Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 3, 2025

Lang2Vision Diffusion: Language-Driven Diffusion for Robotic Action Frame Prediction

Guanyu Chen *

School of Mathematics, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu, 211106, China
*** Corresponding author Email:** 489709134@qq.com

Abstract: We address the challenge of enabling robots to predict the visual outcomes of their actions through Lang2Vision Diffusion (L2V-Diff), a novel adaptation of InstructPix2Pix for robotic action frame prediction. Our framework takes an initial RGB observation paired with natural language instructions and generates photorealistic images of anticipated future states via vision-language conditioned diffusion. The method is fine-tuned on synthetic RoboTwin data (300 episodes across hammering, handover, and stacking tasks), demonstrating strong quantitative performance (mean Structural Similarity: 0.971, Peak Signal-to-Noise Ratio: 37.1 dB). By bridging high-level instructions with pixel-accurate visual prediction, L2V-Diff advances safety-critical robotic applications while eliminating the need for explicit 3D reconstruction or physics simulation.

Keywords: Robotic Vision; Prediction; Diffusion Model.

1. Introduction

Predicting the future visual states of robotic actions is crucial for ensuring safe and efficient task execution in dynamic environments. With the advancement of deep learning and computer vision technologies, a variety of vision-based predictive models have emerged in recent years. These models aim to endow robots with "visual foresight" capabilities[1], enabling them to facilitate decision-making and planning by simulating future scenes. Early approaches, such as the Visual Foresight model[2], employed video prediction techniques to allow robots to anticipate potential outcomes before executing an action, thereby enabling model-free planning and control. However, these methods largely relied on raw visual inputs and lacked the ability to understand high-level semantic instructions.

More recently, researchers have begun exploring ways to integrate natural language instructions with visual inputs in order to enhance the generalization and interactivity of predictive models. For instance, the introduction of the InstructPix2Pix architecture enabled image editing based on textual commands, offering a new perspective on vision-language joint modeling. Meanwhile, emerging frameworks such as PAD (Prediction with Action Diffuser)[3] attempt to unify action prediction and image generation within a single denoising process, achieving joint modeling of future states and action sequences. Additionally, works like ManiTrend[4] leverage causal Transformers to model the relationships between 3D particle dynamics, visual observations, and manipulation actions, further improving the physical plausibility of predictions.

In this work, we propose a novel adaptation of the InstructPix2Pix architecture to model and predict the visual outcomes of robotic interactions with the environment. Our approach integrates both visual and textual modalities as input—taking an initial scene image along with a natural language instruction (e.g., "stack the blocks" or "handover the blocks")—and generates a realistic image depicting the expected future state after the instruction is executed. This enables the system to not only understand high-level semantic commands but also simulate their physical consequences in the visual domain. By leveraging the powerful image-to-image translation capabilities of InstructPix2Pix while adapting it for forward dynamics modeling, our method provides a data-driven way to learn visual foresight without explicit 3D scene reconstruction or physics simulation. The resulting framework can be applied to tasks such as action planning, intention communication, and failure prediction in complex, vision-based environments.

Quantitative evaluation across three robotic manipulation tasks demonstrates strong performance, with mean SSIM of 0.971 and average PSNR of 37.1 dB, confirming the model's effectiveness in both structural preservation and signal fidelity.

2. Related Work

The ability of robots to predict future states is crucial for ensuring safe and efficient operation in complex and dynamic environments. In recent years, substantial progress has been made in leveraging multimodal information—particularly vision and language—for robot task planning, action prediction, and control.

Diffusion models, owing to their strong image and video generation capabilities, have emerged as a promising direction in robotics, especially for tasks such as trajectory generation, grasp prediction, and data augmentation[5]. These models are well suited for handling high-dimensional input and output spaces and for modeling complex multi-modal distributions[6], which is essential for manipulation tasks that often have multiple valid solutions (e.g., trajectory planning)[7]. Recent work has begun to apply diffusion models to robotic manipulation, taking advantage of their probabilistic structure to handle uncertainty and improve robustness. Survey studies have also summarized how diffusion models can be integrated with imitation learning and reinforcement learning, or used to address data scarcity issues in robotics[8]. Frameworks such as Unified World Models (UWM) further attempt to unify policy learning, dynamics modeling, and video prediction through modality-specific diffusion timesteps.

Enabling robots to understand and execute natural language instructions is an intuitive and effective approach to

human–robot interaction[9]. A significant body of research focuses on developing multitask policy models that can interpret language commands and integrate them with visual observations[10], which requires deep fusion of visual and linguistic information.

Some studies use pre-trained vision–language models (VLMs) to control robots by predicting key points or action primitives from language instructions and visual inputs[11]. Other work develops multimodal fusion neural architectures for navigation tasks, combining visual inputs (e.g., RGB images, depth maps) and language instructions via CNNs and GRUs, often enhanced by attention mechanisms to improve feature integration[12].

