ACROSSVAM1.0: PARTICLE WORLD MODELING FOR TEXT-ASSISTED ROBOT VIDEO PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting robot videos requires both precise motion reasoning and preservation of high-frequency appearance, yet monolithic pixel models entangle these objectives and often conceal their progress behind a strong last-frame baseline. We present **AcrossVAM1.0**, a lightweight, text-assisted video action model that factorizes future prediction into object-centric motion and dense appearance. A frozen SAM3-DLP codec decomposes four context frames into semantic particles for the robot, arm, and gripper, together with a background latent. A 0.28M-parameter spatio-temporal Transformer aligns particle identities, rolls their states forward, and is modulated by a frozen OpenCLIP instruction embedding through FiLM. A causal dual-stream decoder combines particle-rendered motion with appearance encoded exclusively from the last observed frame; a residual refiner and learned delivery mask produce five future frames without access to future appearance. On our VRS benchmark constructed from diverse real-robot trajectories, particle dynamics reduce trajectory error by 21.0% over persistence. Across three delivery-mask seeds, AcrossVAM1.0 improves future-frame PSNR/SSIM from 19.97/0.796 to $20.573\pm0.009/0.8004\pm0.0002$, while raw particle generation improves motion-region PSNR from 11.89 to 13.23. The delivered model does not yet beat persistence in LPIPS (0.1304 ± 0.0004 versus 0.122), and correct-versus- shuffled language changes trajectory error by only 2.8–3.1%. We report these limitations alongside oracle, negative-control, multi-seed, and per-robot analyses. The results show that explicit particle dynamics are a promising low-dimensional interface for robot video prediction, while robust language grounding and appearance delivery remain the principal open challenges.

1 INTRODUCTION

A predictive model of how a scene will evolve is useful for robot planning, representation learning, and data generation. Video prediction offers a general interface because it retains the geometry and appearance needed by downstream visual policies (Finn et al., 2016; Ebert et al., 2018; Du et al., 2023). However, real robot videos present an uncomfortable combination: most pixels are static, small articulated parts determine the action, and the same natural-language instruction can admit visually distinct trajectories. A model can therefore achieve competitive global image metrics by copying the last frame, without learning the motion that matters.

Contemporary world models typically compress the whole frame into a dense latent and learn its temporal evolution (Hafner et al., 2023; Babaeizadeh et al., 2022; Huang et al., 2026). This representation is effective but leaves correspondence, dynamics, and texture synthesis coupled. Object-centric models instead represent a scene as slots or particles (Locatello et al., 2020; Wu et al., 2023; Daniel & Tamar, 2024; Daniel et al., 2026). Their explicit state makes motion easier to inspect, but a compact object bottleneck cannot reconstruct fine texture by itself. Moreover, conditioning a video model on language does not guarantee that its trajectory depends on the instruction.

AcrossVAM1.0 assigns different responsibilities to different representations (Figure 1). Semantic particles describe the geometry of the whole robot, arm, and gripper; a compact Transformer predicts their future states; and a separate causal appearance stream restores detail. Language is injected only into the dynamics through feature-wise modulation, which makes its effect measurable at the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

trajectory level. The dense branch observes the last context frame but never a future frame, so future motion cannot leak through the appearance path.

Our contributions are:

- We introduce a semantically anchored particle world model that reduces real-robot video dynamics to a compact trajectory problem. Its dynamics core has only 0.28M trainable parameters and exposes part-level motion for diagnosis.
- We propose a causal motion–appearance decoder that combines predicted particle geometry with context-only dense appearance and explicitly controls the trade-off between sharp static regions and moving-object fidelity.
- We evaluate against persistence, shuffled-language and random controls, oracle particles, multiple seeds, and per-robot splits. The evaluation identifies both a reliable motion benefit and the still-limited degree of language grounding.

2 Related Work

Video prediction for physical interaction. Video prediction has long served as a test bed for learning physical regularities without manual state annotation. Early action-conditioned models predicted future pixels from images and robot controls, demonstrating that visual foresight can support planning and representation learning (Finn et al., 2016; Ebert et al., 2018). This line of work is especially relevant to manipulation because the prediction target preserves object geometry, contact events, and scene appearance in a common observation space. It also exposes a characteristic evaluation problem: robot videos contain large static regions, so copying the last context frame can score well even when articulated motion is wrong. Recent text-guided approaches instead use language to describe a desired evolution and treat predicted video as a policy-like interface (Du et al., 2023). AcrossVAM1.0 follows this conditional formulation but explicitly reports both global image quality and motion-region quality, making improvement over persistence harder to obtain by static copying.

