DRIFT FORCING: SAMPLE-BASED DRIFT FIELDS AS INITIALIZATION FOR AUTOREGRESSIVE VIDEO DIFFUSION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video diffusion distillation makes a bidirectional teacher causal, initializes a few-step causal student and refines it with distribution matching distillation (DMD), in which the bidirectional teacher scores the student's own rollouts. The released initializations imitate the causal teacher one trajectory at a time under teacher forcing: the student is trained on ground-truth prefixes, then handed to a stage that conditions it on its own samples. We ask whether the initialization can instead be trained on those samples directly, with no teacher trajectories, no score targets and no trained critic. Drift Forcing does this with a drift field in the feature space of a frozen causal teacher: the student's own rollouts are attracted toward positive clips, one bidirectional-teacher clip per training prompt, and repelled from one another, in the self-rollout regime that DMD and inference share. The main recipe runs a coarse field over clip-level features and then a finer field over per-chunk features, for about 400 H800-hours on the public Causal-rCM synthetic training set. On Wan2.1-1.3B, with the DMD trainer, data and budget held fixed, the student that DMD makes from the drift initialization reaches the highest VBench Total among all products compared (84.63), above DMD from the released consistency-distilled initializations and from the causal diffusion model with no initialization stage. The lead comes from VBench's dynamic-degree dimension, the share of clips that move. Code is available at `https://anonymous.4open.science/r/DF0E51`.

1 INTRODUCTION

Real-time interactive video generation needs a student that is *autoregressive*, emitting frames causally, and *few-step*, denoising each chunk in a handful of network evaluations. Distilling such students from bidirectional diffusion teachers (Wan et al., 2025) has converged on a three-stage recipe: make the teacher causal (stage 1), *initialize* a few-step causal student (stage 2), then *refine* it with asymmetric distribution matching distillation (DMD, stage 3; Yin et al., 2024a;b), in which the student rolls out causally while a bidirectional teacher scores the result (Yin et al., 2025; Huang et al., 2025; Zhu et al., 2026; Zhao et al., 2026b). Causal-rCM (Zheng et al., 2026b) provides the recipe and checkpoints we build on.

The released stage-2 initializations all imitate the causal teacher one trajectory at a time under teacher forcing: causal ODE regression on a curated set of teacher trajectories, or causal consistency distillation (CD; Zhu et al., 2026; Zhao et al., 2026b; Zheng et al., 2026b). Causal DMD used *as* the initialization is also trained under teacher forcing and remains weaker than causal CD under rollout (Zhao et al., 2026b). The student is trained on ground-truth prefixes although at inference it conditions on its own outputs. The criterion that matters for the slot, moreover, is not the initialization's standalone quality but its *utility for the DMD stage that follows* (Li et al., 2026b; Zhao et al., 2026b). We ask one question: under a fixed downstream DMD recipe, can a few-step initialization be trained on the student's own autoregressive samples, without per-trajectory imitation?

Drift Forcing is our answer. A drift field (Deng et al., 2026) in the feature space of the frozen *causal* teacher attracts the student's own four-step rollouts toward a small library of positive clips and repels

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.

