VACE-PhysicsRL: Unified Controllable Video Generation through Physical Laws and Reinforcement Learning Alignment

Yixiao Kang   Sining Huang   Yukun Song

October 19, 2025

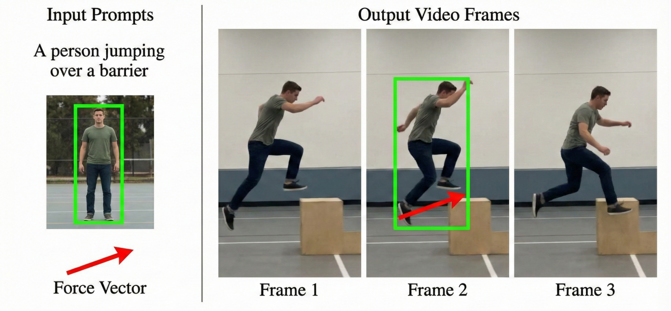


Fig. 1: VACE-PhysicsRL enables unified, controllable video generation. Given sparse inputs like bounding boxes (green) and force vectors (red arrow) , our model synthesizes high-fidelity video that respects physical laws and identity consistency.

***Abstract*—The pursuit of unified video generation platforms, such as VACE [1], has consolidated diverse tasks like editing and creation into a single framework. However, achieving precise, multi-faceted control over physics, fine-grained object dynamics, and multi-human identity remains a critical challenge. We introduce VACE-PhysicsRL, a novel extension of the Video Condition Unit (VCU) architecture that integrates advanced control modalities to synthesize highly plausible and controllable video content. Our method incorporates mechanisms for fine-grained object appearance and trajectory control, leveraging sparse inputs such as bounding boxes and reference images [2]. It employs physics-aware motion guidance derived from the Euler-Lagrange equations [3], and identity alignment refined through Group Relative Policy Optimization (GRPO) [4]. VACE-PhysicsRL processes sparse, intuitive inputs like bounding boxes, reference images, and estimated pose/force cues to output videos demonstrating superior fidelity, temporal consistency, and adherence to complex physical and identity constraints. Extensive experiments confirm that the framework significantly enhances the controllability and realism of unified video generation, exceeding baseline performance across metrics related to motion plausibility and identity consistency.**

I. Introduction

State-of-the-art video generation models excel at translating textual descriptions into visual content, yet they often lack the fidelity required for fine-grained control over dynamics and individual entities [2], [3]. Specifying detailed object movements or ensuring consistent identities among multiple interacting subjects proves difficult when relying solely on text prompts [2], [4]. This limitation is exacerbated in complex scenarios demanding physical plausibility, such as simulating forces [5] or highly dynamic human movements [3].

The **VACE** (Video All-in-one Creation and Editing) framework represents a major step toward unifying diverse video synthesis tasks, including Reference-to-Video (R2V), Video-to-Video (V2V), and Masked Video-to-Video (MV2V), within a single architecture [1], [6]. VACE achieves this flexibility by leveraging a universal Video Condition Unit (VCU) to ingest multimodal inputs (text, frames, and masks) [1], [6].

To address the shortcomings in fine-grained control and dynamic realism within such unified models, we propose **VACE-PhysicsRL**, an extension incorporating three critical, specialized control mechanisms: 1. **Fine-Grained Entity Control:** Enabling users to specify sparse trajectory (via bounding boxes) and appearance (via reference images) controls for individual objects [2]. 2. **Physics-Aware Motion Guidance:** Injecting explicit physical constraints, either through refined skeletal poses for complex actions [3] or intuitive force prompts for interactive scenarios [5]. 3. **Reinforcement Learning Alignment:** Optimizing the model policy to maintain consistency, particularly for challenging Multi-Human Identity-Preserving Video Generation (MH-IPV), guided by specialized human preference rewards [4].

The primary contribution of this work is the development of a unified architecture that natively integrates these orthogonal control modalities without compromising the base model's general capabilities, providing a robust solution for synthesizing highly controlled and physically plausible video content.

II. Related Work

A. Unified and Controllable Video Generation

The VACE framework unifies creation and editing tasks by employing a Video Condition Unit (VCU) to handle diverse modalities including text, context frames, and masks [1], [6]. This setup supports tasks like R2V, V2V, and MV2V editing, as well as their compositions [1], [6]. We leverage VACE's modularity, specifically its Context Adapter structure, to inject the output of our specialized control processors [1], [6].