Latent and diffusion world models. Dense video predictors have progressively combined stochastic latent variables, recurrent state, and learned feature spaces to model uncertainty while reducing the cost of direct pixel prediction. FitVid shows that a carefully regularized latent video model remains a strong reference for stochastic prediction (Babaeizadeh et al., 2022). Learned latent world models can also optimize control-relevant dynamics without reconstructing every observation at every step (Hafner et al., 2023). Diffusion models provide a complementary route to multimodal generation (Ho et al., 2020); Diffusion Forcing extends this view by training sequence models with independently corrupted tokens and flexible conditioning patterns (Chen et al., 2024). These methods obtain expressive dense representations, but motion, correspondence, and high-frequency appearance typically remain entangled in the same latent process. Nano World Models ask how much predictive structure can be retained in deliberately small models (Huang et al., 2026). AcrossVAM1.0 shares this efficiency objective while placing a semantic particle bottleneck between perception and rendering: a small dynamics core predicts geometry, whereas a separate causal stream supplies texture from observed context.

Object-centric representation and dynamics. Object-centric learning decomposes a scene into reusable entities rather than one global feature map. Slot Attention learns exchangeable component representations from images (Locatello et al., 2020), and SAVi++ extends slot-based perception to video with temporal propagation and stronger supervision signals (Elsayed et al., 2022). SlotFormer subsequently predicts future slot states with a Transformer, showing that temporal reasoning can be separated from the decoder (Wu et al., 2023). Particle representations make the spatial state more explicit: DDLP associates appearance with position, scale, depth, and presence, while later latent-particle models study long-horizon dynamics in this structured space (Daniel & Tamar, 2024; Daniel et al., 2026). The benefit is interpretability and a compact motion interface; the cost is that exchangeable entities require temporal correspondence and may discard texture needed for photorealistic rendering. Set prediction commonly addresses permutation ambiguity with bipartite assignment (Carion et al., 2020). AcrossVAM1.0 specializes these ideas to articulated robots: VRS supervision anchors whole-robot, arm, and gripper particles, Hungarian matching supports unordered proposals, and a dense residual code restores appearance that should not be forced through the low-dimensional

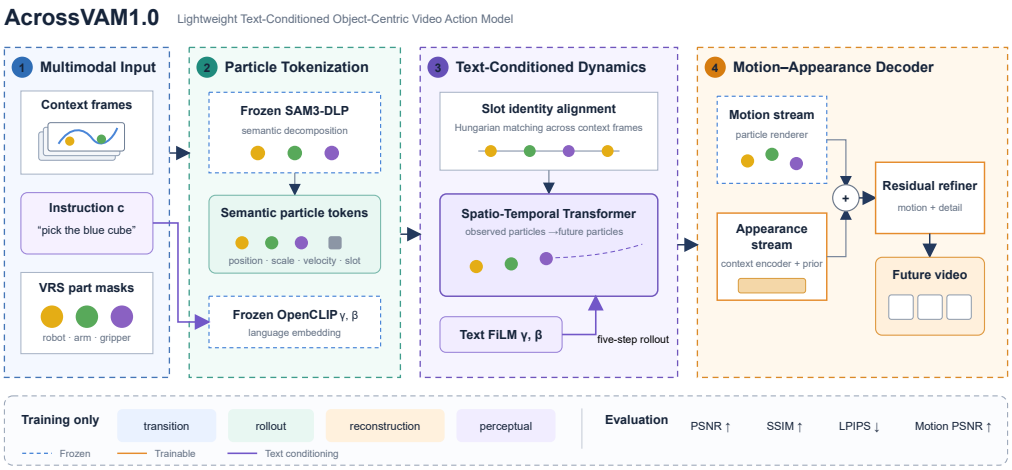

Figure 1: **AcrossVAM1.0 overview.** Four context frames and an instruction are encoded into semantic robot particles and a background latent. Identity-aligned, text-modulated particle dynamics predict five future states. A frozen renderer produces motion structure, while a context-only dense stream supplies appearance; the residual delivery module fuses both predictions. Snowflakes and flames denote frozen and trainable modules. Dashed gray paths are training-only.

motion state. Recent extensions sharpen both axes: OCK explicitly augments object slots with position, velocity, and acceleration (Song et al., 2025), while TextOCVP uses a text-conditioned slot Transformer for language-guided prediction (Villar-Corrales et al., 2026).