Recent approaches also incorporate explicit visual reasoning into Vision-Language-Action (VLA) frameworks by predicting future images as visual goals—i.e., visual chain-of-thought—before generating action sequences to reach those goals. In addition, using 3D flow as an intermediate representation between language-guided future image prediction and fine-grained action generation has emerged as a promising direction. Predictive inverse dynamics models attempt to infer actions by predicting future visual states of the robot, enabling end-to-end learning of vision and action[13].

Inspired by the success of large pretrained models in natural language processing and computer vision, researchers have increasingly explored transferring the representational capabilities of such models (e.g., text-to-image diffusion models) to robot control tasks. By extracting features from or fine-tuning these pretrained models, robots can obtain powerful visual–spatial and semantic representations that significantly enhance the generalization ability of control policies in complex, open-ended environments.

For example, recent work investigates using intermediate-layer features from pretrained diffusion models to learn policies that generalize well to diverse manipulation and navigation tasks[14]. Other research explores adapting large models to robot-specific datasets through fine-tuning[15]; despite the relatively small scale of robotic datasets compared to internet-scale training corpora, fine-tuning has proven far more sample-efficient than training from scratch. Techniques such as DreamBooth and Hypernetworks[16] have also been applied for personalized fine-tuning of text-to-image models for robotics.

In summary, integrating multimodal information—particularly leveraging the generative and representational strengths of large pretrained models and diffusion models—while effectively modeling temporal dependencies constitutes a key research direction for robot action prediction and control.

3. L2V-Diff Architecture

This part details our methodology for training an image generation model to predict future frames of robotic actions in virtual simulation environments. The project focuses on three core tasks: hammering blocks, handing over blocks, and stacking blocks. We adapted the InstructPix2Pix model to process both visual inputs (RGB frames) and textual instructions to generate future state predictions.

The specific methodology encompasses three core components: dataset construction, training procedure, and performance evaluation.

3.1. Dataset Construction

The training and testing datasets were synthesized using the RoboTwin simulation platform, focusing on three fundamental robotic manipulation tasks:

- Block Hammering (block_hammer_beat):
  - Instruction: “Handover the blocks.”
  - Visual content: The robot executes a bimanual transfer, passing one or more blocks between its end-effectors.
- Block Handover (block_handover):
  - Instruction: “Handover the blocks.”
  - Visual content: The robot executes a bimanual transfer, passing one or more blocks between its end-effectors.
- Block Stacking (blocks_stack_easy):
  - Instruction: “Stack blocks.”
  - Visual content: The robot sequentially places blocks to construct a vertical stack or tower.

For each task, we collected 100 independent episodes. Each episode contains the following components:

- Initial observation frame $(\mathrm{t} = 0)$: Captures pre-execution state
- Target frame $(\mathrm{t} = \mathrm{T})$: Documents task completion status
- Paired natural language instruction: Precisely describes the intended action

The final compiled dataset consists of 300 annotated samples (3 tasks × 100 episodes), enforced conversion of images to 3-channel format for model compatibility, providing comprehensive data support for model fine-tuning. Each sample triplet $\mathrm{D}_i = (I_0, I_T, \mathrm{T})$ includes:

- $I_0 \in \mathbb{R}^{H\times W\times 3}$: Initial RGB observation
- $I_T \in \mathbb{R}^{H\times W\times 3}$: Target RGB state
- T : Corresponding textual instruction

where $H = 256$ and $\mathrm{W} = 256$ denote the standardized image resolution.

3.2. Model Architecture and Training

- Input Adaptation:
  - Visual Encoder: UNet-based image encoder processes 256×256 RGB inputs into $32 \times 32 \times 4$ latent representations via AutoencoderKL (compression ratio: $64\times$).
  - Text Encoder: Frozen CLIP ViT-L/14 embeds instructions into 77×768 token embeddings
- Multimodal Fusion:
  - Hybrid Conditioning: Concatenates image latent codes (4D) with text features (4D) to form 8-channel UNet input.
  - Cross-Attention: 8-head transformers $(d_{ctx} = 768)$ operate at resolutions $[4, 2, 1]$.
- Diffusion Process:
  - Noise Schedule: 1000-step DDPM with $\beta : 0.00085 \rightarrow 0.012$
  - Training: EMA-optimized $(\mu = 0.9999)$ with $\mathrm{L} = \mathrm{L}_{\mathrm{l}} + \mathrm{L}_{LPIPS} + \mathrm{L}_{CLIP}$ at effective batch size 32

The model establishes a hierarchical feature extraction-reconstruction pipeline through the synergistic design of Latent Space and UNet. The $32 \times 32 \times 4$ latent space coupled with a UNet architecture of channel dimensions $[320, 640, 1280, 1280]$ enables progressive feature abstraction from local details to global semantics. Cross-modal alignment between text and image is achieved via $77 \times 768 - dimensional$ CLIP text embeddings interacting through an 8-head attention mechanism (each head proc-

essing a 96-dimensional subspace). Multi-resolution attention layers deployed at $[4,2,1]$ scales capture:

• Scene-level semantics at low resolution (4×4 feature maps corresponding to 8×8 input)

• Inter-object relationships at medium resolution (2×2 feature maps for 16×16 input)

• Local detail refinement at high resolution (1×1 feature maps for 32×32 input)

This co-design paradigm optimally balances computational efficiency with generation quality.