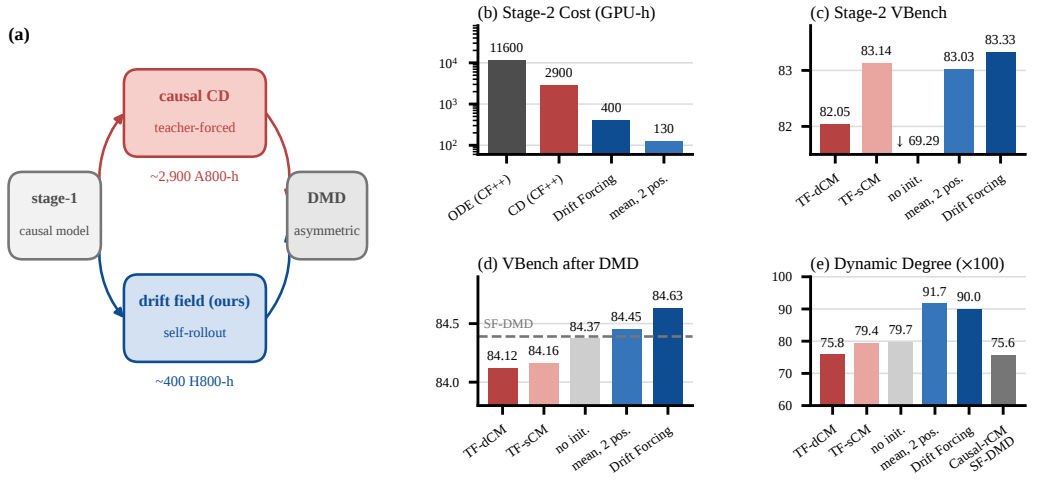

Figure 1: Drift Forcing in one view. (a) The consistency-distillation and drift pipelines share the stage-1 causal model and the asymmetric-DMD stage; only the initialization differs. (b) Stage-2 cost (log scale; ODE and CD costs as reported by Causal Forcing++ in A800-h, drift in H800-h); (c) VBench Total of the initializations (§4.1 protocol; the stage-1 model's four-step score lies below the axis); (d) VBench Total of their DMD products under one trainer, best checkpoint within 2000 DMD iterations, with the Causal-rCM self-forcing DMD (SF-DMD) student dashed; (e) dynamic degree of the same products. Bar labels abbreviate the row names of Table 1: TF-dCM and TF-sCM are the released discrete- and continuous-time consistency-distilled initializations of Causal-rCM (Zheng et al., 2026b), and CF++ is Causal Forcing++ (Zhao et al., 2026b).

them from one another. Because the field is computed from features of sampled clips rather than from teacher trajectories or scores, the teacher enters only as a feature extractor: its activations define the objective, and the gradient flows back through it to the rollout. Stage 2 thus runs in the *same* self-rollout regime as stage 3 and inference: the student conditions on chunks it generated itself, the regime known as self-forcing (Huang et al., 2025). The main recipe runs a coarse mean-pooled (clip-level) field first and a finer chunk-pooled (chunk-level) field second (§3.3), for about 400 H800-hours on clips from the Causal-rCM synthetic training set.

We make three contributions.

- **A sample-based, critic-free, self-rollout initialization** (§3): a drift field inside an autoregressive text-to-video distillation stack at diffusion-transformer (DiT) scale, with the two-stage schedule as the main recipe (Figure 2).
- **A DMD product that leads on VBench** (§4.3): with the DMD trainer, data, batch size and iteration budget held fixed and only the initialization changed, the student that DMD makes from Drift Forcing (two-stage), its *DMD product*, has the highest VBench Total among all products compared. The lead is carried by dynamic degree and paid for in temporal consistency.
- **The field's boundaries** (§4.4, §4.5): the field trains the student under self-rollout and not under teacher forcing, and a motion field raises dynamic degree before and after DMD, though not the VBench Total after it.

2 Background and Related Work

Autoregressive video diffusion distillation. The Causal Forcing line (Zhu et al., 2026; Zhao et al., 2026b; Zheng et al., 2026b) defines the three stages introduced above. Causal-rCM (Zheng et al., 2026b) releases the teacher-forced (TF) and self-forcing (SF) recipes with checkpoints. We use its two teacher-forced consistency-distilled initializations, TF-dCM and TF-sCM (discrete- and continuous-time consistency models), and its self-forcing DMD student trained from TF-dCM, the

Causal-rCM SF-DMD student. Released pipelines within and beyond this line fill the initialization slot with a per-sample objective: ODE or regression warm-up (Yin et al., 2025; Huang et al., 2025; Gamazo Tejero et al., 2026; Yang et al., 2026c), consistency distillation (Zheng et al., 2026b; Yang et al., 2026b), perceptual-space regression (PFM; Zhao et al., 2026a; Qian et al., 2026). None measures what the initialization contributes to coverage of the teacher distribution after DMD. Two recent methods act after the initialization slot instead: OPSD-V (Liu et al., 2026a) post-trains a few-step autoregressive generator on its own rollouts against a teacher given real long-video context, and Salt (Ge et al., 2026) regularizes the composition of denoising updates and conditions on KV-cache quality inside the DMD stage. CMD (Bandyopadhyay et al., 2026) removes the stage by making the scoring teacher causal. Appendix P discusses it and DistillAlign's joint objective (Li et al., 2026b) as alternatives to a separate initialization stage. The Causal Forcing line (Zhu et al., 2026; Zhao et al., 2026b) argues that trajectory-matching stages need a causal teacher while distribution-matching stages should use the stronger bidirectional one. Our design follows that split for a sample-based objective: the bidirectional teacher supplies the positives that define the target, and the causal teacher serves only as the feature extractor, where causality is required (§3.2). Long-video extensions (Liu et al., 2026c; Yang et al., 2026a; Cui et al., 2026; Zhuang et al., 2026; Chen et al., 2026a) and pyramidal inference schedules (Korzhenkov et al., 2026) are orthogonal.