Foundation perception and language conditioning. Large frozen encoders can provide semantic structure without requiring an equally large trainable predictor. Promptable segmentation models such as SAM and SAM2 transfer object masks across diverse scenes and video (Kirillov et al., 2023; Ravi et al., 2024). CLIP and OpenCLIP provide transferable text representations aligned with visual concepts (Radford et al., 2021; Cherti et al., 2023), while FiLM injects a conditioning signal through feature-wise affine modulation (Perez et al., 2018). These components make language conditioning inexpensive, but their presence alone does not demonstrate that a predicted trajectory depends on the instruction: a model may rely on scene priors, dataset bias, or the observed motion prefix. We therefore freeze perception and language towers, apply FiLM only to particle dynamics, and evaluate correct text against shuffled, pseudo-text, and random-label controls. This design separates semantic perception from causal language use and motivates our conservative term *text-assisted* rather than instruction-controlled.

Positioning of AcrossVAM1.0. Prior work generally emphasizes either dense visual fidelity, compact latent dynamics, or object-centric interpretability. AcrossVAM1.0 combines these aims by assigning them to distinct modules: frozen semantic perception defines particles, a 0.28M-parameter Transformer models their motion, and a context-only appearance path performs residual delivery. The closest comparisons are therefore complementary rather than interchangeable: dense predictors test image quality, slot and particle models test structured dynamics, and language-conditioned video models test semantic responsiveness. Our evaluation mirrors this decomposition with persistence, motion-region, particle-oracle, language-control, multi-seed, and cross-robot diagnostics.

3 METHOD

3.1 PROBLEM FORMULATION

Let $\boldsymbol{x}_{1:C} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_C)$ be context RGB frames and c a natural-language instruction. We predict $\hat{\boldsymbol{x}}_{C+1:T}$, with $C = 4$ and $T = 9$. AcrossVAM1.0 factorizes the conditional distribution as

$$p(\boldsymbol{x}_{C+1:T} \mid \boldsymbol{x}_{1:C}, c) = \int p_\phi(\boldsymbol{x}_{C+1:T} \mid \boldsymbol{z}_{C+1:T}, \boldsymbol{x}_C) p_\theta(\boldsymbol{z}_{C+1:T} \mid \boldsymbol{z}_{1:C}, c)\, d\boldsymbol{z}, \tag{1}$$

where $\boldsymbol{z}$ is an object-centric state. Dynamics p_θ is responsible for motion; decoder p_ϕ renders motion and context appearance.

3.2 FROZEN SEMANTIC PARTICLE CODEC

For each frame, a frozen SAM3-DLP encoder receives RGB and VRS part masks and extracts $K = 3$ semantic particles, corresponding to the whole robot, arm, and gripper, plus a spatial background latent $\boldsymbol{b}_t$. Particle i is

$$\boldsymbol{z}_t^i = [\boldsymbol{p}_t^i, \boldsymbol{s}_t^i, d_t^i, q_t^i, \boldsymbol{a}_t^i], \tag{2}$$

where $\boldsymbol{p} \in [0,1]^2$ is position, $\boldsymbol{s} \in \mathbb{R}_+^2$ is scale, d is depth, q is presence, and $\boldsymbol{a}$ is appearance. Explicit geometry forms the prediction target; appearance and background serve the renderer. Freezing the codec prevents the representation from drifting toward an easy pixel-copying solution.

Identity alignment. Unordered proposals require temporal correspondence. For adjacent frames, we form

$$C_{ij} = \lambda_p \|\bar{\boldsymbol{p}}_{t-1}^i - \bar{\boldsymbol{p}}_t^j\|_2 + \lambda_a \left(1 - \cos(\boldsymbol{a}_{t-1}^i, \boldsymbol{a}_t^j)\right) \tag{3}$$

and obtain $\pi_t = \arg\min_\pi \sum_i C_{i,\pi(i)}$ using Hungarian matching. The bar denotes scale-normalized coordinates. In our offline VRS cache, fixed semantic part identifiers already determine the permutation; we bypass matching there and retain it for proposal-based inference.

