complexity, as it demands not only precise recognition of visual patterns but also the ability to perform **step-by-step visual reasoning** that links low-level motion dynamics to high-level evaluative outcomes. Capturing this causal progression from subtle temporal cues and biomechanical coordination to overall performance quality is critical for achieving meaningful and interpretable assessments. This capability is particularly important for high-stakes applications requiring transparency and trust, including athlete training, surgical education, and autonomous robotic skill acquisition.

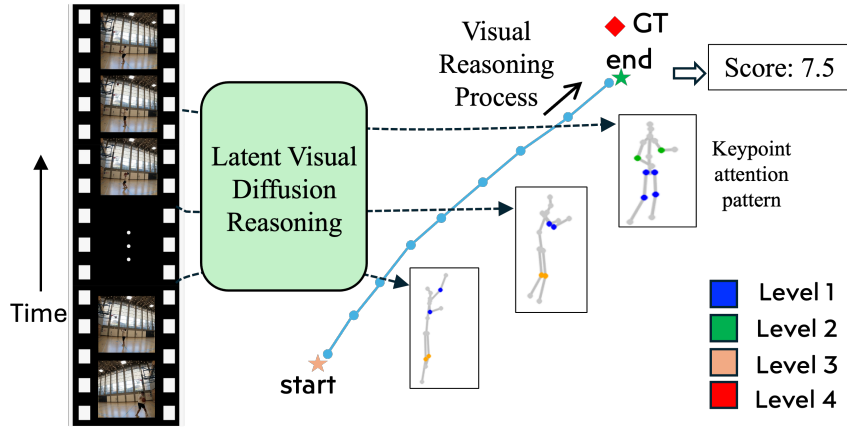


Fig. 1: Overview of Latent Visual Diffusion Reasoning. We formulate visual reasoning as a diffusion process in the latent reasoning space. Our latent visual diffusion reasoning model is able to generate both performance score for a given video and also the visual reasoning process with the evolving body keypoint attention pattern. Different colors on keypoints indicate different attention level at each time stamp.

Recent advances in action quality assessment (AQA) [4,12,28,33] have demonstrated remarkable progress by leveraging deep neural networks to either regress continuous skill scores or predict relative performance rankings. While these models achieve strong predictive accuracy, they largely function as black boxes, which are capable of producing final assessments without exposing the intermediate reasoning steps that lead to those judgments. Consequently, their decision making processes remain opaque, limiting interpretability and hindering the ability of human experts to validate or derive actionable insights from model outputs. This lack of transparency poses a significant barrier to deployment in domains such as surgery and elite sports, where understanding *why* a particular assessment is made is as important as the assessment itself. In such high-stakes settings, accountability, interpretability, and trust are indispensable for real-world adoption of AI-based assessment systems.

To address these limitations, we propose Latent Visual Diffusion Reasoning (LVDR), a novel framework that explicitly models the latent reasoning process underlying fine-grained skill assessment. As illustrated in Figure 1, LVDR not only predicts the overall performance score for a given action but also generates a visual reasoning trajectory that reveals how the model arrives at its judgment. Specifically, LVDR provides interpretable visualizations that capture the evolving keypoint attention patterns throughout the reasoning process. The framework integrates the representational power of diffusion-based generative modeling with the structured exploration capability of Monte Carlo Tree Search (MCTS).

We formulate video reasoning as a generative diffusion process within a latent space. LVDR treats reasoning as a progressive denoising process that evolves over

temporal states. This latent space is structured by a target semantic distribution derived from ground-truth video semantics (e.g., labels or expert commentary). LVDR learns to generate coherent latent trajectories that represent a gradual refinement of semantic understanding. This formulation enables the model to capture the incremental and temporal nature of complex reasoning by explicitly modeling the path from a non-informative prior to a target semantic state.

In reasoning space, the beginning of the video is treated as a noisy initialization of the reasoning process. As time progresses, the diffusion model gradually denoises the latent reasoning embeddings of the early video frames, transforming them into a target distribution that reflects coherent reasoning dynamics. Guided by keypoint-based MCTS, LVDR explores and identifies the most informative reasoning trajectories that contribute to the final evaluation. This design enables LVDR to make the model’s decision-making process both transparent and interpretable, bridging the gap between high predictive performance and human-understandable reasoning in fine-grained skill assessment.

To comprehensively evaluate the effectiveness of LVDR, we conduct experiments across four benchmark datasets spanning diverse skill domains, including multiple sports and surgical procedures. Experimental results demonstrate that LVDR achieves competitive or superior quantitative performance compared to state-of-the-art approaches, while uniquely offering interpretable visual reasoning trajectories that uncover key spatiotemporal patterns underlying skill execution. These results highlight LVDR’s ability not only to assess performance accurately but also to reveal the visual evidence and reasoning processes behind its judgments, paving the way toward more transparent and trustworthy skill assessment models.

In summary, our contributions include:

- We introduce Latent Visual Diffusion Reasoning (LVDR), a framework that models step-by-step visual reasoning as a progressive diffusion process. The diffusion process is performed within a latent reasoning space constructed from the essential semantics of the full video, enabling coherent and interpretable reasoning for fine-grained skill assessment.
- We propose a keypoint-guided Monte Carlo Tree Search mechanism to uncover and visualize latent visual reasoning sequences.
- We demonstrate that LVDR achieves strong performance across multiple sports and surgical datasets while offering unprecedented transparency in its decision-making process.

2 Related Work

2.1 Action Quality Assessment

Action Quality Assessment(AQA) aims to evaluate how well an action is performed. This task is challenging for fine-grained skill activities, which require models to move beyond predicting scores and perform step-by-step visual reasoning. While this field has seen progress in predicting scores accurately by

treating the task as a regression or ranking problem, most existing methods are mainly based on end-to-end deep learning methods and rarely reveal the processes which lead to their judgments. Recent efforts have attempted to address this problem. NS-AQA [14] utilizes neural networks to extract visual symbols from video data and rule-based models to analyze these symbols to generate detailed interpretations. IRIS [13] uses scoring rubrics to segment action sequences and compute technical scores and this can give the explanations of the evaluation process. While these works make the reasoning process transparent, they are built on human-defined rules and fail to learn from data. Interpretability-AQA [6] proposes a new attention loss function to avoid temporal skipping and a weight-score regression module to explain each video segment’s score by giving different weights. While this method can explain how much a video segment contributes to the final score, it still fails to model the sequential, step-by-step reasoning process that a human expert might follow. To address this gap, we introduce a new framework that allows us to model and visualize the latent visual reasoning process, uncovering the critical reasoning sequences that lead to a final judgment while preserving quantitative accuracy.