Mode seeking, mode covering, and motion collapse. Reverse-KL score distillation is mode-seeking, buying per-sample sharpness by narrowing the matched distribution. DistillAlign (Li et al., 2026b) selects initializations by coverage and anchors DMD with a CD term, as DOLLAR (Ding et al., 2025) does for a bidirectional student. f-Distill (Xu et al., 2025), ADM (Lu et al., 2025), SGMD (Wu et al., 2026) and HiAR (Zou et al., 2026a) modify the divergence or the schedule, MSmMS (Cai et al., 2026) adds a supervised flow-matching head, Transition Matching Distillation (Nie et al., 2026) applies DMD to a chain of conditional flows in place of one few-step map, and Data-Forcing Distillation (Chen et al., 2026b) evaluates the teacher score at real samples; DM-Align (Lin et al., 2026) adds a gradient toward preferred samples. Like Drift Forcing, both anchor training to samples, but inside DMD. DUET (Li et al., 2026c) reports that a consistency-initialized student loses its inherited diversity under continued DMD even as per-sample quality improves, and TurboT2VA (Yang et al., 2026b) delays DMD for the same reason, which motivates testing what an initialization carries *through* DMD. On the motion side, DynaForcing (Huang et al., 2026) formalizes *dynamic collapse* in self-forcing DMD, where static content becomes a mode that the DMD gradient cannot push the student out of. Published remedies reweight the DMD gradient, augment the conditioning, anchor to data (Huang et al., 2026; Lu et al., 2026b), perturb the rollout with dual-noise masks (Zhao et al., 2026c) or modify the objective (Liu et al., 2026b; You et al., 2026; Zhang et al., 2026c; Li et al., 2026a); all act during or after DMD, none at the initialization stage.

Drifting generative models and semantic feature spaces. Drifting models (Deng et al., 2026) train one-step generators by moving samples along a field of attraction to data samples and repulsion among generated samples, with a row- and column-normalized double-softmax weighting, the form kept in Eq. 1. A line of theoretical work characterizes fixed points and biases (Zhang et al., 2026a; Gretton et al., 2026; Li & Zhu, 2026) and extends the dynamics (Brown et al., 2026; Kazanskii et al., 2026). Drifting has been used to distill image generators (Zhang et al., 2026b;d), to train robot world models from scratch with U-Nets below 200M parameters (DriftWorld (Lu et al., 2026a)), and as a decoder or action generator (Zou et al., 2026b; Gao et al., 2026a;b). No prior work places a drift field inside a text-to-video distillation stack at DiT scale as the initialization for DMD. Semantic feature spaces are becoming standard in video generation, as generative latents (Zheng et al., 2026a; Xie et al., 2026), alignment targets (Zhang et al., 2025; Lian et al., 2026) and evaluation embeddings (Luo et al., 2025; Ge et al., 2024). Unlike latents and alignment targets, which act on one sample at a time, our field acts on a population of samples, and it must be computable causally and in latent space (§3.2).

3 Drift Forcing

3.1 The three-stage recipe

We inherit the three-stage recipe of the Causal Forcing line (Zhu et al., 2026; Zhao et al., 2026b; Zheng et al., 2026b). Stage 1 adapts bidirectional Wan2.1 models into *causal* diffusion models by