Table 1. Model Specifications

Component	Configuration
Latent Space	$32\times32\times4$
UNet Channels	$[320,640,1280,1280]$
Text Context	77×768
Attention Heads	8 (resolutions $[4,2,1]$)

Table 2. Key Training Configuration

Parameter	Value
Base Learning Rate	1.0×10^{-4}
Effective Batch Size	$32(2\times16)$
EMA Decay (μ)	0.9999
Diffusion Steps	900
Noise Schedule (β)	$0.00085\rightarrow0.012$
Loss Weights	$L_1:0.3, L_{LPIPS}:0.5, L_{CLIP}:0.2$

3.3. Quantitative Evaluation

We employ two well-established metrics for systematic performance assessment:

• Structural Similarity (SSIM):

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \quad (1)$$

where μ denotes local means, σ represents variances, and C_1, C_2 are stability constants. SSIM values range in $[0,1]$ with higher scores indicating better structural preservation.

• Peak Signal-to-Noise Ratio (PSNR):

$$\mathrm{PSNR}=10\cdot\log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \quad (2)$$

where MAX_I is the maximum pixel value (255 for 8-bit images) and MSE denotes mean squared error. Higher PSNR (dB) reflects superior signal fidelity.

The evaluation protocol follows:

1. Compute metrics between all predicted-target image pairs

2. Conduct per-task analysis to identify performance variations

4. Experiment

4.1. Environment Preparation

See Table 3.

4.2. Key Parameters

• Dataset Preparation:

- save_type (raw_data): true

- data_type (RGB): true

- episode_num: 100

Table 3. Server Configuration Details

Category	Configuration Details
GPU	RTX 4090D (24GB) * 2
Deep Learning Framework	PyTorch 2.5.1
Programming Language	Python 3.12 (Ubuntu 22.04)
Computing Platform	CUDA 12.4

• Training:

- max_epochs: 100

- accumulate_grad_batches: 16

- min_resize_res: 256

- max_resize_res: 256

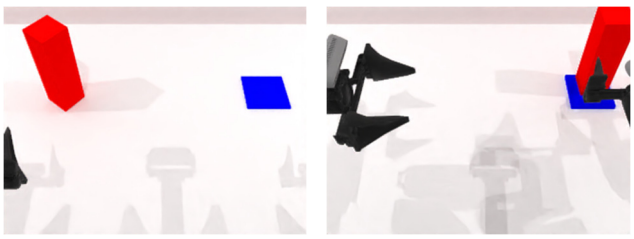
- batch_size: 2 (per GPU)

4.3. Results Visualization

RoboTwin autonomously synthesizes annotated training datasets through physics-based simulation of optimal action sequences (e.g., stack blocks operations), serving as supervisory signals for model training while generating multi-view robotic observations with randomized initial states. Figures 1-3 comparatively illustrate the input frames versus RoboTwin-simulated target frames across different task scenarios.

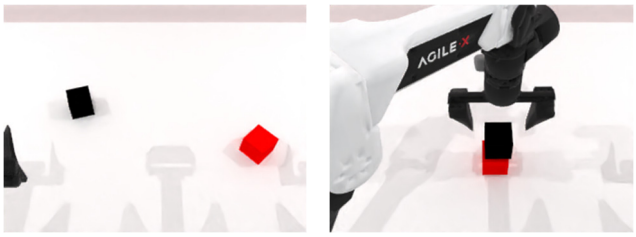
(a) Input Observation (b) RoboTwin Simulation

Figure 1. Task 1: Beat the block with the hammer.



(a) Input Observation (b) RoboTwin Simulation

Figure 2. Task 2: Handover the blocks.



(a) Input Observation (b) RoboTwin Simulation

Figure 3. Task 3: Stack blocks.

The test sets were evaluated using SSIM and PSNR metrics: Block Stack task demonstrates the lowest dual metrics, with a 9.3 dB PSNR degradation compared to the Handover task,

primarily due to coupled texture-luminance degradation induced by multi-object occlusion.