2.2 Monte Carlo Tree Search

Monte Carlo Tree Search (MCTS) is a planning algorithm that strategically balances exploration and exploitation by integrating tree search with stochastic simulations [5]. The algorithm iteratively refines a search tree through four key stages. First, the selection phase traverses the tree from the root node, guided by a policy like Upper Confidence Bounds for Trees (UCT) [11], to select a child node that either maximizes potential score or explores less visited states. Upon reaching a leaf node or an expandable internal node, the expansion phase adds one or more child nodes, representing previously unconsidered actions or plan extensions. Subsequently, the simulation (or rollout) phase estimates the value of the newly added node by performing random or policy-based sampling of actions until a terminal state or a predefined depth is reached. Finally, the backpropagation phase updates the value estimates of the nodes along the path from the newly simulated node back to the root, incorporating the simulation outcome. Through repeated iterations of these four phases, MCTS adaptively focuses its search on promising regions of the state space while maintaining exploration of less certain areas.

3 Method

For a given video $V = \{I_1, \dots, I_T\}$ that contains T video frames $\{I_i\}_{i=1}^T$, our LVDR framework $\mathcal{F}_\theta$ aims to generate: (1) predicted score $\hat{s}$; (2) visual reasoning process $\mathcal{H} = \{H_1, \dots, H_T\}$, where $\{H_i\}_{i=1}^T$ indicate keypoint attention patterns at each timestamp. LVDR includes two major components: (1) diffusion module to learn latent visual reasoning embeddings by modeling the denoising process of

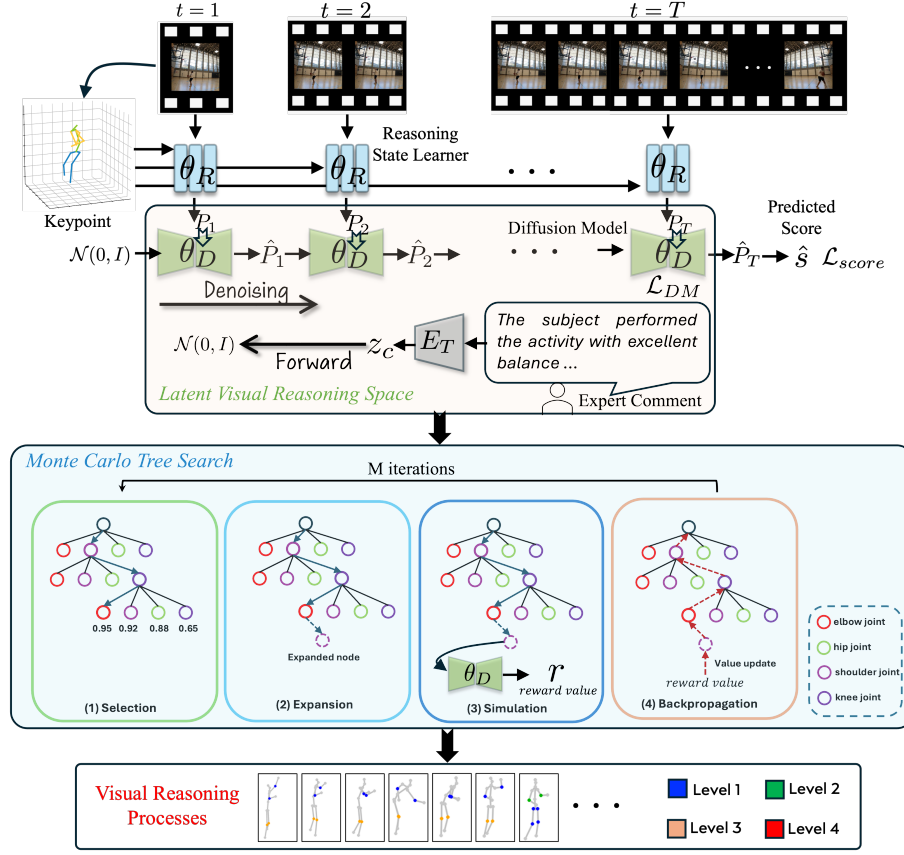


Fig. 2: Overall architecture of Latent Visual Diffusion Reasoning framework. LVDR framework consists of two major components: (1) Latent Visual reasoning space, where we learn visual reasoning embeddings of the given video; (2) Monte Carlo Tree Search to extract explicit visual groundings that correspond to the most influential reasoning steps leading to the final prediction. Keypoint colors represent different levels of attention in the reasoning process. More detailed illustration can be found in Section 3.

temporal reasoning states; (2) Monte-Carlo Tree Search (MCTS) that performs structured exploration in the latent reasoning space to discover informative visual reasoning trajectories.

3.1 Latent Visual Diffusion Reasoning

Visual reasoning aims to provide a detailed illustration of the decision-making process that leads to the final prediction. However, in fine-grained video-based skill assessment, constructing such reasoning is highly challenging because frame-

level annotations describing the reasoning process are nearly impossible to obtain. To address this limitation, we formulate visual reasoning as a diffusion process in a latent reasoning space, where the objective is to learn the target distribution of the reasoning process for the entire video.

Our diffusion model is constructed based on Denoising Diffusion Implicit Models (DDIM) [19]. As shown in Figure 2, we define a forward diffusion process q that gradually adds Gaussian noise to clean text embedding z_c over T discrete time steps, where z_c is derived from expert commentary using a frozen text encoder E_T . During training, we uniformly sample a random step $t \sim \mathcal{U}(\{1, \dots, T\})$. The noised latent z_t at step t can be sampled directly by using the reparameterization:

$$z_t = \sqrt{\bar{\alpha}_t} z_c + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (1)$$

where ϵ is a noise sampled from Gaussian distribution, $\bar{\alpha}_t$ is a noise schedule. We define $\bar{\alpha}_t$ with a cosine function:

$$\bar{\alpha}_t = \cos \left(\frac{(t + 0.5)/T + s}{1 + s} \cdot \frac{\pi}{2} \right)^2, \quad (2)$$

where s is a small offset to prevent $\bar{\alpha}_t$ from approaching zero.