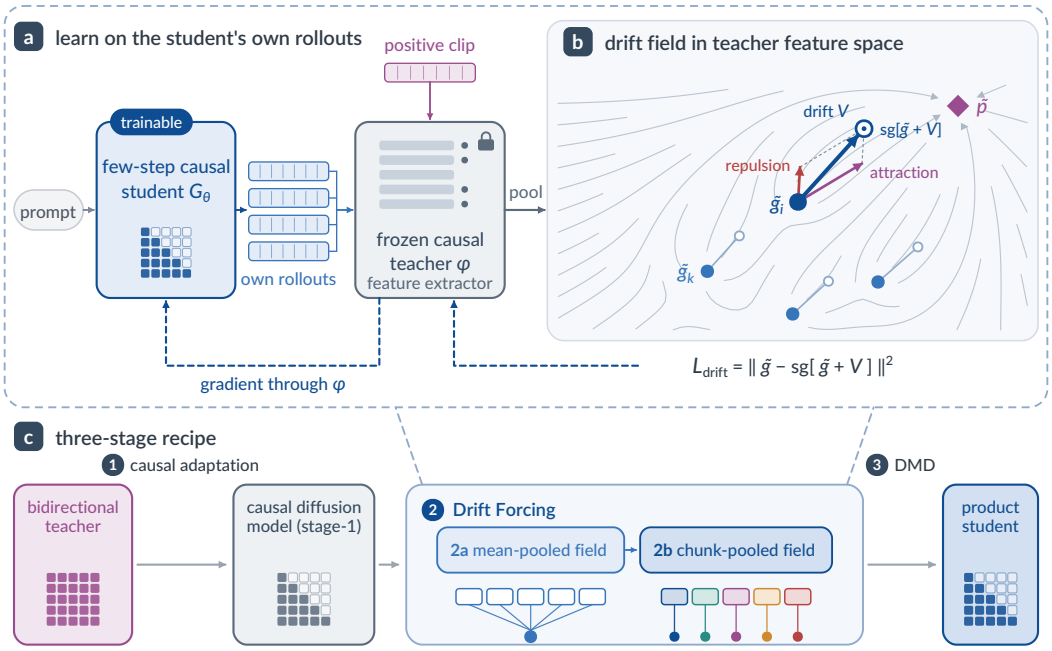

Figure 2: One Drift Forcing training step (a, b) and its place in the three-stage recipe (c). (a) The student rolls out n clips per prompt with its own chunk-wise few-step sampler; the rollouts and the prompt's positive clip from the Causal-rCM synthetic training set pass through the frozen causal teacher, which serves only as a feature extractor. (b) In the teacher's feature space the field on each rollout is attraction toward the positive's features minus repulsion from the sibling rollouts (Eq. 1); the field defines a stop-gradient target, the loss regresses the rollout's features onto it (Eq. 2), and the gradient flows back through the frozen teacher into the student. Field lines illustrate the attraction and repulsion. No teacher trajectory, score target or critic enters. (c) The three-stage recipe. Stage 1 adapts bidirectional Wan2.1 models into causal diffusion models by teacher forcing; the 14B causal model is the frozen teacher ϕ of (a), and the 1.3B one, drawn here, is the student's starting point (the mask icons show which chunks each chunk attends to). Drift Forcing trains the initialization in two stages: the mean-pooled field pools every chunk into one vector per clip (a coarse target) and the chunk-pooled field keeps one vector per chunk (a finer one). The unchanged DMD stage then turns the initialization into its DMD product, which Table 1 compares.

teacher-forced training on chunks of three latent frames: the 14B causal model is the frozen feature extractor ϕ, and the 1.3B one, which we call the stage-1 model, is the student's starting point. Stage 2, the stage Drift Forcing fills, produces a few-step causal student G_θ that denoises each chunk in four steps conditioned on previously generated chunks. Stage 3 refines G_θ with asymmetric DMD, kept as implemented by Causal Forcing++.

3.2 A DRIFT FIELD ON THE STUDENT'S OWN ROLLOUTS

Figure 2 summarizes one stage-2 step. The student produces $n{=}4$ rollouts per prompt with its own four-step chunk-wise sampler, exactly as at inference. The field takes its features from the residual stream of ϕ after blocks $\ell \in \{1, 9, 19, 29\}$ (the *taps*) and uses no other network. Rollouts and positive clips are encoded by the causal teacher, and the activations at tap ℓ are averaged over the tokens of each chunk into one vector per chunk position c (*chunk pooling*); *mean pooling* averages over the whole clip instead, giving one vector per clip. Write g_i for the pooled features of rollout i and p_j for those of positive j at a fixed (ℓ, c); tildes denote the same features rescaled within each group that shares a tap, a chunk position and a prompt, so that coordinates and the distances d below are of order one (exact form in Appendix E). For a bandwidth τ and the matrix d of distances from each rollout (rows) to the positives and to its *siblings*, the other $n{-}1$ rollouts of the same prompt (columns), the pairing weights are the row-column symmetrized softmax