3.3 TEXT-ASSISTED PARTICLE DYNAMICS

The dynamics input concatenates geometry, first-order velocities $\Delta\boldsymbol{p}_t^i, \Delta\boldsymbol{s}_t^i$, depth, presence, and learned slot and time embeddings. A causal spatio-temporal Transformer (Vaswani et al., 2017) alternates temporal reasoning within each slot with interactions among robot parts. It predicts residual updates autoregressively:

$$\hat{\boldsymbol{y}}_{t+1}^i = \hat{\boldsymbol{y}}_t^i + f_\theta(\hat{\boldsymbol{y}}_{1:t}^{1:K}, e_c)^i, \qquad \boldsymbol{y} = [\boldsymbol{p}, \boldsymbol{s}, d, q]. \tag{4}$$

Here e_c is produced by a frozen OpenCLIP ConvNeXt-L/320 text tower. At block ℓ, trainable FiLM maps inject instruction semantics:

$$\tilde{\boldsymbol{h}}_\ell = (1 + \gamma_\ell(e_c)) \odot \mathrm{LN}(\boldsymbol{h}_\ell) + \beta_\ell(e_c). \tag{5}$$

FiLM projections are initialized to zero, so optimization starts from unconditioned dynamics rather than a randomly perturbed model. We combine one-step transition and autoregressive rollout supervision:

$$\mathcal{L}_{\mathrm{dyn}} = \lambda_{\mathrm{tr}} \sum_{t=C}^{T-1} \|\hat{\boldsymbol{y}}_{t+1}^{\mathrm{TF}} - \boldsymbol{y}_{t+1}\|_1 + \lambda_{\mathrm{ro}} \sum_{t=C+1}^{T} \|\hat{\boldsymbol{y}}_t^{\mathrm{AR}} - \boldsymbol{y}_t\|_1. \tag{6}$$

3.4 CAUSAL MOTION–APPEARANCE DECODING

The frozen particle renderer maps predicted particles and context background to a structural RGB proposal $\boldsymbol{r}_t$ and alpha map $\boldsymbol{\alpha}_t$. This proposal moves the correct parts but lacks high-frequency detail. In parallel, a dense encoder extracts $\boldsymbol{h}_C = E_{\mathrm{app}}(\boldsymbol{x}_C)$ from the *last context frame only*. A conditional residual prior predicts future detail from context appearance and particle motion:

$$\hat{\boldsymbol{h}}_t = P_\psi(\boldsymbol{h}_C, \boldsymbol{r}_C, \boldsymbol{\alpha}_C, \boldsymbol{r}_t, \boldsymbol{\alpha}_t, t - C). \tag{7}$$

A residual refiner synthesizes $\boldsymbol{g}_t = R_\phi(\boldsymbol{r}_t, \boldsymbol{\alpha}_t, \hat{\boldsymbol{h}}_t, \boldsymbol{x}_C)$. Finally, a learned confidence mask $\boldsymbol{m}_t \in [0,1]^{H\times W}$ blends generation with persistence:

$$\hat{\boldsymbol{x}}_t = \boldsymbol{m}_t \odot \boldsymbol{g}_t + (1 - \boldsymbol{m}_t) \odot \boldsymbol{x}_C. \tag{8}$$

The mask uses inference-time quantities such as predicted alpha motion and candidate disagreement. The appearance loss is

$$\mathcal{L}_{\text{app}} = \lambda_1 \|\boldsymbol{g}_t - \boldsymbol{x}_t\|_1 + \lambda_s(1 - \text{SSIM}(\boldsymbol{g}_t, \boldsymbol{x}_t)) + \lambda_p \,\text{LPIPS}(\boldsymbol{g}_t, \boldsymbol{x}_t), \tag{9}$$

with extra motion-region weighting during training. Future RGB, masks, and residual codes are targets only. At inference every decoder input is a function of $(\boldsymbol{x}_{1:C}, c, \hat{\boldsymbol{z}}_{C+1:T})$; the dense branch cannot leak future appearance.

