HORIZONLORA: DISTILLING MOTION ADAPTER BANKS FOR TRAJECTORY-CONDITIONED VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon video world models must accommodate changing camera trajectories while reusing a large frozen visual prior. Static LoRA applies one update throughout a rollout, whereas motion-specialized adapter banks retain multiple sets of expert weights. We propose HorizonLoRA, which distills a motion adapter bank into a shared dictionary of normalized rank-one updates and predicts bounded, chunk-specific coefficients from the initial frame, text prompt, ordered camera controls, cumulative commanded pose, and rollout time. A two-stage procedure first fits the dictionary to individual teachers and their mixtures, then trains the coefficient generator with video supervision and Gram-space parameter distillation. The teacher bank is discarded at deployment. We compare five adaptation strategies under 1-chunk, 4-chunk, and 16-chunk recursive rollouts. In the reported comparison, HorizonLoRA reduces deployed adapter storage by 72.4% relative to the learned bank and records lower 16-chunk similarity-aligned trajectory error on the evaluated Sekai and DL3DV samples. Other adapters obtain stronger appearance or boundary-continuity scores. These results characterize the geometry–appearance–storage trade-off of teacher-distilled conditional adaptation for recursive video generation.

1 INTRODUCTION

Video world models predict future visual observations from an initial scene, a semantic prompt, and a sequence of controls. Their practical value depends not only on one generated clip, but also on how scene appearance and camera response evolve when each generated chunk becomes future context. Across a long rollout, camera direction, speed, cumulative pose, and rollout depth all change. This recursive setting calls for adaptation that responds to the requested path and state without creating another feedback loop through synthesized frames.

LoRA (Hu et al., 2022) specializes a large frozen backbone through low-rank updates. A static LoRA applies the same update across scenes, prompts, trajectories, and rollout depths; increasing rank expands capacity without introducing condition dependence. This motivates learning not only reusable update directions, but also a rule for combining them as the requested path evolves.

DimensionX motivates specialized LoRA directors for distinct camera motions (Sun et al., 2024). A bank-based deployment selects or merges motion specialists, retaining their full weights even when each request uses only a combination of them. We construct a *DimensionX-style motion-specialist bank* on the same frozen backbone and use it as offline supervision for a shared representation. The research question is whether the bank's update directions can be consolidated into one dictionary while their condition-dependent combinations are generated at runtime. This is a comparison of adapter representations, rather than a reproduction of the complete DimensionX system.

HorizonLoRA maps the initial scene, semantic prompt, cumulative commanded camera state, ordered future controls, and rollout time to chunk-specific low-rank updates. The formulation takes a first frame, prompt, camera trajectory, and timing as input. The name reflects adaptation conditioned on rollout progress and cumulative commanded state: the same local control can be issued at different points in a recursive sequence. Shared directions remain fixed after training, while their coefficients can change from chunk to chunk.

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.

We instantiate the frozen Base-WM with YUME (Mao et al., 2025) and build a multi-domain adapter corpus from Sekai walking videos (Li et al., 2025a) and DL3DV-480P-1K, a 1000-scene subset of DL3DV (Ling et al., 2024). We split each domain independently at the source-video or scene level and then merge corresponding training, validation, and test partitions. The two held-out domains are reported separately. Adapter-level source exclusion does not establish exclusion from backbone pretraining, so neither domain is presented as a verified unseen-domain test.

We evaluate this representation against Frozen Base-WM, Static LoRA, a learned soft director bank, and Direct HyperLoRA. The learned bank is the primary storage comparator because its specialists provide offline supervision for HorizonLoRA. In the reported comparison, HorizonLoRA stores 72.4% fewer adapter bytes and records lower 16-chunk aligned trajectory error than that bank, whereas the bank and direct hypernetwork obtain stronger appearance or boundary scores. We report all five methods and all three horizons to expose these trade-offs. Separate diagnostics distinguish fitting a denoising target from responding correctly to a signed motion command.

Our contributions are threefold:

- **Teacher-informed shared adaptation.** We distill motion-specialized LoRA teachers into a shared rank-one dictionary, separating reusable update directions from condition-dependent coefficients.
- **Chunk-wise conditional synthesis.** We formulate a coefficient generator conditioned on the initial scene, text, ordered camera controls, cumulative commanded pose, and rollout time, without feeding generated frames into the adapter encoder.
- **Recursive comparison across adaptation strategies.** We compare five adapter configurations across three rollout horizons and report trajectory geometry, appearance, boundary continuity, deployed adapter footprint, and recorded computational costs.

2 Problem Formulation

The user input is

$$\mathcal{I} = (I_0, p, \mathcal{U}, \Delta\mathcal{T}), \tag{1}$$

where I_0 is the first frame, p is the text prompt, $\mathcal{U}$ is the full camera-control sequence, and $\Delta\mathcal{T}$ contains the associated time intervals. Let t index non-overlapping rollout chunks and let $u_t = (u_{t,1}, \ldots, u_{t,H})$ contain H controls. Each control is

$$u_{t,i} = (\delta T_{t,i}, \Delta t_{t,i}), \qquad \delta T_{t,i} \in SE(3), \tag{2}$$

where $\Delta t_{t,i}$ determines timing and speed. A frozen video world model F_θ recursively predicts

$$\hat{x}_t = F_\theta(x_t^{\mathrm{ctx}}, p, u_t; \Delta\theta_t), \qquad x_{t+1}^{\mathrm{ctx}} = \mathcal{C}(x_t^{\mathrm{ctx}}, \hat{x}_t), \tag{3}$$

where $\mathcal{C}$ appends the generated chunk and retains the context window required by Base-WM. Here u_t denotes the requested control, not an assumed native continuous-pose port: YUME realizes camera instructions through action-bearing text, while our adapter encoder receives numerical trajectory features. Generated frames affect future generation through the backbone context, not through the adapter condition encoder.

The adapter condition path receives no generated frame. Its condition is

$$c_t = \left(I_0, p, T_{0\to t}^{\mathrm{cmd}}, u_t, \Delta t_t, e_{\mathrm{time}}(t)\right), \tag{4}$$

where $T_{0\to t}^{\mathrm{cmd}}$ is obtained by composing all commanded increments before chunk t, $\Delta t_t = (\Delta t_{t,1}, \ldots, \Delta t_{t,H})$, and $e_{\mathrm{time}}(t)$ encodes chunk index and elapsed time. The cumulative state distinguishes identical local controls issued at different rollout depths and viewpoints without estimating pose from generated frames.