$a = \sqrt{\mathrm{softmax}_{\mathrm{row}}(-d/\tau) \odot \mathrm{softmax}_{\mathrm{col}}(-d/\tau)}$ (elementwise). The row softmax runs over the positives and the siblings together and the column softmax over the rollouts, with self-pairs masked, so a rollout's attraction mass is shared among its positives and its repulsion mass among its siblings, the split set by the distances. Its entries are a_{ij}^{pos} for positive j and a_{ik}^{sib} for sibling k. The field on rollout i is attraction to positives minus repulsion from siblings, weighted by w_{rep} (1 unless stated):

$$V_\tau(g_i) \;=\; \sum_j a_{ij}^{\mathrm{pos}} \left(\textstyle\sum_k a_{ik}^{\mathrm{sib}} \right) \tilde{p}_j \;-\; w_{\mathrm{rep}} \textstyle\sum_k a_{ik}^{\mathrm{sib}} \left(\sum_j a_{ij}^{\mathrm{pos}} \right) \tilde{g}_k. \tag{1}$$

The field is computed at four bandwidths $\tau \in \mathcal{T}$ (values in Appendix E), each divided by its root-mean-square r_τ over rollouts and feature dimensions, with r_τ floored at r_{soft} (no floor, $r_{\mathrm{soft}}{=}0$, unless stated), and the results summed, $V = \sum_{\tau \in \mathcal{T}} V_\tau / \max(r_\tau, r_{\mathrm{soft}})$. The loss is the drifting fixed-point regression, which moves each rollout's features onto their drifted, stop-gradient target,

$$\mathcal{L}_{\mathrm{drift}} = \left\| \tilde{g} - \mathrm{sg}\left[\tilde{g} + V\right] \right\|_2^2, \tag{2}$$

averaged over taps and chunk positions (sg: stop-gradient; the field is evaluated without gradients, and the target $\mathrm{sg}[\tilde{g} + V]$ is a constant, so the gradient reaches θ only through the $\tilde{g}$ outside the bracket). Eq. 1 is the double-softmax field of Deng et al. (2026), used unchanged (Appendix K). Because the rollouts are generated exactly as at inference and gradients reach θ through them, stage 2 trains under self-rollout with no teacher-forced prefix; Appendix E details the feature pass, the gradient path and the floor.

Why the causal teacher as feature space. *Causality*: the student generates each clip chunk by chunk, so each chunk's feature must depend on that chunk and its past only, and ϕ must be causal; strong external video encoders are bidirectional, and the causal streaming backbones that exist (Yan et al., 2026; 2025) are trained for video understanding. *Latent-space cost*: the gradient of Eq. 2 flows through ϕ into the student, and a pixel-space encoder would route it through the VAE decoder at full resolution (as DriftWorld (Lu et al., 2026a) does with DINO features at low resolution). *Separation from the evaluation encoder*: V-JEPA2 (Assran et al., 2025), in whose feature space we measure precision, coverage and Vendi diversity (Friedman & Dieng, 2023) (§4.5), is never queried by the field. Two controls test this choice: within the teacher family, the 1.3B stage-1 model as ϕ tracks the 14B teacher (§4.4); across encoder families, four external encoders agree with the teacher on how often a student token's nearest neighbor among positives and siblings is a positive (Appendix H).

3.3 Two-stage schedule and positive library

The main initialization, Drift Forcing (two-stage) in every table, is trained from the stage-1 model in two stages, 2a and 2b, of 1,000 iterations each, one field per stage. Stage 2a uses mean pooling with one positive per prompt over the full 4,096-prompt pool at learning rate $4{\times}10^{-6}$ (Appendix F). Stage 2b continues from it with chunk pooling and the floor $r_{\mathrm{soft}}{=}0.10$ at $2{\times}10^{-6}$. The positive library is the field's only data dependency, and for this recipe it costs nothing to build: the positive of each prompt is that prompt's clip in the Causal-rCM synthetic training set, which the consistency-distilled initializations also train on. The initializations trained on a 512-prompt pool (Tables 1 and 18, §4.4) use up to four teacher clips per prompt as positives; no real video enters stage 2. Table 11 (Appendix D) places Drift Forcing in the stage-2 design space. The motion field of Appendix I, a second field on the temporal increments of the chunk features, is not in the main recipe. It appears in §4.5 and in the §4.4 controls that use the chunk-pooled + motion, 4 positives configuration.

4 Experiments

4.1 Setup

All experiments train Wan2.1-1.3B causal students at 480p in the Causal-rCM stack, using its released checkpoints (Zheng et al., 2026b): its 14B causal model is the feature extractor ϕ of §3, and the bidirectional Wan2.1-14B teacher provides the real score in the DMD stage. Every external number is re-measured under our frozen VBench (Huang et al., 2024) protocol (Appendix A). "VBench" means this protocol's Total unless a dimension is named, and "consistency" is the mean of its four normalized temporal-consistency dimensions. Table 1 gives dynamic degree its own column: it is where the products differ most, and at half the weight of the other quality dimensions

a point of it is worth 0.06 points of Total. Two reference instruments complement VBench (§5.2): V-JEPA2 precision and coverage (Kynkäänniemi et al., 2019; Naeem et al., 2020) and Vendi for the distribution axis (§4.5), and Fréchet video distance (FVD) of 128 generated clips against the positive library, which monitors training trajectories (Appendix E). A *product* is what an initialization becomes after DMD; *teacher-fit* is closeness to the teacher's samples on FVD, precision and coverage; "it." abbreviates iterations.