4 Experiments

4.1 Setup

Data and protocol. We construct the Video Robot Segmentation (VRS) benchmark from diverse DROID-style real-robot trajectories (Khazatsky et al., 2024), augmented with masks for whole robot, arm, and gripper and a natural-language command. The manifest contains 2,746 labeled clips: 2,526/106/105 are assigned to train/validation/test and nine records are excluded from the fixed protocol. We resize frames to 128×128, sample nine-frame windows at stride three, observe four frames, and predict five. The future-video test contains 92 clips and 460 predicted frames; codec reconstruction uses 48 clips and 432 frames; trajectory diagnostics use 368 held-out windows. Platforms include Franka, WidowX, Mobile ALOHA, Google Robot, Fanuc, Sawyer, Stretch, UR5, Kuka IIWA, and xArm.

Metrics and selection. We report PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) over all future pixels, plus PSNR in a dilated motion/foreground region. Checkpoints and blending thresholds are selected on validation data; test is evaluated once. Persistence copies $\boldsymbol{x}_C$. The particle oracle uses ground-truth future particles, and the pixel oracle chooses generation or persistence per pixel; both are upper bounds, not deployable methods.

Implementation. The particle Transformer has width 64, two layers, four heads, and 0.28M parameters. The complete trainable future-prediction stack has approximately 9.5M parameters, excluding frozen SAM3-DLP and OpenCLIP backbones. We train codec, dynamics, appearance, and delivery modules in stages. Appendix A gives details.

4.2 Future video prediction

Table 1 exposes the central trade-off. Raw particle-conditioned generation improves motion PSNR by 1.34 dB over persistence, but degrades global SSIM because it modifies easy static pixels. Residual delivery recovers static fidelity and improves global PSNR by 0.61 dB and SSIM by 0.0044 while retaining most of the motion advantage. Dispersion across delivery-mask seeds is small (0.009 dB PSNR and 0.0002 SSIM), but this sweep shares all upstream checkpoints and therefore measures delivery robustness rather than full end-to-end uncertainty. LPIPS remains 0.0082 worse than persistence. The 1.55 dB oracle gap indicates that candidate quality is not the only bottleneck: deciding where to synthesize remains imperfect.

Recent public baselines. We surveyed public methods released in the past five years: the dense stochastic predictor FitVid, slot-based SlotFormer, particle-based DDLP, slot-based TextOCVP, and particle-based LPWM. All provide public implementations, but none reports the fixed VRS split, four-frame context, five-frame horizon, and motion-region metric used here. We therefore separate measured results (Table 1) from the reproduction queue (Table 2) rather than mixing incommensurate numbers from BAIR, CLEVRER, OBJ3D, or LanguageTable. Reproduction follows the priority LPWM-Language, TextOCVP, FitVid, DDLP, and SlotFormer.

Table 1: Future prediction on 92 held-out clips (460 future frames). AcrossVAM1.0 reports mean ± sample standard deviation over three independently initialized delivery masks; upstream dynamics, appearance, and diffusion-prior checkpoints are shared. Only results evaluated on the identical VRS protocol are included; scores reported on other datasets are not transplanted. Bold marks the best evaluated deployable result.

Method	PSNR↑	SSIM↑	LPIPS↓	Motion PSNR↑
Persistence	19.97	0.796	**0.122**	11.89
Particle-conditioned generation	19.93	0.763	0.138	**13.23**
+ fixed particle-alpha delivery	20.17	0.785	0.134	12.93
AcrossVAM1.0 DiffPrior (3 seeds)	**20.573±0.009**	**0.8004±0.0002**	0.1304±0.0004	12.870±0.032
Per-pixel selection oracle	22.12	0.833	0.114	14.09

Table 2: Public-baseline reproduction protocol. This table tracks comparability and execution status, not quantitative results. Every method will be retrained on the fixed VRS split with four context and five future frames and evaluated on the same 92 clips and 460 future frames as Table 1.

Priority	Method	Published	Conditioning	Structured state / status
1	LPWM-Language (Daniel et al., 2026)	2026	text	particles / in progress
2	TextOCVP (Villar-Corrales et al., 2026)	2026	text	slots / queued
3	FitVid (Babaeizadeh et al., 2022)	2022	frames	dense latent / queued
4	DDLP (Daniel & Tamar, 2024)	2024	frames	particles / queued
5	SlotFormer (Wu et al., 2023)	2023	frames	slots / queued

4.3 Qualitative behavior and delivery robustness

The examples in Figure 2 illustrate why global and motion metrics must be read together. Raw generation changes the articulated region but also perturbs static texture. The blend mask localizes many edits, yet errors remain near thin links, occlusion boundaries, and newly revealed pixels. The qualitative failure pattern is consistent with the gap to the pixel-selection oracle.