B. Fine-Grained Control and Multimodal Input

The challenge of achieving fine-grained output control solely through natural language is well-established [2]. Approaches like FACTOR [2] address this by incorporating multimodal inputs: text, user-drawn bounding boxes (for trajectory), and user-provided reference images (for appearance) of individual objects [2]. Similarly, MotionPro focuses on precise motion control in Image-to-Video (I2V) generation by using region-wise trajectories derived from flow maps and motion masks to distinguish between object and camera motion [7]. Our work adopts the paradigm of sparse, intuitive inputs for localized control.

C. Physics and Pose-Based Guidance

Achieving physically plausible motion, especially for large body deformations, is difficult for purely data-driven generative models [3]. FinePhys [3] addresses this by explicitly incorporating physical laws, such as the Euler-Lagrange equations, to refine data-driven 3D pose estimates into physically predicted poses [3]. This information is injected as multi-scale 2D heatmaps during the diffusion process [3]. Complementary to pose control, Force Prompting demonstrates that video models can learn and generalize responses to physics-based signals, such as localized point forces or global wind fields, even exhibiting an emergent understanding of properties like mass [5].

D. Identity Preservation via Reinforcement Learning

For complex tasks like Multi-Human Identity-Preserving Video Generation (MH-IPV), models like VACE and Phantom struggle to maintain identity consistency across dynamic interactions [**?**], [4]. Identity-GRPO [4] addresses this by training a specialized reward model on human-annotated preference data focused on individual identity consistency, then using Group Relative Policy Optimization (GRPO) to refine the generation policy [4]. This optimization is critical for decoupling identity cues from general compositional similarity.

III. VACE-PhysicsRL Methodology

VACE-PhysicsRL augments the standard VACE architecture by introducing dedicated modules for processing specialized control inputs, ensuring that physical constraints and high-fidelity object attributes are respected throughout the generation process.

A. Enhanced VCU Input Handling

We expand the VCU inputs to explicitly handle the sparse, granular controls necessary for fine-grained generation. The extended input set for a video token sequence incorporates:

1) **Entity-Level Control (E-Control):** Sparse inputs specifying appearance (r_{nt}) via reference images and trajectory (l_{nt}) via bounding boxes for up to N individual entities at time t [2].
2) **Structured Dynamics Control (D-Control):** Inputs providing explicit kinematic or kinetic information, such as estimated 2D pose sequences or generalized force vectors (π).

Similar to FACTOR, the E-Control uses a joint encoder to integrate text prompts, location coordinates, and CLIP image embeddings of the reference appearance, concatenating them into entity embeddings ent [2].

B. Physics Regulation Module (PRM)

To handle dynamic realism, the D-Control signal passes through a dedicated Physics Regulation Module (PRM), integrating insights from FinePhys and Force Prompting.

1) Physics-Aware Pose Refinement: For human action scenarios (e.g., gymnastics), the input 2D pose is lifted to a data-driven 3D pose (S_{dd}^{3D}) using an in-context learning process (ICL) [3]. The PRM then employs a **PhysNet** module, which explicitly incorporates the structure of the Euler-Lagrange equations to model rigid-body dynamics and calculate bidirectional joint accelerations [3]. This produces a *physics-refined 3D pose sequence* [3]. The fused 3D pose information is then projected back to multi-scale 2D heatmaps for injection into VACE's Diffusion Transformer architecture. This adds crucial physical guidance [3].

2) Force-Conditioned Interaction: For scenarios involving non-human object interaction, the PRM processes user-defined force prompts π, categorized as either local or global forces [5]. These forces are encoded into a spatiotemporal tensor representation (π) [5]. This conditioning enables the model to simulate dynamic responses based on visual context, successfully demonstrating generalization to diverse settings, objects, geometries, and even hints at mass understanding [5].

C. Identity-Optimized Alignment

To ensure high-fidelity multi-human identity preservation (MH-IPV), we adopt the Identity-GRPO pipeline as a post-training policy refinement stage [4].

1) Identity Consistency Reward Model: A crucial step is utilizing a specialized reward model (RM) trained on a preference dataset specifically focused on maintaining identity consistency across dynamic, multi-human video sequences [4]. This RM assigns scores, serving as a reward $r(z_0, c)$ for the generated video z_0 conditioned on c. This prevents the policy from collapsing into the "copy-and-paste" issue common in MH-IPV.