Stage 3 uses the Causal Forcing++ DMD trainer with VidProM prompts (Wang & Yang, 2024). Stage-3 iterations are DMD trainer iterations; each iteration updates DMD's fake-score network once, and every fifth iteration also updates the student. Every initialization is reported at its best evaluated checkpoint within 2000 DMD iterations and at iteration 1000, the stage-3 length of Causal Forcing++, which we call the matched budget. The matched budget is the primary comparison; the "best" column is descriptive, and the two-stage product is first in both. Appendix E lists the evaluated checkpoints. A small-data variant of the DMD stage is reported in Appendix C.

4.2 The initialization on its own

Before DMD, on the frozen protocol, the Drift Forcing (two-stage) initialization and the mean-pooled, 2 positives initialization score within 1.3 VBench points of the consistency-distilled initializations, the two-stage initialization above both, and about 14 points above the stage-1 model rendered in four steps, the start of the no-initialization run (Table 10, Appendix D). **Cost.** Stage 2 of the main recipe is about 400 H800-hours (Table 3), including the forward and backward passes of the 14B feature extractor. Causal Forcing++ reports about 2,900 A800-hours for its consistency-distillation stage (Appendix D). Stage 3 costs about 200 H800-hours for every initialization, so the two-stage product at iteration 1000 costs about 600 H800-hours beyond the stage-1 model.

4.3 Main result: what DMD makes of the initialization

The runs of Table 1 hold the DMD trainer, prompt set and batch fixed and change only the initialization. At the matched budget Drift Forcing (two-stage) gives the highest VBench Total among all products compared, above the products of the consistency-distilled initializations and of the stage-1 model under the same trainer, and above the Causal-rCM SF-DMD student, a fixed reference checkpoint. We call its DMD product the two-stage product.

Against the TF-dCM product, dynamic degree accounts for more than the whole lead: the two-stage product's dynamic degree is about 14 points higher, which is worth 0.9 points of Total against a lead of 0.5 points, while its consistency is about one point lower (Figure 3). The no-initialization product has no stage-2 cost and comes within a third of a point of the two-stage product; here too the dynamic-degree gap is worth more than the lead (0.6 against 0.3 points). The no-initialization product's shortfall is the one Causal Forcing++ reports for an initialization without few-step capability: DMD refines what its initialization renders (Zhao et al., 2026b). In their frame-wise regime the shortfall shows as blur and near stillness; in ours, on VBench, it shows as missing motion: the no-initialization product renders fewer of its clips as moving and is smoother in time (Appendix Q). What the two-stage initialization buys is therefore motion that DMD alone does not produce, about ten points of dynamic degree, for under one point of consistency and the stage-2 compute.

Within the first 2000 DMD iterations, the two-stage and mean-pooled field only products score highest at iteration 1000, while the TF-dCM product's Total moves by less than 0.1 (Figure 4, Appendix B). Of the seven runs extended to iteration 3000, three score higher at 2500 than at any earlier checkpoint, and none is higher at 3000 than at its best within 2000 (Table 7).

4.4 Why this recipe: controls and ablations

The field trains under self-rollout and fails under teacher forcing. A *placement* is the set of samples the field acts on; the field fails under teacher forcing: a fully teacher-forced variant is inert from the stage-1 model and destructive from TF-dCM, where FVD more than triples within a few hundred iterations, and fixing the first one to three chunks to ground truth freezes those chunks instead of training them. The failure is structural (Appendix G): under teacher forcing the rollout side of the field is anchored to a ground-truth prefix rather than sampled from the student, so the within-prompt spread the field needs collapses. Under self-rollout the same field holds the Causal-