Block Hammer Beat task exhibits a striking divergence between peak SSIM (0.982) and valley PSNR (36.5 dB), revealing its unique trade-off characteristic between high-frequency detail loss and structural integrity preservation.

The final validation loss stabilizes at 0.042 (Figure 4), confirming robust convergence under data-limited conditions.

The visualization of robotic action frame prediction is shown in Figure 5-7.

Table 4. Performance Metrics for Different Tasks

Task	SSIM	PSNR (dB)
Block Hammer Beat	0.982	36.5
Block Handover	0.974	42.1
Block Stack (Easy)	0.956	32.8

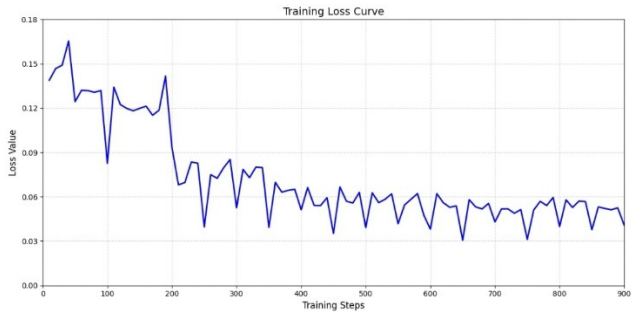


Figure 4. Training Loss

(a) Input (b) Target (c) Prediction

Figure 5. Comparative analysis of stack blocks task execution: (a) Original input sequence with uniform values, (b) RoboTwin-simulated target sequence, and (c) Model-predicted sequence after fine-tuning.

(a) Input (b) Target (c) Prediction

Figure 6. Comparative analysis of handover-the-blocks task execution: (a) Original input sequence with uniform values, (b) RoboTwin-simulated target sequence, and (c) Model-predicted sequence after fine-tuning.

(a) Input (b) Target (c) Prediction

Figure 7. Comparative analysis of stack blocks task execution: (a) Original input sequence with uniform values, (b) RoboTwin-simulated target sequence, and (c) Model-predicted sequence after fine-tuning.

5. Conclusion

This work establishes multimodal fine-tuning as an effective paradigm for robotic action prediction, demonstrating how instruction-conditioned diffusion models (e.g., adapted InstructPix2Pix) can reliably forecast visual states in dynamic environments by jointly processing RGB observations and textual commands. The results highlight the viability of pretrained diffusion models for safety-sensitive robotic applications requiring explainable visual predictions.

References

[1] Chelsea Finn, Sergey Levine. Deep visual foresight for planning robot motion. In 2017 IEEE International Conference on Robotics and Automation (ICRA), 2786–27993, 2017.

[2] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. arXiv preprint arXiv:1812.00568, 2018.

[3] Tsung-Wei Ke, Nikolaos Gkanatsios, Katerina Fragkiadaki. 3D Diffuser Actor: Policy diffusion with 3D scene representations. arXiv preprint arXiv:2402.10885, 2024.

[4] Yuxin He, Qiang Nie. ManiTrend: Bridging Future Generation and Action Prediction with 3D Flow for Robotic Manipulation. arXiv preprint arXiv:2502.10028, 2025.

[5] Rosa Wolf, Yitian Shi, Sheng Liu, Rania Rayyes. Diffusion Models for Robotic Manipulation: A Survey. arXiv preprint arXiv:2504.08438, 2025.

[6] Jessica E. Liang. Diffusion Models for Robotics. In Proceedings of the AAAI Conference on Artificial Intelligence, 39:29587–29589, 2025.

[7] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, Abhishek Gupta. Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets. arXiv preprint arXiv:2504.02792, 2025.

[8] Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, Xiaodan Liang. VidMan: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. Advances in Neural Information Processing Systems, 37:41051–41075, 2024.

[9] Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, et al. Multimodal Fusion and Vision-Language Models: A Survey for Robot Vision. arXiv preprint arXiv:2504.02477, 2025.

[10] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. arXiv preprint arXiv:2503.22020, 2025.

[11] Gunshi Gupta, Karmesh Yadav, Yarin Gal, Dhruv Batra, Zsolt Kira, Cong Lu, Tim G. J. Rudner. Pre-trained text-to-image diffusion models are versatile representation learners for control. Advances in Neural Information Processing Systems, 37:74182–74210, 2024.

[12] Yiping Zhang, Kolja Wilker. Visual-and-Language Multimodal Fusion for Sweeping Robot Navigation Based on CNN and GRU. Journal of Organizational and End User Computing (JOEUC), 36(1):1–21, 2024.

[13] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. arXiv preprint arXiv:2412.15109, 2024.

[14] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman. DreamBooth: Fine-tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22500–22510, 2023.

[15] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023.

[16] David Ha, Andrew Dai, Quoc V. Le. Hypernetworks. arXiv preprint arXiv:1609.09106, 2016.