Specifically, each video frame I_i is associated with a corresponding latent visual reasoning embedding $\hat{P}_i$. At the beginning of the video, the reasoning embedding P_1 is inherently noisy due to the limited contextual information available from the first frame. As the video progresses, the diffusion model iteratively denoises these latent embeddings over time, producing progressively refined reasoning representations that converge toward the target distribution P_T . This formulation enables LVDR to model the evolution of reasoning as a continuous denoising trajectory, effectively capturing how the model’s internal reasoning process becomes more coherent and goal-directed as it observes more temporal context.

We first extract 3D keypoints $\{K_i\}$ for each video frame I_i . Then we take $\{K_i\}$ and I_i as inputs to the Reasoning State Learner θ_R to generate the initial visual reasoning embedding $P_i = \theta_R(I_i, K_i) \in \mathbb{R}^d$, where d is the dimension of visual reasoning space. Then we employ a diffusion model θ_D to learn the actual visual reasoning embedding $\hat{P}_i = \theta_D(\hat{P}_{i-1}, P_i)$ based on its previous visual reasoning embedding $\hat{P}_{i-1} \in \mathbb{R}^d$ and conditioned on its own initial embedding P_i . Diffusion model θ_D is implemented with a diffusion transformer architecture and all θ_D at every timestamp share the same weights. The learning objective of diffusion model θ_D is:

$$\mathcal{L}_{DM} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t, \hat{P}_t} \left[\|\epsilon - \epsilon_\theta(P_t, t, \hat{P}_t)\|^2 \right], \quad (3)$$

where ϵ denotes noise and ϵ_θ indicates denoising autoencoder.

With the learned visual reasoning embedding $\hat{P}_t$, we are able to obtain the predicted score $\hat{s} = \text{FC}(\hat{P}_t) \in \mathbb{R}$, where FC represents a fully connected layer. The loss function for regressing the predicted score is:

$$\mathcal{L}_{score} = \|s - \hat{s}\|^2, \quad (4)$$

where s denotes the ground-truth score of the given video. Thus, the total training objective is:

$$\mathcal{L} = \mathcal{L}_{DM} + \lambda \mathcal{L}_{score}, \quad (5)$$

where λ indicates the hyperparameter for weight control and we take it as 1 for implementation.

3.2 Monte-Carlo Tree Search (MCTS)

With the proposed latent visual diffusion reasoning module, we obtain a sequence of latent reasoning embeddings that capture the model’s step-by-step visual reasoning process. The remaining challenge is to transform this implicit reasoning trajectory into an explicitly interpretable form. To address this, we incorporate Monte Carlo Tree Search (MCTS), a principled strategy for balancing exploration and exploitation in sequential decision-making. By integrating MCTS into our framework, we identify and extract explicit visual groundings that correspond to the most influential reasoning steps leading to the final prediction. This design not only enhances interpretability but also provides a transparent and structured view of how the model derives its decisions.

Human referees rely on structured, sequential visual-processing strategies to accurately assess a performer’s actions, systematically attending to different regions of interest over time. For example, a referee may first focus on upper-body kinematics before shifting attention to lower-limb movements, forming a final judgment based on this ordered sequence of visual fixations. Inspired by this temporal and selective nature of human visual perception, we model these visual-processing strategies as combinatorial search spaces. Specifically, we formulate the referee’s visual process as a structured search problem and employ a keypoint-guided MCTS to capture and predict their dynamic attention patterns. For each video frame, we construct a MCTS, as illustrated in Figure 2, to identify the most critical keypoints that contribute to the visual reasoning process.

In the constructed MCTS, each node represents a state (represented with circles of different colors in Figure 2), each edge represents a transition from one state to another. To start with, we first define the tree nodes based on the particular problem we are solving. In sports scenarios, we define tree nodes based on the four major joints that control human movement: *elbow*, *shoulder*, *hip*, and *knee*. Therefore, every tree node in sports scenarios consists of four possible actions. In surgical scenarios, since surgical actions are performed via surgical instruments, we define tree nodes based on the three major parts of instruments: *tip*, *body* and *tail*. Therefore, every tree node in surgical scenarios consists of three possible actions. We use sports scenarios as running examples and include illustrations on surgical scenarios in supplement materials.

For the MCTS of each video frame, we implement the keypoint-guided tree with a finite depth K . Each leaf node involves $K - 1$ steps and corresponds to a unique trajectory through the assessment process, representing a sequence of consecutive. To represent each node, we utilize the three key physical information

of each joint: *angle*, *speed*, and *position*. Formally, each node is represented in the form of (a, v, p) where $a \in [0, 1]$ indicates angle, $v \in \mathbb{R}$ denotes the joint speed and $p = (x, y, z)$ indicates the position of the joint, where $x \in [0, 1]$, $y \in [0, 1]$ and $z \in [0, 1]$ are normalized 3D coordinates. This representation models the hierarchical relationships between major joints and their associated kinematic properties, providing a structured representation of the domain knowledge relevant to skilled human activities.

The decision tree of judgment is constructed iteratively when MCTS is executed over M iterations at each judgment step. Each iteration includes four phases as shown in Figure 2:

- 1. Selection:** Starting from the current root node, we traverse the tree by repeatedly selecting a child node according to the standard Upper Confidence Bound for Trees (UCT) criterion, which balances exploration and exploitation. This process continues until reaching a leaf node or a partially expanded node, ensuring that the search prioritizes both promising and underexplored branches.
- 2. Expansion:** Upon reaching such a node, the expansion step generates new child nodes by exploring previously unvisited paths that partition the remaining possibility space. This operation incrementally refines and narrows the search space, enabling the tree to progressively represent a richer set of candidate reasoning trajectories.
- 3. Simulation:** A rollout policy is employed to estimate the expected reward of the selected node by leveraging the learned diffusion model θ_D . Beginning from this node, the policy simulates a sequence of interactions until either a terminal state is reached or a predefined rollout depth is satisfied. This simulation provides an approximate evaluation of the node’s downstream potential within the reasoning space.
- 4. Backpropagation:** Once the simulation concludes at a leaf node u , the resulting expected reward is propagated backward through the tree. All ancestor nodes along the path from u to the initially selected node are updated accordingly, ensuring that both favorable and unfavorable simulated outcomes influence future selection decisions.

After M iteration, the selected path v^* has the highest expected reward:

$$v^* = \arg \max_{v'} R_e(v'), \quad (6)$$

where R_e indicates the expected reward of the given path.