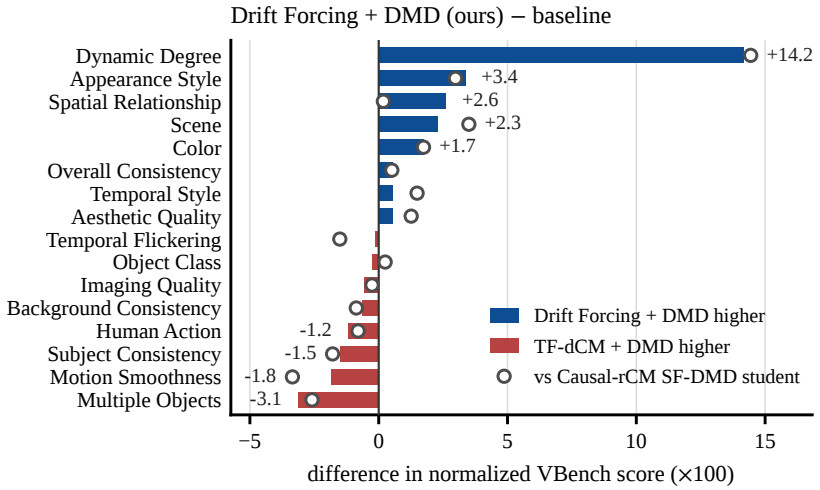

Figure 3: Per-dimension VBench difference at iteration 1000 (normalized scores $\times 100$): the two-stage product (Drift Forcing + DMD) minus the TF-dCM product under the same trainer (bars; blue where the two-stage product is higher, red where it is lower) and minus the Causal-rCM SF-DMD student (circles).

	VBench (frozen protocol)					reference instruments		
Initialization	Total, it. 1000 ↑	best (it.) ↑	dyn. ↑	cons. ↑	Sem. ↑	FVD ↓	P ↑	C ↑
drift, one positive per prompt over 4,096 prompts								
Drift Forcing (two-stage)	**84.63**	**84.63** (1000)	90.0	93.15	81.65	965.3	.504	.137
chunk-pooled field only	83.84	84.13 (2000)	90.3	92.72	80.71	933.9	.461	.180
mean-pooled field only	83.94	83.94 (1000)	**94.2**	91.86	80.92	978.2	.441	.170
drift, 2–4 positives per prompt over 512 prompts								
mean-pooled, 2 positives	84.41	84.45 (1875)	91.7	92.85	81.21	847.0	.389	.170
chunk-pooled, 4 positives	84.08	84.17 (2000)	91.1	92.50	81.17	936.1	.426	.178
chunk-pooled + motion, 4 positives	83.57	83.99 (2000)	92.2	92.19	80.58	901.0	.473	.162
no drift stage								
no initialization (stage-1 model)	84.37	84.37 (1000)	79.7	93.83	**81.79**	922.5	.553	.170
TF-dCM	84.12	84.12 (1000)	75.8	**94.17**	80.92	832.5	**.596**	**.223**
TF-sCM	84.16	84.16 (1000)	79.4	93.50	81.16	**806.8**	.527	.213
Causal-rCM SF-DMD student	84.39	n/a	75.6	95.03	80.85	800.6	.506	.260

Table 1: Initialization utility under one DMD recipe: VBench Total at DMD iteration 1000 and the best Total within 2000 DMD iterations (iteration in parentheses); dyn., cons., Sem., FVD, P, C: dynamic degree, consistency, Semantic score, FVD and V-JEPA2 precision and coverage at the "best" checkpoint. Blue: the main recipe (§3.3). Bold and underline: best and second best above the rule; the Causal-rCM SF-DMD student is a fixed checkpoint.

rCM SF-DMD student in place, and a run from the stage-1 model descends monotonically on FVD (Appendix J). **The two-stage schedule beats either field alone.** Run for the same 2,000 stage-2 iterations, the chunk-pooled field only initialization and the mean-pooled field only initialization both give products with a lower Total than the two-stage product (Table 1). The mean-pooled field only initialization scores higher on VBench before DMD than the two-stage initialization (Tables 10 and 12), but its product has more motion and less consistency at iteration 1000 and a Total more than a point lower from iteration 2000 on (Table 7), whereas runs that end with a chunk-pooled field keep theirs. **Within the teacher family, extractor capacity and the floor cost nothing; repulsion does.** With the 1.3B causal teacher as ϕ, the run matches the 14B run's FVD by stage-2 iteration 1200 and its VBench Total within a fifth of a point (Appendix G). The floor r_{soft} (§3) removes the $1/r_\tau$ singularity, lowers FVD and leaves precision, coverage and VBench Total unchanged at stage-2 iteration 1000; it is part of stage 2b of the main recipe. Repulsion is not dispensable: the two-stage, attraction only initialization ends stage 2 equal on VBench but behind on teacher-fit, and its DMD product has lower precision, coverage, consistency and Total (Table 2, Appendix G.3).