We group the first-frame, prompt, trajectory, cumulative-state, and time encoders into v_{ϕ_1} and write the layer-aware parameter generator as g_{ϕ_2}. The dynamic-adaptation path is

$$z_t = v_{\phi_1}\left(I_0, p, T_{0\to t}^{\mathrm{cmd}}, u_t, \Delta t_t, e_{\mathrm{time}}(t)\right), \tag{5}$$

$$\Gamma_t = g_{\phi_2}(z_t, \{p_\ell\}_{\ell\in\mathcal{S}}), \tag{6}$$

$$\Delta\theta_t = \mathcal{A}(\Gamma_t; \mathcal{D}), \tag{7}$$

$$\hat{x}_t = F_\theta\left(x_t^{\mathrm{ctx}}, p, u_t; \Delta\theta_t\right). \tag{8}$$

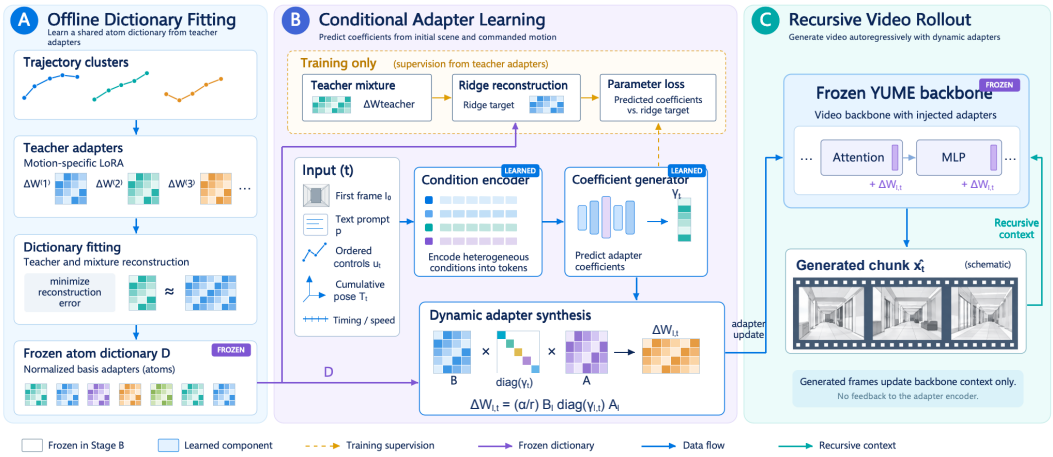


Figure 1: Overview of HorizonLoRA. (a) Stage A fits and freezes a normalized atom dictionary from trajectory-clustered LoRA teachers. (b) Stage B uses the dictionary for ridge reconstruction and dynamic adapter synthesis; the generator predicts coefficients from the first frame, prompt, ordered controls, cumulative pose, and timing. Dashed links denote training supervision. The synthesis includes the global LoRA scale α/r. Corridor thumbnails are illustrative schematics, not experimental outputs. (c) Updates are injected into frozen YUME; generated chunks update backbone context only.

Here $\Gamma_t = \{\gamma_{\ell,t}\}_{\ell\in\mathcal{S}}$ contains the coefficients for all adapted layers, $\mathcal{D} = \{(A_\ell, B_\ell)\}_{\ell\in\mathcal{S}}$ is the frozen atom dictionary, and $\mathcal{A}$ maps diagonal coefficients into layer-wise LoRA updates. Unlike an unconstrained hypernetwork, g_{ϕ_2} generates bounded coefficients over a fixed dictionary rather than arbitrary model parameters. Section 3.3 defines the layer-wise realization.

A static LoRA uses one update $\Delta W_\ell = \alpha B_\ell A_\ell / r$ for all conditions. A director bank instead stores a finite set $\{\Delta W_\ell^{(j)}\}_{j=1}^{J}$ and requires a routing or merging rule. HorizonLoRA targets the intermediate design space: a shared, low-dimensional coefficient field $c_t \mapsto \Delta W_{\ell,t}$ with a deployed dictionary independent of J. This is a representational hypothesis to be tested empirically, not a guarantee that a condition-specific optimum exists or is recovered.

3 HorizonLoRA

3.1 Overview

HorizonLoRA contains a first-frame encoder, prompt encoder, order-sensitive trajectory encoder, cumulative-state encoder, condition mixer, layer-aware coefficient generator, and normalized rank-one atom dictionaries. Dictionary fitting and conditional-generator training are separated: Stage A learns atoms that reconstruct static teacher updates and their mixtures; Stage B freezes those atoms and trains condition-dependent coefficients. Figure 1 separates the adapter condition path from Base-WM autoregressive feedback.

3.2 Condition Representation and Ordered Encoding

Frozen Base-WM visual and text features are pooled into cached scene and prompt tokens. Each future-control token encodes translation, Lie-algebra rotation, time interval, linear and angular speed, and relative position. A temporal transformer with a summary token preserves order; it does not impose permutation invariance. Cumulative commanded pose is composed in temporal order using body-frame increments. A mixer fuses scene, prompt, cumulative-pose, ordered-control, and rollout-time tokens into h_t. Only the command, trajectory, timing, and coefficient paths are recomputed per chunk. Generated frames never enter this condition path. Appendix D gives the full encoding equations.

3.3 TRAJECTORY-CONDITIONED DIAGONAL-CORE LORA

For each adapted layer ℓ, let $A_\ell \in \mathbb{R}^{r\times d_{\text{in}}}$ and $B_\ell \in \mathbb{R}^{d_{\text{out}}\times r}$ contain normalized rank-one directions. A shared generator combines h_t with a learned layer embedding p_ℓ and predicts r bounded coefficients:

$$\zeta_{\ell,t} = G_\phi(h_t, p_\ell), \qquad \gamma_{\ell,t} = a_{\max}\tanh(H_\gamma(\zeta_{\ell,t})), \tag{9}$$

$$\Delta W_{\ell,t} = \frac{\alpha}{r} B_\ell \operatorname{diag}(\gamma_{\ell,t}) A_\ell. \tag{10}$$

Each row $a_{\ell,k}$ of A_ℓ and corresponding column $b_{\ell,k}$ of B_ℓ is renormalized to unit Euclidean norm after each Stage A update. Each update lies in a shared matrix subspace of dimension at most r. The dictionary therefore determines which matrix directions are available; the coefficient generator determines how they are combined for a condition. The coefficients are nonlinear functions of the scene, text, ordered trajectory, cumulative pose, timing, and rollout state, although synthesis is linear in the fixed atoms. A fixed coefficient vector gives a static update in the same span. Changing coefficients enables conditional adaptation without loading another teacher or increasing deployed dictionary size with teacher count J. The shared-span constraint trades the expressiveness of unrestricted rank-r updates for a reusable deployment representation.

3.4 OFFLINE DICTIONARY FITTING AND RECONSTRUCTION

Stage A trains $J = 8$ trajectory-clustered rank-r teachers. Soft routing over current-chunk ordered-control descriptors forms a teacher mixture $M^T_{\ell,t} = \sum_j \pi_{t,j}\Delta W^{(j)}_\ell$. For normalized atoms $D_{\ell,k} = b_{\ell,k}a_{\ell,k}^\top$, define

$$G_{kk'} = \langle D_{\ell,k}, D_{\ell,k'}\rangle_F, \qquad v_k = \langle D_{\ell,k}, M^T_{\ell,t}\rangle_F, \qquad \beta = (G+\epsilon I)^{-1}v. \tag{11}$$

The anchor $\widetilde{\Pi}^\epsilon_{\mathcal{D}_\ell}(M^T_{\ell,t}) = \sum_k \beta_k D_{\ell,k}$ is a ridge reconstruction, not an orthogonal projection. Proposition 1 in the appendix establishes uniqueness and perturbation stability for $\epsilon > 0$. Stage A minimizes normalized reconstruction errors for individual teachers and mixtures, plus a Gram-orthogonality penalty; atoms are initialized from principal-component directions followed by rank-one SVD, normalized after each update, and then frozen. Stage B distills effective updates through

$$\mathcal{L}_{\text{param}} = \sum_{\ell\in\mathcal{S}} \frac{\|\Delta W_{\ell,t} - \widetilde{\Pi}^\epsilon_{\mathcal{D}_\ell}(M^T_{\ell,t})\|_F^2}{\|M^T_{\ell,t}\|_F^2 + \epsilon_d}. \tag{12}$$

Matching updates avoids the non-identifiability of LoRA factors. Gram-space identities avoid dense adapter materialization; anchor norms expose cancellation among teachers. Appendix C retains the complete fitting objectives, routing rule, and solver details.