The earlier DensePrior variant retains a 0.50 dB average PSNR gain over persistence and a 1.02 dB motion-region gain, with low dispersion. SSIM is statistically near the persistence baseline and LPIPS remains worse. One seed selected its initialization under the validation rule, further indicating that appearance refinement is less stable than particle dynamics.

The stronger DiffPrior delivery is similarly stable across mask initialization: all three seeds improve over persistence in PSNR and SSIM, and the selected checkpoints occur at 1,500–3,000 steps (Table 4). Its motion-region PSNR remains 0.98 dB above persistence but 0.36 dB below raw generation, quantifying the cost of suppressing changes outside the predicted delivery region. Because the upstream particle and appearance models remain fixed, we report this result as a delivery-stage sweep rather than implying full end-to-end multi-seed evidence.

4.4 Representation and dynamics

Adding a $32 \times 32 \times 4$ residual code raises reconstruction PSNR by 7.01 dB (Table 5), motivating a dual stream rather than forcing texture through object state. The small dynamics model reliably beats trajectory persistence across seeds. Diagnostics show that residualized box coordinates and raw appearance features introduce shortcuts; removing them and using stride-three motion raises error reduction from low single digits to 21.0%.

4.5 Does the model use language?

The positive multi-seed gap in Table 6, together with near-zero gaps for kinematic pseudo-text and random labels, suggests that FiLM captures a real semantic signal. The effect is nevertheless small and below our 10% target. We therefore describe AcrossVAM1.0 as *text-assisted*, not instruction-controlled. The lower verb recall of automatically generated captions also explains why a larger language encoder does not automatically produce a larger causal effect.

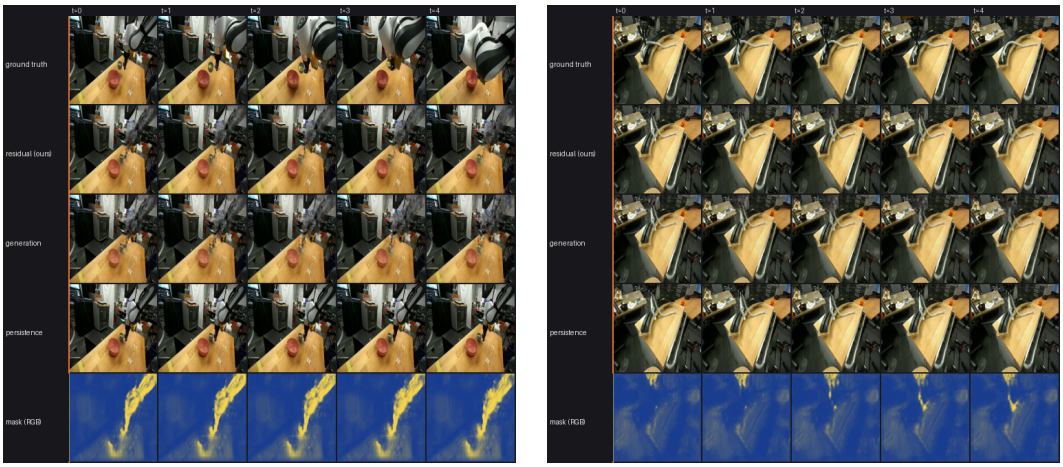

Figure 2: **Representative held-out Franka predictions.** Each panel shows five future steps; rows are ground truth, residual delivery, raw generation, persistence, and the learned blend mask visualized with a blue-to-yellow RGB colormap. The left example has a compact interaction region, whereas the right example contains heavier clutter and occlusion. These cases are selected from the fixed seed-0 demo set after checkpoint selection; they are not used for quantitative model selection.

Table 3: Earlier DensePrior delivery-stage robustness over three appearance/refinement seeds on the same 460-frame test set. The shared particle dynamics checkpoint is frozen, so this is not a full end-to-end seed sweep. Values are mean ± sample standard deviation.