During inference, we first apply the trained LVDR model to the test videos to obtain the corresponding latent visual reasoning embeddings. We then perform MCTS on these embeddings to derive explicit keypoint-based visual groundings that reveal the model’s step-by-step reasoning process.

4 Experiments

4.1 Datasets

We evaluate our Latent Diffusion Reasoning model on four datasets: (1) **EgoExo4D** [9] dataset comprises multi-view videos of daily and sports activities; for sports

Models	JIGSAWS		FitnessAQA			
	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$	F1 (OK)	F1 (OE)	F1 (SF)	F1 (SI)
MAGR [33]	0.48	10.60	0.564	0.357	0.796	0.157
MAGR++ [31]	0.55	10.20	0.485	0.022	0.818	0.000
LLaVA-Video [29]	0.41	40.61	0.392	0.533	0.795	0.257
Gemini 2.5 [8]	0.10	33.13	0.105	0.128	0.513	0.170
GPT-4o [15]	0.09	71.61	0.000	0.046	0.000	0.000
Qwen-2.5-VL-7B [3]	0.03	67.48	0.404	0.082	0.000	0.000
Ours(Δ Diffusion)	0.65	10.73	0.768	0.497	0.725	0.920
Ours	0.69	9.93	0.747	0.818	0.813	0.920

Table 1: Action score prediction results in JIGSAWS and FitnessAQA dataset. OK: OHP Knee; OE: OHP Elbow; SF: Squat Forward; SI: Squat Inward.

skill assessment, we use three action categories: basketball, soccer and climbing; (2) **JIGSAWS** [7] dataset comprises videos of robot-assisted surgical tasks; used for surgical skills assessment; (3) **FitnessAQA** [16] dataset comprises videos with expert annotations to evaluate the performance of athletes; used for fitness skills assessment; (4) **Cataract-101** [18] dataset comprises videos of cataract surgeries; used for cataract surgical skill assessment. Detailed introduction of datasets can be found in supplement materials.

4.2 Evaluation Metrics

Models	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$
FineParser [27]	0.41	29.36
MAGR [33]	0.73	10.14
MAGR++ [31]	0.71	4.17
Qwen-2.5-VL-7B [3]	0.46	44.26
LLaVA-Video [29]	0.21	16.93
GPT-4o [15]	0.54	11.87
Gemini 2.5 [8]	0.61	5.57
Ours(Δ Diffusion)	0.83	2.60
Ours	0.88	1.69

Table 2: Action score prediction results in EgoExo4D dataset.

We employ two metrics to qualify the performance of models: Spearman’s Rank Correlation (ρ) [26] and relative l_2 distance $R - \ell_2(\times 100)$ [26]. For the FitnessAQA dataset, we adopt the F1-score following the prior work, while for the Cataract-101 dataset, model performance is assessed using accuracy.

4.3 Implementation Details

Training Details. We utilize visual encoder and text encoder from InternVideo2.5 [20] to extract visual and text features. Our diffusion model is a

Transformer-based architecture with 12 layers and 16 attention heads. The dimension of the hidden state is 512. We use Adam optimizer with a learning rate of $1e - 4$ and a batch size of 8 to train our diffusion model and score-prediction module.

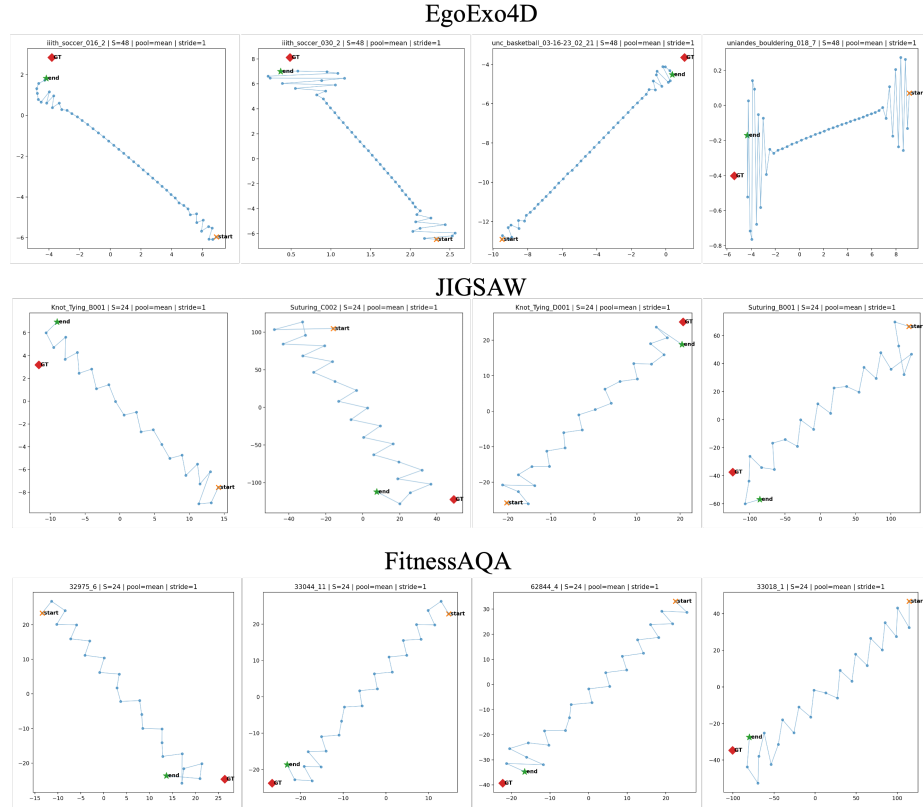


Fig. 3: Visual reasoning trajectory visualization results. We present visual reasoning trajectories on 2D space with three datasets. Trajectories show the progressive reasoning process from start point to end point.

MCTS Inference Details. For the MCTS inference, the depth K is set to 4 in the sports scenarios and 3 in the surgical scenarios. We use the Upper Confidence Bound for Trees(UCT) algorithm for node selection. The reward r is calculated as the cosine similarity between the embedding of the path’s keypoints and the pre-computed target embedding for that step. All experiments are conducted on a single NVIDIA L40S GPU.