3.5 STRUCTURAL PROPERTIES OF THE ADAPTER

The parameterization makes its capacity and amplitude explicit. Write $\mathcal{D}(q) = \sum_k q_k D_k$ and $q = (\alpha/r)\gamma$ for one layer. Since every atom has unit Frobenius norm and $|\gamma_k| \le a_{\max}$,

$$\|\Delta W\|_F \le \frac{|\alpha|}{r}\sum_{k=1}^{r}|\gamma_k| \le |\alpha| a_{\max}. \tag{13}$$

Thus increasing the number of atoms need not increase the worst-case update norm when the global scale is fixed. This is a parameter-space bound; it is not an output-stability guarantee for the recursive backbone.

The ridge anchor and the bounded generator solve different approximation problems. For $f_\epsilon(q) = \|M - \mathcal{D}(q)\|_F^2 + \epsilon\|q\|_2^2$, let β be its unconstrained minimizer and let $\mathcal{B} = [-c, c]^r$, $c = |\alpha|a_{\max}/r$. Completing the square gives

$$f_\epsilon(q) - f_\epsilon(\beta) = (q-\beta)^\top(G+\epsilon I)(q-\beta). \tag{14}$$

If $q_{\mathcal{B}}$ minimizes this objective over the box, the excess objective separates into $f_\epsilon(q_{\mathcal{B}}) - f_\epsilon(\beta)$ and $f_\epsilon(q) - f_\epsilon(q_{\mathcal{B}})$, both nonnegative. The first measures bounded representational approximation; the second measures conditional approximation within the box. Appendix B proves the identities and handles dependent atoms. This distinction motivates source-specific coefficient fitting as a diagnostic without claiming that either term has been measured in the reported comparison.

Algorithm 1: HorizonLoRA Training

Build group-disjoint Sekai and DL3DV splits; merge corresponding partitions.
Train $J = 8$ trajectory-cluster teachers on the train split.
Fit normalized atoms using $\mathcal{L}_A$; renormalize each step.
Freeze Base-WM and the fitted atom dictionaries.
Warm up the conditional generator with ground-truth context.
for each pair of consecutive rollout chunks **do**
    Encode I_0, p, cumulative command, controls, timing, and rollout time.
    Form the soft teacher mixture and ridge-reconstructed anchor.
    Predict bounded diagonal coefficients.
    Optimize $\mathcal{L}_{\text{video}}$, $\mathcal{L}_{\text{param}}$, $\mathcal{L}_{\text{smooth}}$, and $\mathcal{L}_{\text{update}}$.
    Build the second context from $\text{sg}(\hat{x}_t)$ and repeat.
end for
Save split manifests, checkpoints, metrics, and costs as immutable outputs.

3.6 Conditional Training Objectives

Stage B freezes the backbone and dictionary and trains the conditional generator with

$$\mathcal{L}_B = \mathcal{L}_{\text{video}} + \lambda_p \mathcal{L}_{\text{param}} + \lambda_s \mathcal{L}_{\text{smooth}} + \lambda_u \mathcal{L}_{\text{update}}. \tag{15}$$

Here $\mathcal{L}_{\text{video}}$ is YUME's existing flow-matching velocity-target loss, and $\mathcal{L}_{\text{param}}$ matches effective updates to ridge-reconstructed teacher mixtures. The update penalty limits relative Frobenius norm; the smoothness hinge permits adapter changes in proportion to normalized command-state changes. Neither is a formal trust-region or output-Lipschitz guarantee. Teacher forcing precedes two-chunk recursive fine-tuning, with the first generated chunk detached before reuse as context. The reported checkpoint is Stage B step 6810 after two epochs; longer rollouts are evaluation conditions. Appendix E defines the penalties and training context.

3.7 Deployment and Complexity

A rank-r dictionary stores $P_L = \sum_{\ell \in \mathcal{S}} r(d_{\text{in},\ell} + d_{\text{out},\ell})$ parameters. HorizonLoRA adds the condition encoders, mixer, layer embeddings, and coefficient generator, giving a total P_G independent of teacher count J. Ignoring router storage, the symbolic bank-storage break-even is $J^* = \lfloor P_G / P_L \rfloor + 1$. Updates are applied through low-rank factors without materializing dense matrices. Reported artifact sizes and scheduler costs, including teacher training and dictionary fitting, appear in Table 2. Appendix F details parameter accounting; Algorithm 1 summarizes training.

Deployment objects. The required adaptation state consists of the frozen atom factors, the trainable first-frame and prompt projections, the trajectory and cumulative-state encoders, the mixer, the coefficient generator, and layer/time embeddings. Frozen image and text features already belonging to Base-WM are reused, while additional learned projections belong to the adapter accounting. The teachers and their training-time reconstruction targets are not required to predict coefficients at deployment. Appendix F enumerates these parameter groups, and distinguishes the intended module set from the unresolved membership of historical stored artifacts.

Chunk-wise application. A chunk shares one set of coefficients across its denoising evaluations. For an activation matrix X, the adapter is applied as

$$X(W_\ell + \Delta W_{\ell,t})^\top = X W_\ell^\top + \frac{\alpha}{r} \left((X A_\ell^\top) \odot \gamma_{\ell,t} \right) B_\ell^\top. \tag{16}$$

The first-frame and text tokens can be cached over the rollout; only commanded-state, trajectory, time, and coefficient calculations change with the chunk. This describes reuse in the implementation, not a measured latency reduction. Static LoRA, a routed bank, and a generated adapter differ in what they store and recompute, so artifact bytes alone cannot determine their runtime ordering.

4 EXPERIMENTS

4.1 DATA LINEAGE AND SPLITS

Table 1 retains the reported mixed-test comparison. The historical report attributes it to two-epoch checkpoints.These counts and the binding of each table cell to its source-level output remain unconfirmed in the available records. we used 500 clips for evaluation; a configured or submitted evaluation does not establish completed coverage. We retain the historical attribution rather than asserting full-test-set coverage or substituting a larger count.

Each domain uses a 70/15/15 split before corresponding partitions are merged. Sekai groups source videos and near-duplicate components using first/middle/last-frame embeddings and perceptual hashes; DL3DV groups complete scenes. Assignment approximately balances scene labels, path length, speed, and curvature. Assertions check source/component and scene disjointness, domain-namespaced identifiers, unique paths, and complete modality tuples.