Method	PSNR↑	SSIM↑	LPIPS↓	Motion PSNR↑
Persistence	19.966	0.796	**0.122**	11.894
Raw DensePrior generation	19.540 ± 0.274	0.718 ± 0.018	0.160 ± 0.011	**13.057** ± 0.077
AcrossVAM1.0 joint residual	**20.464** ± 0.112	0.795 ± 0.005	0.129 ± 0.001	12.911 ± 0.082
Pixel-selection oracle	21.940 ± 0.071	0.830 ± 0.001	0.112 ± 0.001	13.907 ± 0.074

Cross-robot behavior. The language-conditioned particle model improves on Franka (+21.1%, 136 windows), WidowX (+19.0%, 68), and UR5 (+18.4%, 8), but degrades on Mobile ALOHA (−16.1%, 60), Google Robot (−5.2%, 32), and several small subsets. The non-Franka aggregate remains positive (+10.0%, 232 windows), but strict leave-one-robot-out Mobile ALOHA experiments do not beat persistence.

4.6 Ablations and Interpretation

Table 7 summarizes the component and delivery ablations. Raw geometric state is preferable to codec box residuals: the latter exhibit a 34.4% velocity sign-flip rate and obscure physical displacement. Stride three makes meaningful motion visible at 128-pixel resolution. Appearance is useful for rendering but harmful as a dynamics shortcut. On the delivery side, fixed particle alpha improves PSNR by 0.23 dB over raw generation, whereas the learned causal mask adds another 0.41 dB and restores SSIM above persistence. This improvement trades 0.36 dB of motion PSNR for substantially better static-region fidelity; the oracle shows that mask calibration still leaves 1.55 dB of PSNR headroom. Table 6 provides the complementary language ablation. Slot matching has negligible effect in the fixed-ID cache, as expected, but remains necessary for unordered proposals.

5 Limitations

AcrossVAM1.0 is not yet a general-purpose action model. It predicts only five future frames at 128^2 resolution and uses part masks during codec construction. The frozen OpenCLIP tower is large even though the trainable dynamics are small, so *lightweight* refers to the learned predictive core rather than total deployment memory. Global gains over persistence are modest, public-model baselines

Table 4: Exact DiffPrior delivery sweep. Each learned mask is independently initialized and selected by validation SSIM; the test set is evaluated once after selection. Particle rollouts, appearance features, and DiffPrior candidates are shared across seeds.

Seed	Selected step	PSNR↑	SSIM↑	LPIPS↓	Motion PSNR↑
0	1,500	20.563	0.8006	0.1305	12.833
1	3,000	20.578	0.8004	0.1299	12.886
2	1,500	20.578	0.8002	0.1307	12.890
Mean ± s.d.	–	20.573±0.009	0.8004±0.0002	0.1304±0.0004	12.870±0.032

Table 5: Representation and dynamics diagnostics. Float counts measure the extra dense code per frame; improvement is trajectory-error reduction over persistence.

Representation	Floats	PSNR↑	SSIM↑	LPIPS↓
Base particle codec	160	20.94	0.637	0.362
Generic dense code	1,024	23.62	0.747	0.310
Particle + residual code	4,096	**27.95**	**0.877**	**0.148**

Dynamics model	Val improvement	Test improvement
Width-64, 2-layer particle Transformer	14.1%	**21.0%**
three-seed test range	–	[20.6%, 21.4%]

are not yet retrained on the identical VRS split, LPIPS is worse, language sensitivity is limited, and cross-embodiment generalization is inconsistent. The pixel oracle also reveals delivery calibration as a major bottleneck. Future work should learn uncertainty-aware blending, collect counterfactual instruction pairs, replace mask supervision with proposal-time segmentation, and evaluate closed-loop control rather than video metrics alone.

6 Conclusion

We introduced AcrossVAM1.0, a compact text-assisted world model that treats robot video prediction as semantic particle dynamics plus causal appearance synthesis. The model improves trajectories and motion-region prediction while residual delivery recovers much of the static visual fidelity. Strong controls show both what works—low-dimensional motion modeling—and what remains unresolved—perceptual delivery, language dependence, and cross-robot transfer. This separation provides a foundation for interpretable video action models without hiding failure modes behind aggregate image metrics.

Table 6: Language controls. Gap is the relative trajectory-error increase under shuffled rather than correct text. Values are means over three seeds.