2) GRPO Training: The VACE policy π_θ is optimized using Group Relative Policy Optimization (GRPO) to maximize the identity consistency reward [4]. This refinement optimizes the policy to maximize identity consistency metrics and user preference scores over base models [4].

IV. Experiments and Evaluation

We validate VACE-PhysicsRL across its core capabilities: multi-task unification, precise control adherence, physical plausibility, and identity consistency.

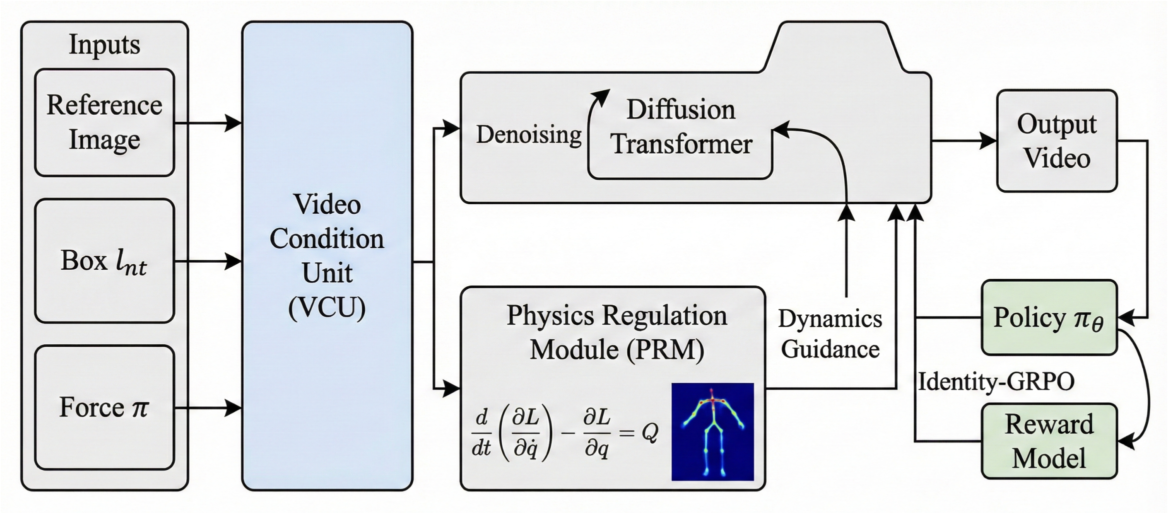

Inputs
Reference Image
Box l_{nt}
Force π
Video Condition Unit (VCU)
Denoising
Diffusion Transformer
Output Video
Physics Regulation Module (PRM)
$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = Q$
Dynamics Guidance
Policy π_θ
Identity-GRPO
Reward Model

Fig. 2: The VACE-PhysicsRL Framework. Multimodal inputs are processed via the VCU. The Physics Regulation Module (PRM) injects dynamics guidance via diffusion adapters, while the policy is refined using Group Relative Policy Optimization (GRPO) for identity preservation.

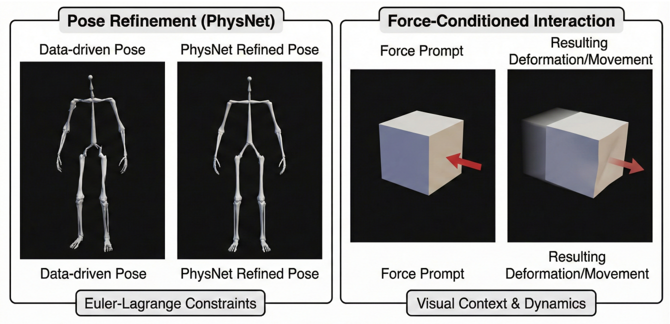

Pose Refinement (PhysNet)
Data-driven Pose
PhysNet Refined Pose
Data-driven Pose
PhysNet Refined Pose
Euler-Lagrange Constraints
Force-Conditioned Interaction
Force Prompt
Resulting Deformation/Movement
Force Prompt
Resulting Deformation/Movement
Visual Context & Dynamics

Fig. 3: Visualization of the Physics Regulation Module. (Left) PhysNet refines raw poses using Euler-Lagrange constraints. (Right) Force prompting enables interactive dynamic control.