Both domains use one camera convention. Translation increments are divided by their source/scene group's robust median non-zero step length; additional normalizers use mixed training only. DL3DV prompts use frozen POI, environment, and lighting metadata, and numerical controls retain frame-wise poses. Missing metadata or invalid poses cause isolation rather than a fallback prompt.

All adaptive methods use a fixed training sampler with $P(\mathrm{Sekai}) = 0.75$ and $P(\mathrm{DL3DV}) = 0.25$. Hyperparameters, checkpoints, trajectory thresholds, and metric calibration are selected by a pre-declared equal-domain macro objective on mixed validation. The frozen configuration is then evaluated once on mixed test. Sekai is called *adapter-held-out*, because YUME pretraining may have included it; DL3DV is called *target-domain held-out*, because its training partition contributes to adapter learning. Ground-truth future views are targets, not model inputs. Each sample provides a first frame, prompt, camera increments, time intervals, cumulative commanded poses, chunk indices, elapsed time, and Base-WM context frames.

4.2 ROLLOUT PROTOCOL AND BASELINES

The Base-WM is frozen YUME (Mao et al., 2025). Main evaluation uses fully recursive rollouts of 1, 4, and 16 chunks; each horizon measures a generated prefix, with no ground-truth future frame reintroduced between chunks. Adapter conditions always use the first frame, prompt, commanded state, future controls, timing, and rollout time rather than generated-frame pose estimates.

The main comparison contains four adaptation methods: Static LoRA, Learned Soft Director Bank, Direct HyperLoRA, and HorizonLoRA. They represent a fixed low-rank update, routing among stored specialists, direct conditional parameter generation, and teacher-distilled coefficient generation, respectively. These are implementations on a common backbone, not comparisons against complete external world-model systems. The learned bank receives exactly the same condition c_t:

$$r_t = \mathrm{softmax}(R_\psi(c_t)), \qquad \Delta W_{\ell,t} = \sum_{j=1}^{J} r_{t,j} \Delta W_\ell^{(j)}. \tag{17}$$

All adaptive methods share mixed training data, the 75/25 domain sampler, per-domain sample-access budget, generation seed, and evaluator. Conditional methods have access to the same declared condition information; Static LoRA uses one update regardless of that condition. Their realized training costs differ and are reported rather than normalized away. The bank uses the same $J = 8$ teachers as HorizonLoRA; teacher clusters and their normalizer are fitted only on mixed training.

Table 2 reports stored adapter artifacts and recorded GPU-hours, including teacher and dictionary preparation where applicable. Learned Bank's recorded time also includes inline evaluation, so this column is not a strictly matched measure of training-only efficiency.

Implementation configuration. The available P0 launchers explicitly set rank and LoRA scale to 128 and target the attention projections `q,k,v,o`. Their wrapper uses 480×832 frames and a 33-frame context window. The recorded code paths differ in recursive scheduling and visual conditioning, so these are recovered implementation settings rather than a newly verified configuration for every table cell.

Quality over 1/4/16 recursive chunks					
Method	S-ATE ↓	S-EBD ↓	D-ATE ↓	D-EBD ↓	DINO† ↑
Static LoRA	0.0010/0.0152/0.0427	–/0.275/0.361	0.350/**1.484**/3.116	–/**0.283**/–	0.775/0.653/0.504
Learned Bank	0.0007/0.0141/0.0302	–/0.232/0.352	0.433/1.585/3.558	–/0.386/–	**0.855**/**0.675**/0.563
Direct HyperLoRA	**0.0006**/0.0141/0.0350	–/**0.215**/**0.306**	0.587/1.631/2.922	–/0.411/–	0.758/0.636/**0.606**
HorizonLoRA	0.0008/**0.0085**/**0.0281**	–/0.313/0.347	**0.088**/2.184/**2.857**	–/0.487/–	0.729/0.575/0.538

Table 1: Reported recursive quality for four adaptation methods on Sekai adapter-held-out (S) and DL3DV-480P-1K target-domain-held-out (D) samples. Cells retain all 1-/4-/16-chunk values; bold marks the best reported value among these four methods at each horizon. Historical scope: N_{Sekai}=15 sources and N_{DL3DV}=1 scene, subject to the record qualification in Section 4.1. †DINO is the unweighted domain macro at 1/4 chunks and Sekai-only at 16 chunks. Dashes denote undefined single-chunk EBD or unavailable multi-chunk features. Reported pose-estimation success is 1.00 throughout.

4.3 Metrics and Statistical Inference

The primary camera metric is similarity-aligned absolute trajectory error (Sim(3)-ATE). We recover generated-view trajectories with MegaSaM (Li et al., 2025b) and apply similarity alignment because metric scale is not reliably available. Sim(3)-ATE measures trajectory agreement after similarity alignment; it is not a standalone test of signed command response or metric-scale accuracy. Pose-estimation success is

$$\text{Success} = \frac{\text{valid pose sequences}}{\text{all evaluated sequences}}. \tag{18}$$

Pose-estimation success measures extraction coverage, not command success. The historical report records success for every evaluated sequence, so its reported conditional and failure-aware ATE coincide.

The main temporal metric is excess boundary discontinuity (EBD). For a chunk boundary at frame b, using normalized DINO feature distance d_ψ,

$$\text{EBD}_b = d_\psi(\hat{x}_{b-1}, \hat{x}_b) - d_\psi(x^{\text{gt}}_{b-1}, x^{\text{gt}}_b). \tag{19}$$

Frames are matched at the same temporal separation, and EBD is averaged over valid boundaries. Single-chunk EBD is undefined because there is no internal chunk boundary. Missing multi-chunk features remain unavailable, not zero. DINO similarity measures perceptual appearance. PSNR, SSIM, LPIPS, and FVD (Unterthiner et al., 2018) are not reported because the corresponding validated measurements are unavailable.

ATE and EBD are reported separately by domain. DINO is the unweighted mean of the Sekai and DL3DV aggregates at one and four chunks; at 16 chunks it covers Sekai only because the DL3DV feature is unavailable. Consequently, these three DINO values do not constitute a common-domain trend. The reported comparison uses one shared generation seed, as attributed in Section 4.1, and has no confidence intervals. We interpret differences as descriptive comparisons, not statistically established gains. Source variability and training-seed variability are distinct; neither uncertainty is reconstructed from the displayed means.

4.4 Main Comparison and Cost Accounting

We compare adapter footprint against the learned bank, then examine geometry, appearance, and boundary continuity across horizons. Both tables retain the historical numerical entries for the four adaptation methods; attribution and verification scope are stated in Section 4.1.

Adapter storage relative to the learned bank. HorizonLoRA represents adaptation using a shared atom dictionary and a condition-to-coefficient generator rather than the full motion-specialist bank. Table 2 reports an adapter-artifact footprint of 811.1 MiB, compared with $2{,}934.0$ MiB for Learned Bank: a 72.4% reduction calculated from those displayed totals. The reported footprint is 27.6% of the bank's, or approximately $1/3.62$ as many stored adapter bytes. This is an adapter-only artifact comparison, not a reduction of the common frozen backbone or a runtime-memory measurement. Teacher construction remains part of the offline cost.