4.4 Quantitative Results

We evaluate the quantitative performance of our LVDR model on four datasets spanning both sports and surgical domains. As shown in Table 2, our model achieves state-of-the-art performance on the EgoExo4D dataset, outperforming both domain-specific approaches (e.g., MAGR [33]) and large vision-language models (e.g., GPT-4o [15]) in predicting performance scores. Consistent improve-

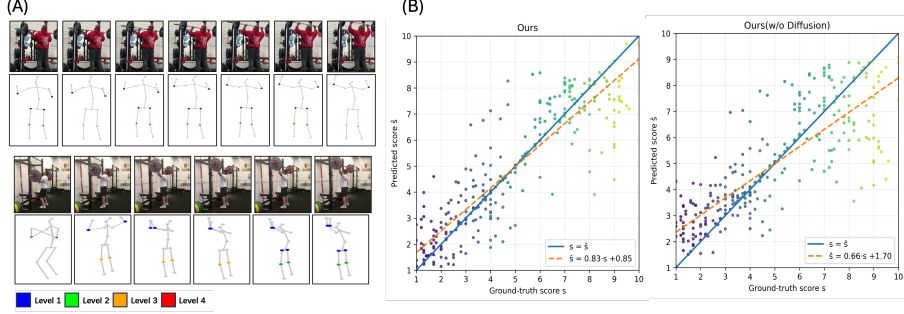


Fig. 4: (A) Keypoint visualization results of FitnessAQA dataset. Keypoint colors represent different levels of attention in the reasoning process. **(B) Scatter plots on EgoExo4D dataset.** We compare the scatter plots for prediction score performance on EgoExo4D dataset with and without diffusion module.

Models	Soccer		Basketball		Climbing	
	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$
MAGR [33]	0.52	8.43	0.72	7.90	0.74	13.32
MAGR++ [31]	0.59	4.44	0.58	6.54	0.67	7.45
FineParser [27]	0.71	45.52	0.12	24.43	0.72	13.18
LLaVA-Video [29]	0.07	29.64	0.31	13.58	0.48	20.23
GPT-4o [15]	0.18	38.64	0.51	10.17	0.75	16.7
Gemini 2.5 [8]	0.64	14.21	0.63	7.01	0.70	4.34
Qwen-2.5-VL-7B [3]	0.70	69.49	0.48	11.37	0.43	59.67
Ours(Δ Diffusion)	0.97	0.67	0.78	4.13	0.83	3.18
Ours	0.94	0.81	0.88	2.08	0.87	2.16

Table 3: Quantitative results of score prediction performance. Experiments on the EgoExo4D dataset for the three sports categories (Soccer, Basketball, Climbing).

ments are also observed in Table 1 and Table 5, further demonstrating the robustness and generalizability of our LVDR framework across diverse video-based skill assessment benchmarks. In Table 3 and Table 4, we also present score prediction performance of individual sports categories (soccer, basketball and climbing) and surgery actions (knot-tying, needle-passing and suturing). In those individual categories, our model also surpasses competing methods in most cases.

4.5 Visual Reasoning Trajectory Evaluation

User Study. We invited a domain expert in soccer and basketball to evaluate the correctness of the keypoint reasoning processes generated by MCTS at each video frame. The expert assigned one of three labels: Correct (C); Wrong (W); Not Sure (N). We assess a total of 50 soccer and basketball videos and report

Models	Knot-Tying		Needle-Passing		Suturing	
	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$
MAGR [33]	0.45	8.84	0.56	13.37	0.42	9.60
MAGR++ [31]	0.50	9.86	0.60	11.10	0.55	9.64
LLaVA-Video [29]	0.45	31.74	0.17	48.69	0.61	41.41
Gemini 2.5 [8]	0.05	29.07	0.14	50.59	0.11	19.73
GPT-4o [15]	0.27	72.09	0.08	119.51	0.07	23.23
Qwen-2.5-VL-7B [3]	0.47	72.81	0.26	108.88	0.29	20.73
Ours(Δ Diffusion)	0.77	5.89	0.53	20.15	0.64	6.15
Ours	0.78	5.37	0.60	18.28	0.68	6.14

Table 4: Quantitative results of score prediction performance on the JIGSAWS dataset. Three surgical activities are included: Knot-Tying, Needle-Passing, Suturing.

percentage results of each category: 86% C; 8% W; 6% N. These results indicate strong alignment between the generated keypoint reasoning processes and human expert judgment.

Keypoint Masking Experiments. To further verify that MCTS identifies critical motion cues for skill assessment, we conduct keypoint masking experiments from two complementary perspectives: **(1)** masking the keypoints selected by MCTS, and **(2)** masking the keypoints *not* selected by MCTS.

Models	Acc $\uparrow$
MAGR [33]	0.62
MAGR++ [31]	0.62
Qwen-2.5-VL-7B [3]	0.53
LLaVA-Video [29]	0.48
Ours(Δ Diffusion)	0.67
Ours	0.71

Table 5: Action score prediction results in Cataract-101 dataset

<i>EgoExo4D</i>	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$	<i>Cataract-101</i>	ACC $\uparrow$	<i>FitnessAQA</i>	F1 (OHP Elbow) $\uparrow$
Ours (Δ MS)	0.86	2.16	Ours (Δ MS)	0.53	Ours (Δ MS)	0.282
Ours (Δ MUNS)	0.88	1.78	Ours (Δ MUNS)	0.71	Ours (Δ MUNS)	0.818
Ours	0.88	1.69	Ours	0.71	Ours	0.818

Table 6: Score prediction results for EgoExo4D, Cataract-101 and FitnessAQA datasets by (1) masking MCTS selected keypoints; (2) masking MCTS unselected keypoints. Results of masking MCTS selected keypoints are highlighted in **red, indicating the difference with the original model performance. Δ MS: mask MCTS selected keypoints; Δ MUNS: mask MCTS unselected keypoints.**



Fig. 5: Keypoint visualization results in soccer and basketball. Keypoint colors represent different levels of attention in the visual reasoning process. Detailed illustration can be found in Section 4.8.

4.6 Inference Latency

In Table 7, we report the inference speed per frame under different numbers of MCTS iterations. The results demonstrate that our model maintains **efficient inference** across these settings, demonstrating its suitability for real-time deployment with fast MCTS-based reasoning.

4.7 Ablation Study

Furthermore, we conduct an ablation study to assess the contribution of the diffusion module within our LVDR framework. As shown in Table 2, 1, 5, 3 and 4, removing the diffusion model leads to a consistent performance degradation across all datasets, confirming its essential role in modeling progressive visual

Ours (w/o MCTS)	Ours (75 MCTS)	Ours (150 MCTS)	Ours (225 MCTS)
0.006s	0.048s	0.094s	0.129s

Table 7: Inference speed per frame. The first column represents our method with only score prediction (no MCTS). Column 2-4 indicate our model with 75, 150 and 225 MCTS iterations, respectively.