Text source	Correct–shuffled gap↑	Seed range	Verb recall
Source annotation	2.8%	[1.5, 4.9]	92.9%
Rule-corrected annotation	**3.1%**	[2.3, 4.2]	92.9%
Vision-language caption	2.9%	[1.2, 4.0]	47.7%
Kinematic pseudo-text	0.3%	–	–
Random label control	0.2%	–	–

Table 7: Component ablations. Top: held-out particle trajectory-error reduction over persistence. Bottom: delivery alternatives on the same 92-clip/460-frame future-video test as Table 1.

Configuration	Improvement↑
Residualized boxes + appearance in dynamics	3.9%
Clean boxes, stride 1	4.2%
Clean boxes, stride 3	15.1%
Clean boxes, no appearance shortcut	18.5%
Final width-64, 2-layer model	**21.0%**

Delivery configuration	PSNR↑	SSIM↑	LPIPS↓	Motion PSNR↑
No delivery (raw generation)	19.935	0.7634	0.1381	**13.227**
Fixed particle-alpha mask	20.168	0.7850	0.1338	12.931
Learned causal mask	**20.573**	**0.8004**	**0.1304**	12.870
Pixel-selection oracle	22.121	0.8327	0.1140	14.092

REFERENCES

Mohammad Babaeizadeh, Mohammad Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. In *International Conference on Learning Representations*, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tal Daniel and Aviv Tamar. Deep dynamic latent particles for unsupervised object-centric video prediction. *Transactions on Machine Learning Research*, 2024.

Tal Daniel, Carl Qi, Dan Haramati, Amir Zadeh, Chuan Li, Aviv Tamar, Deepak Pathak, and David Held. Latent particle world models: Self-supervised object-centric stochastic dynamics modeling. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lTaPtGiUUc.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

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

Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Thomas Kipf, and Michael C. Mozer. Savi++: Towards end-to-end object-centric learning from real-world videos. *arXiv preprint arXiv:2206.07764*, 2022.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, 2016.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Siqiao Huang, Partha Kaushik, Michael Chen, Hengkai Pan, Kaiwen Geng, Omar Chehab, Fernando Moreno-Pino, and Max Simchowitz. Nano world models: A minimalist approach to future video prediction. *arXiv preprint arXiv:2605.23993*, 2026.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Srirama, Lawrence Chen, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Yeon-Ji Song, Jaein Kim, Suhyung Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. Ock: Unsupervised dynamic video prediction with object-centric kinematics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11359–11368, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Angel Villar-Corrales, Gjergj Plepi, and Sven Behnke. TextOCVP: Object-centric video prediction with language guidance. *Transactions on Machine Learning Research*, 2026.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. In *International Conference on Learning Representations*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

A Reproducibility Details

Fixed protocol. All model selection uses 2,526 training and 106 validation clips. The 105 test-manifest clips are never used for checkpoint or threshold selection. Stages contain fewer clips when they require complete RGB, mask, particle-cache, and text records; each main table reports the effective sample count.

Optimization stages. We (i) train and freeze the particle codec; (ii) cache particles; (iii) train transition and rollout dynamics; (iv) train the dense prior and residual refiner using rollout particles; and (v) train the inference-safe blend mask. The dynamics configuration has dimension 64, two layers, four heads, nine time steps, four context steps, and three slots. Future prediction is autoregressive.

Leakage audit. The appearance encoder is evaluated once on $\boldsymbol{x}_C$. Future RGB, masks, oracle particles, and detail codes are removed from inference inputs and used only as targets or upper bounds. Text shuffling occurs within the evaluation split.

B Additional Per-Robot Results

Table 8: Per-robot trajectory diagnostics. Positive is better than persistence.

Robot	Windows	Improvement
Franka	136	+21.1%
WidowX	68	+19.0%
Mobile ALOHA	60	−16.1%
Google Robot	32	−5.2%
Fanuc Mate	28	−9.0%
Sawyer	20	−12.9%
UR5	8	+18.4%
All non-Franka	232	+10.0%

C Broader Impact

Video world models can support data-efficient robot learning and safer offline analysis, but inaccurate rollouts can encourage unsafe actions if treated as physical guarantees. AcrossVAM1.0 is a research prototype and should not be used as a safety controller. Its data may encode embodiment, laboratory, and annotation biases. Part-level visualization can expose failures, but does not eliminate the need for real-world validation and appropriate safety constraints.