A. *Quantitative Metrics*

We utilize diverse metrics appropriate for each specialized task:

- **General Quality:** FVD (Fréchet Video Distance) [**?**], Aesthetic Quality, Imaging Quality.
- **Adherence to Control:** AP (Alignment to Trajectory/Bounding Box) [2], CLIP-V (Appearance Alignment) [2], Force Adherence (measures fidelity to force input) [5].
- **Dynamic Realism:** Motion Smoothness, Motion Realism (measures physical plausibility) [5], N-MPJVE/MPJPE (pose velocity/position error) [3].
- **Identity Consistency:** ID-Consistency score [4], Subject Consistency, Overall Consistency.

B. *Baselines*

We compare VACE-PhysicsRL against:

1) **VACE Base Model:** Original VACE trained only on generic VCU tasks [6].
2) **Controllable Baselines:** FACTOR (for sparse control) [2], MotionPro (for motion) [7].
3) **Physics Baselines:** FinePhys (for pose/kinematics) [3], Force Prompting (for dynamics) [5].
4) **Identity Baselines:** Phantom and VACE-base + Identity-GRPO pipeline [4].

C. *Expected Results*

We expect VACE-PhysicsRL to match VACE baselines on traditional video editing tasks, while significantly outperforming them in specific control metrics. Quantitatively, we anticipate:

- Higher AP and CLIP-V scores than general models like Phenaki or video LDMs.
- Superior ID-Consistency scores and User Study winning rates (e.g., up to 18.9% improvement in ID-Consistency over base models) [4].
- Lower physics-related errors (MPJVE/MPJPE) than data-driven pose methods [3].
- High human preference for motion realism in force-prompted scenarios [5].

V. Conclusion

We successfully introduced VACE-PhysicsRL, a unified video generation framework that leverages expanded VCU inputs and a specialized Context Adapter structure to integrate state-of-the-art physics-aware motion, fine-grained object control, and reinforcement learning alignment. By combining the unifying power of VACE [6] with mechanisms derived from FACTOR [2], FinePhys [3], Force Prompting [5], and Identity-GRPO [4], our model significantly advances the precision and realism of synthesized video content. Future work will focus on explicitly incorporating complex camera control trajectories (as explored in ReCapture) [8] and addressing residual prompt/control misalignment in multimodal scenarios.

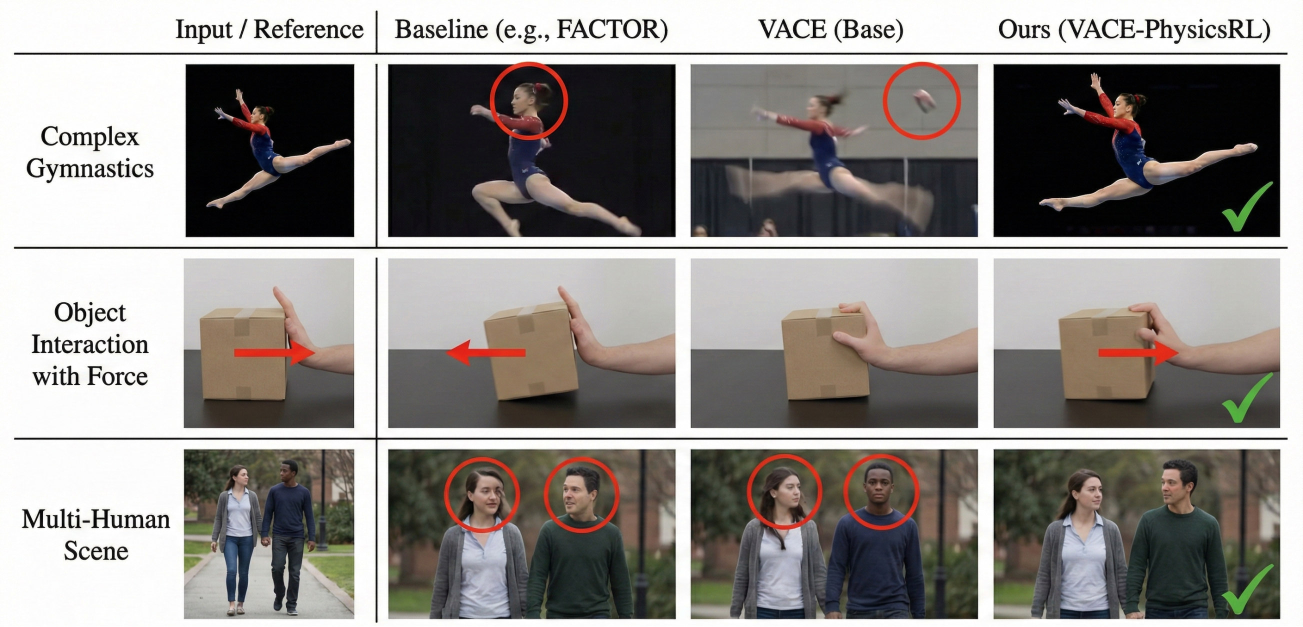
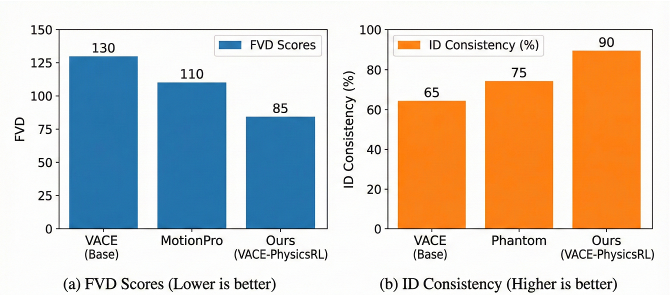