EgoExo4D dataset	$\rho \uparrow$	$R - \ell_2(\times 100) \downarrow$	Catatact-101 dataset	Acc $\uparrow$
$\lambda = 0.5$	0.86	3.86	$\lambda = 0.5$	0.67
$\lambda = 2$	0.86	2.07	$\lambda = 2$	0.67
$\lambda = 1$ (Ours)	0.88	1.69	$\lambda = 1$ (Ours)	0.71
liner schedule	0.85	2.25	liner schedule	0.67
cosine schedule (Ours)	0.88	1.69	cosine schedule	0.71

Table 8: Score prediction results with different loss weights and noise scheduler in diffusion model.

reasoning. In addition, the scatter plots in Fig. 4-B further illustrate the performance gap between LVDR variants with and without the diffusion component, reinforcing the effectiveness of diffusion-based reasoning in our framework.

In Table 8, we report results of additional ablation studies including different loss function weights in Eq. 5 weights and noise scheduler in diffusion model.

4.8 Qualitative Results

We visualize the latent visual reasoning trajectories in Figure 3. Each trajectory depicts the temporal evolution of the model’s reasoning state from the initial step to the final prediction. At early steps, the reasoning state lies far from the ground-truth representation, reflecting higher uncertainty in the model’s understanding. As reasoning progresses, the trajectory moves steadily closer to the ground-truth point, indicating a gradual and consistent refinement of the inferred semantics. This visualization demonstrates that our LVDR framework successfully learns a coherent and interpretable reasoning path within the latent visual reasoning space.

To assess the reliability of MCTS, we visualize the keypoint selection results in Figure 4-A and Fig. 5. The color of each keypoint reflects how many times it is selected during the MCTS procedure: Level 1 indicates a single selection, whereas Level 4 indicates four selections. We further highlight in the text the essential semantic cues that correspond to these visualizations (e.g., the description “footwork and movement” in Figure 5 (B) aligns with the strong attention on the knee joints). In Figure 5 (A), the pink boxes illustrate a case where all four joints are selected once, which matches the video frame showing the ball

at mid-air height with no distinctive gesture yet performed, thus requiring uniformly distributed attention. In Figure 4-A, we present the keypoint selection patterns from the FitnessAQA dataset. Here, the knee joints receive the highest selection frequency (shown in orange and green), which aligns perfectly with the fact that the performer is executing squats, where knee articulation is the most critical component of the motion. These qualitative observations collectively demonstrate that our MCTS-based reasoning module consistently identifies the essential keypoints at each timestep, providing faithful and interpretable visual groundings that lead to the model’s final prediction.

5 Conclusion

In this work, we presented LVDR, a novel framework for interpretable fine-grained skill assessment in sports and surgical videos. By formulating visual reasoning as a diffusion process in a latent reasoning space and integrating keypoint-guided Monte Carlo Tree Search (MCTS), LVDR not only delivers accurate performance predictions but also reveals the underlying reasoning trajectories that lead to these assessments. Extensive experiments across diverse datasets demonstrate that our approach achieves competitive quantitative results while providing interpretable insights into the spatiotemporal patterns critical for skill evaluation. Looking forward, we will extend LVDR to handle multi-person scenarios, integrating richer contextual information, and exploring real-time interpretable assessment for live applications in high-stakes domains such as surgical training and elite sports.

Acknowledgement

This work was supported in part by the Virginia Commonwealth University start-up fund.

References

1. Ashutosh, K., Grauman, K.: Learning skill-attributes for transferable assessment in video. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
2. Ashutosh, K., Nagarajan, T., Pavlakos, G., Kitani, K., Grauman, K.: Expertaf: Expert actionable feedback from video. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13582–13594 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), <https://arxiv.org/abs/2502.13923>
4. Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: European conference on computer vision. pp. 422–438. Springer (2022)
5. Coulom, R.: Efficient selectivity and backup operators in monte-carlo tree search. In: International conference on computers and games. pp. 72–83. Springer (2006)
6. Dong, X., Liu, X., Li, W., Adeyemi-Ejeye, A., Gilbert, A.: Interpretable long-term action quality assessment. arXiv preprint arXiv:2408.11687 (2024)
7. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI workshop: M2cai. vol. 3, p. 3 (2014)
8. Google DeepMind: Gemini 2.5. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025> (2025)
9. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanov, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives (2024), <https://arxiv.org/abs/2311.18259>
10. Han, R., Zhou, K., Chen, S., Atapour-Abarghouei, A., Shum, H.P.: Calfow: Enhancing long-term action quality assessment with causal counterfactual flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8231–8241 (2026)
11. Kocsis, L., Szepesvári, C.: Bandit based monte-carlo planning. In: European conference on machine learning. pp. 282–293. Springer (2006)
12. Majeedi, A., Gajjala, V.R., GNVV, S.S.S.N., Li, Y.: Rica²: Rubric-informed, calibrated assessment of actions. In: European Conference on Computer Vision. pp. 143–161. Springer (2024)