Reported adapter footprint and computational cost		
Method	Deploy MiB ↓	Total GPU-hours ↓
Static LoRA	360.1	66.8
Learned Bank	2,934.0	236.7
Direct HyperLoRA	1,072.7	51.5
HorizonLoRA	811.1	118.9

Table 2: Reported adapter-artifact footprint and recorded computational cost. MiB denotes bytes divided by 2^{20}, excluding the common frozen backbone; exact deployment-file membership remains to be reconciled with the reported totals. GPU-hours are approximate scheduler-derived records, not matched training-only timings: Learned Bank includes inline evaluation, and HorizonLoRA includes teacher training, dictionary fitting, and conditional training.

Long-horizon geometry. At 16 chunks, HorizonLoRA records Sim(3)-ATE values of 0.0281 on Sekai and 2.857 on the evaluated DL3DV samples, compared with 0.0302 and 3.558 for Learned Bank. These are reductions of approximately 7.0% and 19.7% relative to the displayed bank values. Static LoRA and Direct HyperLoRA yield Sekai values of 0.0427 and 0.0350, and DL3DV values of 3.116 and 2.922. The bank comparison directly addresses the teacher-to-deployment representation change; the other adapters place this result within fixed and directly generated adaptation. These comparisons concern aligned trajectory geometry and do not isolate an individual conditioning or training component.

Mid-horizon behavior. The ordering is not uniform across horizons. At one chunk, HorizonLoRA has Sekai ATE 0.0008, while Direct reaches 0.0006; HorizonLoRA has the lowest reported D-ATE (0.088) among the four adapters. At four chunks, HorizonLoRA has Sekai error 0.0085, compared with approximately 0.014–0.015 for the other adapters. On DL3DV, however, HorizonLoRA's four-chunk ATE is 2.184, versus 1.484, 1.585, and 1.631 for Static LoRA, Learned Bank, and Direct HyperLoRA. The stated historical DL3DV scope does not support a stable domain-population ranking. We retain this less favorable horizon alongside the 16-chunk results.

Appearance and boundary continuity. Appearance and boundary metrics favor other adapters in several settings. Learned Bank leads short-horizon appearance (macro DINO 0.855 at $H{=}1$). Direct HyperLoRA leads Sekai-only DINO at $H{=}16$ (0.606 versus HorizonLoRA's 0.538) and Sekai EBD at $H{=}4$ / 16 (0.215 / 0.306, versus 0.313 / 0.347). HorizonLoRA's appearance values are 0.729 / 0.575 / 0.538, with the domain coverage specified above. Static LoRA obtains the lowest four-chunk D-EBD (0.283), alongside higher long-horizon Sekai ATE. Aligned geometry, perceptual similarity, and boundary continuity therefore describe different aspects of the generated sequence rather than one interchangeable quality measure.

Takeaway. The principal comparison is a deployment representation trade-off: HorizonLoRA reuses teacher-informed directions without retaining the full specialist bank. The reported geometry and footprint favor it relative to that bank at the longest evaluated horizon, while Static LoRA and Direct HyperLoRA expose different shorter-horizon, appearance, and cost trade-offs. Table 2 also records 118.9 GPU-hours for HorizonLoRA and 236.7 for Learned Bank, with different accounting scopes; Static LoRA and Direct HyperLoRA record 66.8 and 51.5. These records describe computational expenditure, not a measured training speedup. No latency, throughput, peak-memory, or full-system storage advantage is inferred.

4.5 Additional Diagnostic Scope

Additional diagnostics on a simplified trajectory-residual variant show that lower denoising loss does not consistently imply correct bidirectional motion. They use separate checkpoints and protocols from the main model, and neither establish nor refute the behavior of the complete multimodal formulation.

5 Related Work

Parameter-efficient and conditional adaptation. LoRA learns low-rank factors (Hu et al., 2022); hypernetworks generate parameters from conditions (Ha et al., 2017). VeRA shares frozen random matrices and learns scaling vectors (Kopiczko et al., 2024), while NOLA uses linear combinations of random bases to parameterize low-rank factors (Koohpayegani et al., 2024). HorizonLoRA instead fits a rank-one dictionary to motion-specialized teachers and predicts its coefficients from scene, semantics, ordered trajectory, and rollout state. Fixed bases alone are not the novelty: the proposed contribution is their teacher-supervised, state-conditioned use during recursive generation. Conditional LoRA for video understanding (Jacob et al., 2026) addresses input-dependent adaptation but not this recursive camera-conditioned deployment setting.

World models and camera control. YUME (Mao et al., 2025) provides an autoregressive visual prior and an action-text interface; Sekai and DL3DV provide the adaptation domains (Li et al., 2025a; Ling et al., 2024). CameraCtrl (He et al., 2024) introduces explicit camera conditioning, whereas DimensionX (Sun et al., 2024) motivates motion-specialized LoRA directors. Our comparison concerns alternative adapter representations on one backbone, not reproduction of these complete systems. The learned bank is the direct comparator for amortized storage; Frozen, Static LoRA, and Direct HyperLoRA expose different geometry–appearance and cost profiles. Appendix G reviews the broader literature.

6 Discussion and Limitations

HorizonLoRA separates a teacher-informed matrix span from conditional coefficient synthesis. This representation is useful when deployment must accommodate multiple motion conditions without storing every specialist. Its fixed span can omit useful updates, and teacher mixtures can partially cancel. Conditioning on commanded rather than recovered camera state also means that the adapter does not directly correct realized rollout drift.

The evaluation concerns alternative adapter representations on a frozen YUME backbone. It does not isolate the contribution of every conditioning and training component. Similarity-aligned trajectory error is not a standalone measure of signed command following, and boundary similarity can favor blur or repeated frames. Source-disjoint adapter splits do not establish exclusion from backbone pretraining. The diagnostic appendix preserves the distinction between improved denoising fit and bidirectional response; sample scope, missing metrics, and record attribution are specified with the experimental settings rather than pooled across protocols.

7 Conclusion

HorizonLoRA distills motion-specialized LoRA teachers into a shared dictionary and predicts chunk-specific coefficients for recursive video world models. Its two-stage formulation separates reusable update directions from conditional synthesis and uses Gram-space distillation to compare effective updates without materializing dense adapter matrices. In the reported comparison, this representation reduces deployed adapter storage relative to the learned bank and records lower 16-chunk aligned trajectory error, alongside trade-offs in appearance and boundary continuity. The study examines teacher-distilled coefficient prediction as a compact representation for conditional adaptation without retaining a full motion-specialist bank at deployment.

Reproducibility Statement

The appendices provide condition encodings, dictionary objectives, solver identities, training and deployment equations, and a consolidated account of the separate fitting and command-response diagnostics.

AI Use Statement

Generative AI GPT assisted with manuscript editing, literature checks, and LaTeX formatting and validation.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

David Ha and Jürgen Schmidhuber. World models. In *Advances in Neural Information Processing Systems*, 2018.

David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022a.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022b.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2019.

Anthony Hu, Lloyd Russell, Harrison Yeo, Zak Murez, Georgy Fedoseev, Alex Kendall, and Jamie Shotton. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Darryl Cherian Jacob, Xinyu Liu, Kai Wang, and Pan He. Copra: Conditional parameter adaptation with reinforcement learning for video anomaly detection. https://arxiv.org/abs/2605.15325, 2026.

Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems*, 2016.

Soroush Abbasi Koohpayegani, K. L. Navaneet, Parsa Nooralinejad, Soheil Kolouri, and Hamed Pirsiavash. NOLA: Compressing LoRA using linear combination of random basis. In *International Conference on Learning Representations*, 2024.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025a.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Uriel Singer, Adam Polyak, Thomas Hayes, Xiaoyue Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. https://chenshuo20.github.io/DimensionX/, 2024.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Li, Menghan Chen, Meng Zhou, Yingya Zhang, Yujun Wang, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023b.

Jinbo Xing, Menghan Xia, Yong Liu, Yuechen Zhang, Yuxiang Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023a.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023b.

A RIDGE-OPTIMAL DICTIONARY RECONSTRUCTION

Proposition 1 (unique ridge reconstruction and stability). For one adapted layer, let $D_k = b_k a_k^\top \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ for $k = 1, \dots, r$, define $\mathcal{D}(\beta) = \sum_{k=1}^r \beta_k D_k$, and let

$$G_{kk'} = \langle D_k, D_{k'} \rangle_F, \qquad v_k = \langle D_k, M \rangle_F.$$

For every $\epsilon > 0$, the objective

$$f_\epsilon(\beta) = \|M - \mathcal{D}(\beta)\|_F^2 + \epsilon \|\beta\|_2^2$$

is strictly convex and has the unique minimizer

$$\beta^\star = (G + \epsilon I)^{-1} v.$$

For two target inner-product vectors v and v' computed with the same dictionary,

$$\|\beta^\star(v) - \beta^\star(v')\|_2 \le \frac{1}{\epsilon} \|v - v'\|_2.$$

Proof. Expanding the squared Frobenius norm gives

$$\begin{aligned} f_\epsilon(\beta) &= \|M\|_F^2 - 2 \sum_k \beta_k \langle D_k, M \rangle_F \\ &\quad + \sum_{k,k'} \beta_k \beta_{k'} \langle D_k, D_{k'} \rangle_F \\ &\quad + \epsilon \beta^\top \beta \\ &= \|M\|_F^2 - 2\beta^\top v + \beta^\top (G + \epsilon I) \beta. \end{aligned}$$

The matrix G is positive semidefinite because it is the Gram matrix of the atoms under the Frobenius inner product. Hence $G + \epsilon I$ is positive definite for every $\epsilon > 0$, and f_ϵ is strictly convex. Its unique stationary point satisfies

$$\nabla f_\epsilon(\beta) = 2(G + \epsilon I)\beta - 2v = 0,$$

which yields $\beta^\star = (G + \epsilon I)^{-1} v$.

For two target inner-product vectors,

$$\beta^\star(v) - \beta^\star(v') = (G + \epsilon I)^{-1}(v - v').$$

Since $G \succeq 0$,

$$\|(G + \epsilon I)^{-1}\|_2 = \frac{1}{\lambda_{\min}(G) + \epsilon} \le \frac{1}{\epsilon}.$$

Submultiplicativity of the spectral norm gives the stated stability bound. □

Corollary 1 (Gram-space realization). Let $A \in \mathbb{R}^{r \times d_{\text{in}}}$ contain $a_k^\top$ as its k-th row and let $B \in \mathbb{R}^{d_{\text{out}} \times r}$ contain b_k as its k-th column. Then

$$G = (B^\top B) \odot (A A^\top).$$

Moreover, for any coefficient vectors β, β',

$$\|\mathcal{D}(\beta) - \mathcal{D}(\beta')\|_F^2 = (\beta - \beta')^\top G (\beta - \beta').$$

Derivation. For each pair of atoms,

$$\begin{aligned} G_{kk'} &= \operatorname{tr}\left((b_k a_k^\top)^\top b_{k'} a_{k'}^\top \right) \\ &= (b_k^\top b_{k'})(a_k^\top a_{k'}), \end{aligned}$$

which is the stated Hadamard identity. Applying the Gram definition to $\mathcal{D}(\beta) - \mathcal{D}(\beta') = \sum_k (\beta_k - \beta'_k) D_k$ gives the distance identity. Thus ridge reconstruction and parameter-distillation distances can be evaluated in the r-dimensional coefficient space without materializing dense adapter matrices.

Scope. Proposition 1 concerns the stability of the reconstructed adapter coefficients for a fixed dictionary. It does not imply that the complete Base-WM output is Lipschitz in the adapter condition.

B Bounded Reconstruction and Conditional Approximation

For a fixed layer, let $\mathcal{D}(q) = \sum_k q_k D_k$, $G_{ij} = \langle D_i, D_j \rangle_F$, and $v_i = \langle D_i, M \rangle_F$. The effective generator coefficients are $q = (\alpha/r)\gamma$, so they lie in $\mathcal{B} = [-c, c]^r$ with $c = |\alpha| a_{\max}/r$. A finite tanh output lies in the interior; the closed box provides a best-case reachability bound.

Corollary 2 (bounded approximation gaps). Let $\beta = (G + \epsilon I)^{-1} v$ minimize $f_\epsilon(q) = \|M - \mathcal{D}(q)\|_F^2 + \epsilon \|q\|_2^2$ and let $q_{\mathcal{B}}$ minimize it over $\mathcal{B}$. Then every $q \in \mathcal{B}$ satisfies

$$f_\epsilon(q) - f_\epsilon(\beta) = (q - \beta)^\top (G + \epsilon I)(q - \beta), \tag{20}$$

$$= \big[f_\epsilon(q_{\mathcal{B}}) - f_\epsilon(\beta)\big] + \big[f_\epsilon(q) - f_\epsilon(q_{\mathcal{B}})\big]. \tag{21}$$

Both bracketed terms are nonnegative. Completing the square proves the first identity; adding and subtracting $f_\epsilon(q_{\mathcal{B}})$ proves the second. The minimizer on the box exists and is unique because the box is compact and the objective strictly convex. It is a projection in the $(G + \epsilon I)$ metric, not necessarily coordinatewise clipping.

For unit-Frobenius atoms, $\|\mathcal{D}(q)\|_F \le \sum_k |q_k| \|D_k\|_F \le rc = |\alpha| a_{\max}$. This is a parameter bound, not a claim about long-horizon output stability. If the atoms are independent, exact realization of $\mathcal{D}(\beta)$ requires $q = \beta$; with dependent atoms it requires only $q - \beta \in \ker \mathcal{D}$. Thus an out-of-box anchor coefficient alone does not prove that its matrix update is unreachable. These statements characterize the regularized reconstruction objective, not the denoising objective or the empirical cause of a failed control screen. Neither gap is measured by the reported main comparison.

Computational interpretation. Once the dictionary is frozen, G and its regularized factorization can be reused across teacher targets. For a new target, only v changes. Distillation compares effective updates through Gram-space distances rather than matching non-identifiable low-rank factors. The teacher bank is required for offline supervision but not for coefficient prediction at deployment.