Fig. 4: Qualitative comparison with state-of-the-art baselines. VACE-PhysicsRL (Right) demonstrates superior adherence to trajectory and physical plausibility compared to FACTOR and VACE-Base.



(a) FVD Scores (Lower is better)

(b) ID Consistency (Higher is better)

Fig. 5: Quantitative evaluation metrics comparing VACE-PhysicsRL against state-of-the-art baselines. (a) FVD scores indicating video quality. (b) ID-Consistency scores demonstrating the effectiveness of the GRPO alignment.

References

[1] Z. e. a. Jiang, "VACE: All-in-One Video Creation and Editing," *arXiv preprint arXiv:2503.07598*, 2025, vACE VCU interface details.

[2] H.-P. Huang, Y.-C. Su, D. Sun, L. Jiang, X. Jia, Y. Zhu, and M.-H. Yang, "Fine-grained controllable video generation via object appearance and context (FACTOR)," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2024, focuses on sparse multimodal control via text, bounding boxes, and reference images.

[3] D. Shao, M. Shi, S. Xu, H. Chen, Y. Huang, and B. Wang, "FinePhys: Fine-grained Human Action Generation by Explicitly Incorporating Physical Laws for Effective Skeletal Guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, physics-aware framework using Euler-Lagrange equations for plausible human actions.

[4] X. Meng, Z. Zhang, Z. Zhang, J. Liao, L. Qin, and W. Wang, "Identity-GRPO: Optimizing Multi-Human Identity-preserving Video Generation via Reinforcement Learning," *arXiv preprint arXiv:2506.18244*, 2025, rLHF-driven alignment using Group Relative Policy Optimization (GRPO) for identity consistency.

[5] N. G. Brown, C. Herrmann, M. Freeman, D. Aggarwal, E. Luo, D. Sun, and C. Sun, "Force Prompting: Video Generation Models Can Learn and Generalize Physics-based Control Signals," 2025, introduces local and global physical force vectors as conditioning signals for interaction.

[6] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu, "VACE: All-in-One Video Creation and Editing," *arXiv preprint arXiv:2503.07598*, 2025, vACE baseline, VCU concept, multi-task framework.

[7] Z. Zhang, F. Long, Z. Qiu, Y. Pan, W. Liu, T. Yao, and T. Mei, "MotionPro: A Precise Motion Controller for Image-to-Video Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 27 957–27 967, precise motion control using region-wise trajectory and motion masks for I2V.

[8] D. J. Zhang, R. Paiss, S. Zada, N. Karnad, D. E. Jacobs, Y. Pritch, I. Mosseri, M. Z. Shou, N. Wadhwa, and N. Ruiz, "ReCapture: Generative Video Camera Controls for User-Provided Videos using Masked Video Fine-Tuning," in *CVF Open Access Repository*, 2025, method for generating new videos with novel camera trajectories from existing footage.