13. Matsuyama, H., Kawaguchi, N., Lim, B.Y.: Iris: Interpretable rubric-informed segmentation for action quality assessment. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. pp. 368–378 (2023)
14. Okamoto, L., Parmar, P.: Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 3204–3213 (2024)
15. OpenAI: Gpt-4o. <https://platform.openai.com/docs/models> (2024)
16. Parmar, P., Gharat, A., Rhodin, H.: Domain knowledge-informed self-supervised representations for workout form assessment. In: European Conference on Computer Vision. pp. 105–123. Springer (2022)
17. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
18. Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M.J., Putzgruber, D.: Cataract-101: video dataset of 101 cataract surgeries. In: Proceedings of the 9th ACM multimedia systems conference. pp. 421–425 (2018)
19. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
20. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., Dou, M., Chen, K., Wang, W., Qiao, Y., Wang, Y., Wang, L.: Intern-video2.5: Empowering video mllms with long and rich context modeling (2025), <https://arxiv.org/abs/2501.12386>
21. Wang, Y., Xi, N., Meng, J., Yuan, J.: Interaction-centric spatio-temporal context reasoning for multi-person video hoi recognition. In: European Conference on Computer Vision. pp. 419–435. Springer (2024)
22. Xi, N., Meng, J., Yuan, J.: Forest graph convolutional network for surgical action triplet recognition in endoscopic videos. IEEE Transactions on Circuits and Systems for Video Technology **32**(12), 8550–8561 (2022)
23. Xi, N., Meng, J., Yuan, J.: 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)
24. Xi, N., Meng, J., Yuan, J.: Open set video hoi detection from action-centric chain-of-look prompting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3079–3089 (2023)
25. Xu, H., Ke, X., Li, Y., Xu, R., Wu, H., Lin, X., Guo, W.: Vision-language action knowledge learning for semantic-aware action quality assessment. In: European Conference on Computer Vision. pp. 423–440. Springer (2024)
26. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2949–2958 (2022)
27. Xu, J., Yin, S., Zhao, G., Wang, Z., Peng, Y.: Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment (2024), <https://arxiv.org/abs/2405.06887>
28. Yun, W., Qi, M., Peng, F., Ma, H.: Semi-supervised teacher-reference-student architecture for action quality assessment. In: European Conference on Computer Vision. pp. 161–178. Springer (2024)
29. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data (2025), <https://arxiv.org/abs/2410.02713>

30. Zhao, S., Wang, Z., Luan, T., Jia, J., Zhu, W., Luo, J., Yuan, J., Xi, N.: Pp-motion: Physical-perceptual fidelity evaluation for human motion generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6840–6849 (2025)
31. Zhou, K., Pan, Q., Zhang, X., Shum, H.P.H., Li, F.W.B., Liang, X., Wang, L.: Continual action quality assessment via adaptive manifold-aligned graph regularization (2025), <https://arxiv.org/abs/2510.06842>
32. Zhou, K., Shum, H.P., Li, F.W., Zhang, X., Liang, X.: Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. *IEEE transactions on image processing* **34**, 3718–3732 (2025)
33. Zhou, K., Wang, L., Zhang, X., Shum, H.P.H., Li, F.W.B., Li, J., Liang, X.: MAGR: Manifold-Aligned Graph Regularization for Continual Action Quality Assessment, p. 375–392. Springer Nature Switzerland (Nov 2024). https://doi.org/10.1007/978-3-031-73247-8_22, http://dx.doi.org/10.1007/978-3-031-73247-8_22

Latent Visual Diffusion Reasoning with Monte Carlo Tree Search

Supplementary Materials

1 Datasets Details

1.1 EgoExo4D

The EgoExo4D [9] dataset is a large-scale, multimodal, and multiview dataset, providing time-synchronized videos from both egocentric (first-person) and exocentric (third-person) views. We utilize this dataset for sports skills assessment, focusing on three types of sports: soccer, basketball and climbing. We use the action quality ratings, which are provided by the dataset, as the ground truth for our assessment task.

1.2 JIGSAWS

The JIGSAWS [7] dataset is a widely used dataset for robot-assisted surgical skill assessment. It comprises recordings of three fundamental surgical tasks (knot-tying, needle-passing, suturing) performed by surgeons on bench-top models, all commonly included in surgical training curricula. We utilize this dataset for surgical skills assessment.

1.3 FitnessAQA

The FitnessAQA [16] dataset is specifically designed to evaluate the performance of athletes in fitness exercises. It contains videos and expert annotations for three exercises: BackSquat, Barbell Row and Overhead (shoulder) Press. We utilize this dataset for fitness skills assessment.

1.4 Cataract-101

The Cataract-101 [18] dataset consists of 101 videos from surgical microscopes during real-world cataract surgeries. The videos are annotated with overall skill ratings (novice/expert). Our task is to predict the overall skill rating from the video.

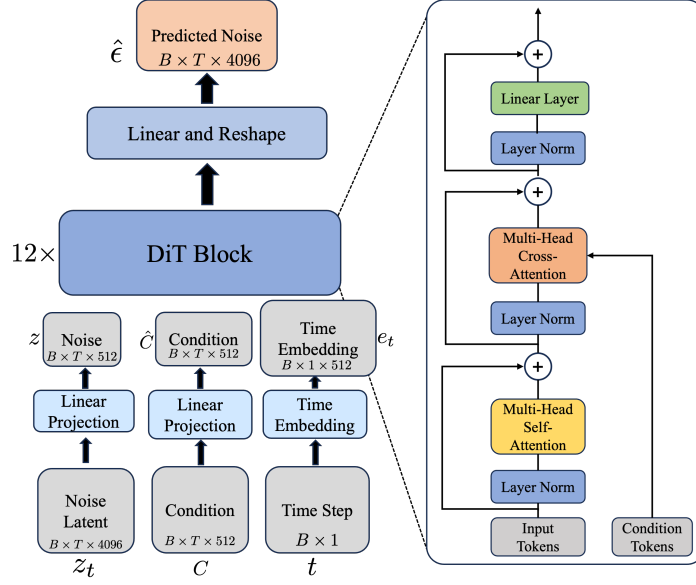


Fig. 1: Detailed Architecture of Diffusion Transformers. We utilize 12 DiT blocks to implement diffusion model. The detailed architecture of DiTs is shown on the right side.

2 More on Monte Carlo Tree Search (MCTS)

In our implementation of MCTS, the selection phase is critical for balancing the trade-off between exploring new paths and exploiting paths that are already known to be promising. We employ the Upper Confidence Bound for Tree (UCT) [11] algorithm in this phase. The core idea of the UCT algorithm is to treat each node selection as a multi-armed bandit problem, where the algorithm must choose the child node that maximizes a specific score. During the tree traversal, when MCTS must select a child node c from a parent node p , it calculates a UCT score for every child. The UCT formula we use is as follows:

$$UCT(c) = \frac{W(c)}{N(c)} + C \cdot \sqrt{\frac{\ln N(p)}{N(c)}} \quad (1)$$

where $UCT(c)$ is the score calculated for the child node c ; $W(c)$ is the total reward accumulated from all simulations that have passed through the node c ; $N(c)$ is the total number of times node c has been visited; C is the exploration constant, which we set to $C = \sqrt{2}$ in our experiments; and $N(p)$ is the total number of times the parent node p has been visited.

3 More on Latent Visual Diffusion Reasoning Framework

3.1 Detailed Architecture of Diffusion Transformers for Diffusion Module.

Following Section 3.1 of the main paper, we instantiate our diffusion backbone using the Diffusion Transformer (DiTs) architecture [17], illustrated in Fig. 1.