C Dictionary Fitting and Reconstruction Details

We cluster the mixed training split by normalized camera-trajectory descriptors and train $J = 8$ static LoRA teachers, one per cluster. The fixed 75/25 domain sampler also governs teacher sample access, and cluster statistics are computed overall and per domain. These form a DimensionX-style motion-specialist bank on the same frozen backbone, used only for offline supervision.

Hard teacher routing would introduce discontinuities near cluster boundaries. Instead, trajectory distances define weights $\pi_{t,j}$ over the J teachers and form a raw soft anchor

$$\pi_{t,j} \propto \exp(-d(\rho(u_t), \mu_j)/\tau_T)\,, \tag{22}$$

$$M_{\ell,t}^T = \sum_{j=1}^{J} \pi_{t,j} \Delta W_\ell^{(j)}. \tag{23}$$

Here $\rho(u_t)$ summarizes only the ordered controls in the current chunk, so teacher routing and the student use the same future-control horizon. The sum $M_{\ell,t}^T$ can have rank as large as Jr. Let $D_{\ell,k} = b_{\ell,k} a_{\ell,k}^\top$ denote normalized atoms. Ridge-regularized dictionary reconstruction computes

$$\begin{aligned} G_{kk'} &= \langle D_{\ell,k}, D_{\ell,k'} \rangle_F, \\ v_k &= \langle D_{\ell,k}, M_{\ell,t}^T \rangle_F, \\ \beta &= (G + \epsilon I)^{-1} v, \end{aligned} \tag{24}$$

and

$$\widetilde{\Pi}_{\mathcal{D}_\ell}^{\epsilon}(M) = \sum_k \beta_k D_{\ell,k}. \tag{25}$$

For $\epsilon > 0$, $\widetilde{\Pi}^{\epsilon}$ is a ridge reconstruction rather than an orthogonal projection. Cholesky factorization is used when stable, with a documented solve fallback and Gram condition-number logging. Proposition 1 in Appendix A proves that this ridge reconstruction is the unique regularized minimizer, gives its perturbation-stability bound, and derives the Gram-space identities used to avoid dense adapter materialization.

Reconstructing mixtures alone does not ensure coverage of individual teachers. Stage A therefore optimizes

$$\mathcal{L}_{\text{teacher}} = \sum_{\ell,j} \frac{\|\Delta W_\ell^{(j)} - \widetilde{\Pi}^{\epsilon}_{\mathcal{D}_\ell}(\Delta W_\ell^{(j)})\|_F^2}{\|\Delta W_\ell^{(j)}\|_F^2 + \epsilon_d}, \tag{26}$$

$$\mathcal{L}_{\text{mixture}} = \sum_{\ell,t} \frac{\|M_{\ell,t}^T - \widetilde{\Pi}^{\epsilon}_{\mathcal{D}_\ell}(M_{\ell,t}^T)\|_F^2}{\|M_{\ell,t}^T\|_F^2 + \epsilon_d}, \tag{27}$$

$$\mathcal{L}_{\text{orth}} = \sum_{\ell} \|G_\ell - I\|_F^2. \tag{28}$$

The Stage A objective is

$$\mathcal{L}_A = \mathcal{L}_{\text{teacher}} + \lambda_{\text{mix}}\mathcal{L}_{\text{mixture}} + \lambda_o\mathcal{L}_{\text{orth}}. \tag{29}$$

Atoms are initialized from principal-component directions of vectorized teacher updates followed by rank-one singular-value decomposition, optimized with $\mathcal{L}_A$, normalized after every update, and frozen for Stage B. During fitting, we monitor teacher and mixture reconstruction residuals, Gram condition number, effective dictionary rank, and the soft-anchor norm distribution.

With the dictionary fixed, conditional parameter distillation is

$$\mathcal{L}_{\text{param}} = \sum_{\ell \in \mathcal{S}} \frac{\|\Delta W_{\ell,t} - \widetilde{\Pi}^{\epsilon}_{\mathcal{D}_\ell}(M_{\ell,t}^T)\|_F^2}{\|M_{\ell,t}^T\|_F^2 + \epsilon_d}. \tag{30}$$

The loss is applied to effective updates rather than LoRA factors because factorization is not identifiable. Soft-anchor norms are recorded to expose cancellation among teachers, and low-rank Gram identities evaluate distances without materializing dense matrices.

D Condition Encoding Details

A visual encoder E_I converts the first frame into a scene token $e_I = E_I(I_0)$. It reuses frozen Base-WM image or variational-autoencoder features and trains only a pooling projection. A prompt encoder similarly reuses the frozen Base-WM text encoder and pools its output as $e_p = E_p(p)$. Both tokens are cached for the full rollout. If Base-WM requires a multi-frame initial context, the implementation may repeat I_0, but the user input remains one frame.

For camera-conditioned rollout, let $\delta T_{t,i} = (\delta R_{t,i}, \delta p_{t,i})$ be a body-frame increment. We represent its motion as

$$\xi_{t,i} = [\, \delta p_{t,i}, \operatorname{Log}(\delta R_{t,i}) \,] \in \mathbb{R}^6. \tag{31}$$

Linear and angular speeds are

$$v_{t,i} = \frac{\|\delta p_{t,i}\|_2}{\Delta t_{t,i}}, \qquad \omega_{t,i} = \frac{\|\operatorname{Log}(\delta R_{t,i})\|_2}{\Delta t_{t,i}}. \tag{32}$$

Each trajectory token $v_{t,i}$ concatenates normalized Fourier features $\mathrm{FF}(\xi_{t,i})$, $\mathrm{FF}(\Delta t_{t,i})$, $v_{t,i}$, $\omega_{t,i}$, and a relative-position embedding. A lightweight temporal transformer prepends a learned summary token and avoids permutation-invariant mean pooling:

$$e_t^u = E_u([\mathrm{CLS}, v_{t,1}, \ldots, v_{t,H}])_{\mathrm{CLS}}. \tag{33}$$

Ordered tokens and explicit time intervals permit sensitivity to reversal, shuffling, and speed changes; unlike mean pooling, the encoder does not impose permutation invariance. Whether the learned coefficients use these distinctions remains an empirical question.

The cumulative commanded pose is computed only from requested controls. For body-frame increments, the update is $T^{\mathrm{cmd}}_{k+1} = T^{\mathrm{cmd}}_{k}\delta T_k$, so

$$T^{\mathrm{cmd}}_{0\to t} = \prod_{s=0}^{t-1}\prod_{i=1}^{H}\delta T_{s,i}, \tag{34}$$

where multiplication is evaluated from left to right in temporal order. Translation and Lie-algebra rotation features map this pose to e^{cmd}_t. The fused condition is

$$h_t = \mathrm{Mixer}([e_I; e_p; e^{\mathrm{cmd}}_t; e^{u}_t; e_{\mathrm{time}}(t)]). \tag{35}$$

Only e^{cmd}_t, e^{u}_t, time features, and the coefficient generator are recomputed per chunk.

The adapter path deliberately excludes generated frames, preventing visual errors from changing adapter parameters through a second feedback loop. Replacing E_I with an observed-history encoder is a future extension when real temporal observations are available.

E Training Objective Details

YUME uses a flow-matching objective with a velocity target, as implemented by its existing training path. HorizonLoRA does not change this objective; it only changes the effective weights of selected layers:

$$\mathcal{L}_{\mathrm{video}} = \mathbf{E}_\tau \| v_{\theta,\phi}(z_\tau, \tau, x^{\mathrm{ctx}}_t, p, u_t) - v^*_\tau \|_2^2, \tag{36}$$

where v^*_τ is the Base-WM path-sampler velocity target. To limit the magnitude of effective updates, we impose a *relative update-norm penalty*

$$\mathcal{L}_{\mathrm{update}} = \sum_{\ell\in\mathcal{S}} \left[\frac{\|\Delta W_{\ell,t}\|_F}{\|W_\ell\|_F} - \eta \right]_+^2. \tag{37}$$

This penalty is not a formal trust region. We log atom norms, relative Frobenius and spectral update norms, and coefficient distributions, with finite-value assertions.

Let s_t concatenate fixed normalized descriptors of $T^{\mathrm{cmd}}_{0\to t}$, u_t, and elapsed time. Adjacent non-overlapping chunks, including rollout boundaries, are constrained by

$$\mathcal{L}_{\mathrm{smooth}} = \sum_{t,\ell} \left[\frac{\|\Delta W_{\ell,t} - \Delta W_{\ell,t-1}\|_F}{\sqrt{d_{\mathrm{in}} d_{\mathrm{out}}}} - \kappa \|s_t - s_{t-1}\|_2 \right]_+^2. \tag{38}$$

We call this *control-relative adapter smoothness*. It constrains parameter variation only and does not claim that Base-WM outputs satisfy a Lipschitz property. The hinge permits update change in proportion to normalized commanded-state change, including turns and speed shifts. The Stage B objective is

$$\mathcal{L}_B = \mathcal{L}_{\mathrm{video}} + \lambda_p \mathcal{L}_{\mathrm{param}} + \lambda_s \mathcal{L}_{\mathrm{smooth}} + \lambda_u \mathcal{L}_{\mathrm{update}}. \tag{39}$$

Training context. Stage B begins with teacher forcing, using ground-truth context $x^{\mathrm{ctx}}_t = x^{\mathrm{gt}}_t$ to stabilize the coefficient generator. A short recursive fine-tuning phase then unrolls two consecutive chunks from the same source trajectory. The first prediction is detached before it becomes the second context:

$$x^{\mathrm{ctx}}_{t+1} = \mathcal{C}(x^{\mathrm{ctx}}_t, \mathrm{sg}(\hat{x}_t)). \tag{40}$$

The reported Stage B checkpoint is the two-epoch step-6810 artifact. Longer rollouts are evaluation conditions rather than training-time backpropagation horizons.

F Deployment Accounting Details

For adapted layers $\mathcal{S}$, one rank-r LoRA stores

$$P_L = \sum_{\ell \in \mathcal{S}} r(d_{\text{in},\ell} + d_{\text{out},\ell}). \tag{41}$$

Let P_{img}, P_{text}, P_{traj}, P_{cmd}, P_{mix}, P_{PG}, P_{layer}, and P_{time} denote the trainable deployment parameters of the first-frame pooling projection, prompt projection, ordered trajectory encoder, cumulative-state encoder, condition mixer, parameter generator, layer embeddings, and time encoder, respectively. Here

$$P_{\text{PG}} = |\phi_2| = |G_\phi| + |H_\gamma|, \tag{42}$$

and $P_{\text{time}} = 0$ when fixed sinusoidal or Fourier time features are used. HorizonLoRA's total deployed adapter storage is therefore

$$\begin{aligned} P_G = P_L + P_{\text{img}} + P_{\text{text}} + P_{\text{traj}} + P_{\text{cmd}} + P_{\text{mix}} \\ + P_{\text{PG}} + P_{\text{layer}} + P_{\text{time}}. \end{aligned} \tag{43}$$

The symbol P_G includes both the frozen atom dictionary and all condition-to-parameter-generation modules; it should not be confused with P_{PG}, which counts the parameter generator alone. Ignoring router storage, its symbolic storage break-even is the smallest integer

$$J^* = \lfloor P_G / P_L \rfloor + 1 \tag{44}$$

for which a J-director bank stores more adapter parameters. This formula is not a measured advantage; the paper reports adapter MiB and recorded computational cost, including teachers and dictionary fitting.

Equation 16 gives the low-rank application without forming a dense ΔW. Table 2 retains reported adapter-artifact sizes and scheduler-derived computational costs. The symbolic parameter count above specifies the intended deployment modules; it does not independently certify the membership or dtype of each historical artifact total. The low-rank application uses transient coefficients rather than dense adapter matrices.

Algorithm 1 in the main text summarizes this training procedure.

G Extended Related Work

Video world models. Latent world models support prediction and planning through compact learned dynamics (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2023). Recent generative systems extend world modeling to high-dimensional interactive observations and autonomous driving (Bruce et al., 2024; Hu et al., 2023). Sekai provides long egocentric exploration videos and camera annotations (Li et al., 2025a), while YUME uses this data to build an interactive autoregressive world generator (Mao et al., 2025). HorizonLoRA keeps the world-model backbone frozen and focuses on adaptive long-horizon specialization rather than changing its world representation or rollout architecture.

Video generation and control. Diffusion-based video models build on denoising and score modeling, latent diffusion, and diffusion transformers (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Peebles & Xie, 2023). Video extensions include direct diffusion, cascaded generation, variable-length token prediction, transformer pretraining, and latent video diffusion (Ho et al., 2022a; Singer et al., 2022; Ho et al., 2022b; Villegas et al., 2022; Hong et al., 2023; Blattmann et al., 2023b;a). Controllable systems introduce motion modules, compositional conditions, explicit camera controllers, and image animation priors (Guo et al., 2023; Wang et al., 2023a;b; He et al., 2024; Xing et al., 2023). DimensionX learns separate spatial and temporal LoRA directors and multiple motion-specific S-Directors (Sun et al., 2024). HorizonLoRA studies the narrower question of amortizing a DimensionX-style trajectory-director bank into conditional coefficients for a frozen world-model backbone.

Conditional parameter adaptation. Adapters, prompts, low-dimensional updates, LoRA, adaptive rank allocation, and quantized LoRA reduce the cost of model specialization (Houlsby et al., 2019; Li & Liang, 2021; Lester et al., 2021; Aghajanyan et al., 2021; Hu et al., 2022; Zhang et al., 2023b; Dettmers et al., 2023). Lightweight adaptation is also widely used for visual personalization and conditional generation (Gal et al., 2023; Ruiz et al., 2023; Zhang et al., 2023a). Hypernetworks and dynamic filters generate parameters from task or input representations (Ha et al., 2017; Jia et al., 2016), while mixture-style adapters route among reusable experts. Recent conditional LoRA work demonstrates input-dependent adaptation in video understanding (Jacob et al., 2026). HorizonLoRA focuses on scene–semantic–trajectory conditioning for recursive video rollout. Relative to a learned soft director router, it discards the full teacher bank after training and predicts coefficients over a fitted fixed dictionary. The reported comparison gives lower 16-chunk aligned trajectory error for this representation relative to the bank; individual training and conditioning choices are not isolated by that comparison.