Input processing begins by linearly projecting the noisy latent $z_t \in \mathbb{R}^{B \times T \times 4096}$ into a hidden embedding space of dimension $D = 512$, resulting in feature representations $z \in \mathbb{R}^{B \times T \times D}$. Simultaneously, the diffusion timestep $t \in \mathbb{R}^{B \times 1}$ is mapped to sinusoidal positional encodings ($D = 512$) and processed by a learnable Multi-Layer Perceptron (MLP) to yield the time embedding $e_t \in \mathbb{R}^{B \times 1 \times D}$.

The backbone comprises 12 stacked DiT blocks. Within each block, we employ a masked multi-head self-attention mechanism to strictly enforce temporal causality. This is followed by a cross-attention layer that integrates the conditional context $C \in \mathbb{R}^{B \times T \times D}$ to guide the generation process. Finally, the processed features undergo layer normalization and a final linear projection to predict the noise component $\hat{e} \in \mathbb{R}^{B \times T \times 4096}$.

4 MCTS Inference Algorithm

In Algorithm 1, we present details on test-time pre-computation of MCTS target embedding $\mathbf{e}_t^*$.

Algorithm 1 Test-time Pre-computation of MCTS Target Embeddings

Require: test video $\mathbf{V}$; human keypoints $\mathbf{K}$; trained reasoning state learner θ_R ; trained diffusion denoiser ϵ_θ ; denoising steps T ; canonical human keypoint order $\mathbf{c}$

Output: Target embedding $\mathcal{E} = \{\mathbf{e}_t^*\}_{t=1}^T$

```

1:  $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $\mathcal{E} \leftarrow \emptyset$ 
2: for  $t = T, T-1, \dots, 1$  do
3:    $\mathbf{x}_t = \theta_R(\mathbf{V}, \mathbf{K}, \mathbf{c})$  ▷ Construct the default condition prompt
4:    $\hat{\mathbf{e}}_t \leftarrow \epsilon_\theta(\mathbf{z}_t, t, \mathbf{x}_t)$  ▷ Predict noise
5:    $\hat{\mathbf{z}}_{0,t} \leftarrow (\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \hat{\mathbf{e}}_t) / \sqrt{\bar{\alpha}_t}$  ▷ Estimate denoised embedding
    $\mathbf{e}_t^* \leftarrow \hat{\mathbf{z}}_{0,t}$  ▷ MCTS target embedding  $\mathbf{e}_t^*$  for step  $t$ 
6:    $\mathcal{E}[t] \leftarrow \mathbf{e}_t^*$  ▷ Cache the target embedding  $\mathbf{e}_t^*$ 
7:    $\alpha_t \leftarrow \bar{\alpha}_t / \bar{\alpha}_{t-1}$ ,  $\beta_t \leftarrow 1 - \alpha_t$ 
8:    $\mathbf{z}_{t-1} \leftarrow \left( \mathbf{z}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\mathbf{e}}_t \right) / \sqrt{\alpha_t}$  ▷ Deterministic DDIM update
9: end for
10: return  $\mathcal{E} = \{\mathbf{e}_t^*\}_{t=1}^T$ 

```

5 More Visual Reasoning Trajectory Results

In Fig. 2, we present more reasoning trajectory visualization results on the EgoExo4D dataset and JIGSAW dataset.

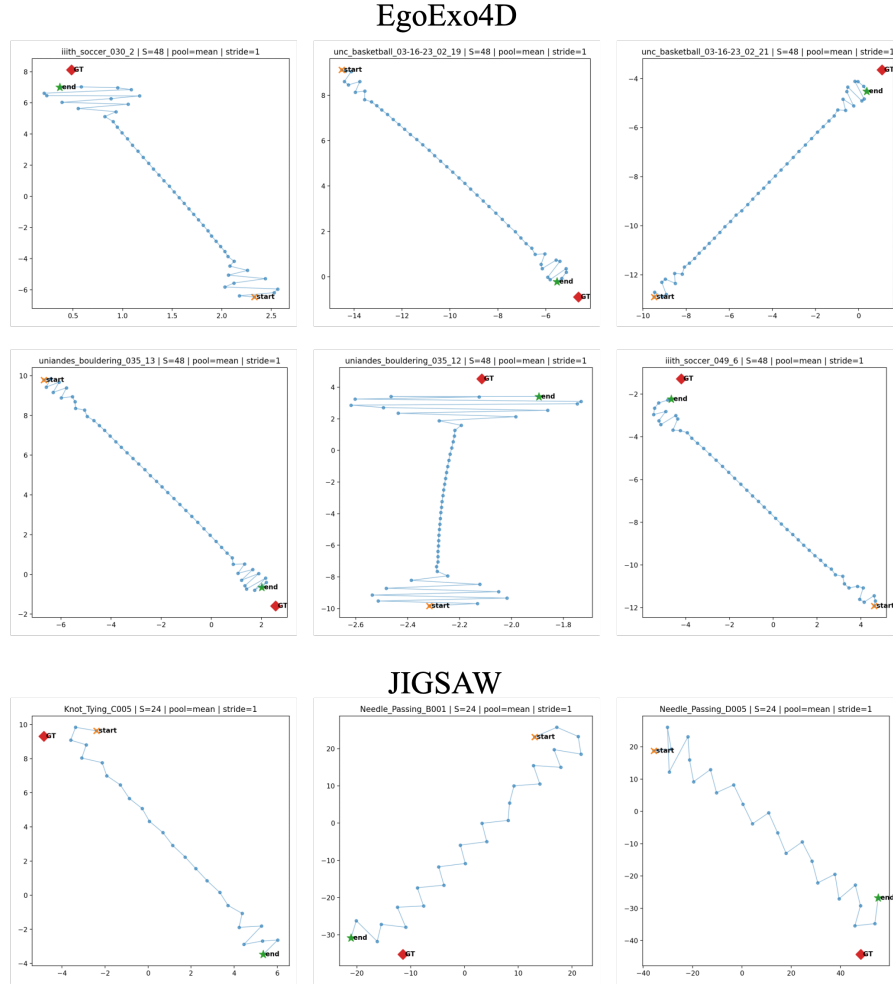


Fig. 2: Visual reasoning trajectory visualization results. We present visual reasoning trajectories on 2D space with three datasets. Trajectories show the progressive reasoning process from start point to end point.