Initialization	precision ↑		coverage ↑		Vendi ↑	
	before	after DMD	before	after DMD	before	after DMD
our DMD recipe, products at DMD iteration 1000						
Drift Forcing (two-stage)	.678	.504	.473	.137	1.321	1.267
two-stage, attraction only	.629	.410	.436	.109	1.300	1.249
chunk-pooled field only	.680	.506	.514	.106	1.332	1.248
mean-pooled field only	.652	.441	.492	.170	1.317	1.280
mean-pooled, 2 positives	.518	.412	.318	.109	1.324	1.251
mean-pooled, 4 positives	.565	.473	.375	.102	1.324	1.248
chunk-pooled, 4 positives	.617	.383	.412	.123	1.334	1.269
chunk-pooled + motion, 4 positives	.617	.453	.441	.072	1.333	1.238
no initialization (stage-1 model)	.104	.553	.049	.170	1.290	1.270
TF-dCM	.789	.596	.586	.223	1.324	1.290
TF-sCM	.742	.527	.537	.213	1.332	1.287
released pipelines, at their published training length						
TF-dCM → Causal-rCM SF-DMD student	.789	.506	.586	.260	1.324	1.299
TF-sCM → Causal-rCM sCM-initialized SF-DMD student	.742	.660	.537	.393	1.332	1.308
DistillAlign initialization → its distilled student	.549	.443	.398	.252	1.318	1.325

Table 2: V-JEPA2 precision, coverage and Vendi diversity before and after DMD, against a 512-clip bidirectional-teacher reference (teacher Vendi 1.380; products at iteration 2000 in Table 8). Noise floors: precision 0.0097, coverage 0.0078, Vendi 0.0001.

The library benefits from clean positives and a wide pool. A positive downsampled from a 720p teacher library reaches a shallower FVD minimum than the library clip alone; freshly generated positives change when the minimum is reached, not its value; and the wider pool (4,096 prompts, the main recipe's) scores best on coverage and stage-2 VBench (Appendix G.1). **For stage 2, the stage-1 model is as good a starting point as TF-dCM.** With field and data fixed, the chunk-pooled + motion, 4 positives configuration reaches nearly the same VBench from the stage-1 model and from TF-dCM (Appendix G), and the run from TF-dCM gives up part of TF-dCM's teacher-fit.

4.5 Beyond the Total: coverage, motion and stability

Coverage after DMD follows the initialization's. VBench scores samples, not coverage, so we measure precision, coverage and Vendi diversity in V-JEPA2 feature space (Assran et al., 2025; Friedman & Dieng, 2023) against a fixed reference of bidirectional-teacher clips (protocol in Appendix E, noise floors in the caption of Table 2); Table 2 pairs every initialization with its DMD product. From every few-step initialization the DMD stage lowers precision and coverage, and the Causal-rCM and DistillAlign pipelines show the same movement. The TF-dCM and TF-sCM initializations enter with the most coverage and their products keep the most, and every product has more coverage at iteration 2000 than at 1000 (Table 8, Appendix B); Appendix L relates the coverage gap between drift and consistency-distilled initializations, and its loss through DMD, to the two objectives.

A motion field carries motion through DMD but not VBench Total. The field of Eq. 1 matches where chunk features sit, not how they move; the motion field (§3.3, Appendix I) addresses this. Three initializations (mean-pooled, 4 positives; chunk-pooled, 4 positives; chunk-pooled + motion, 4 positives) share stage-2 iterations, data and library and differ one step at a time; the probe row of Table 18 (Appendix I.1) gives the noise scale. At stage-2 iteration 1500 the three Totals lie within 0.07 points of one another. Decomposed by dimension (Table 18, Appendix I.1), the Total hides two opposite effects: chunk pooling buys subject consistency at the price of dynamic degree, and the motion field returns most of that motion while holding consistency (Appendix I.1). Under DMD the extra motion remains (Table 20); on VBench Total, however, the chunk-pooled + motion, 4 positives product scores below both the chunk-pooled, 4 positives product and the two-stage product at iterations 1000 and 2000. The motion field needs at least two positives per prompt, so the main recipe, with one, does not use it (Appendix I). **The motion gain costs consistency, and raters separate the products on stability.** Against the Causal-rCM SF-DMD student the two-stage product's consistency deficit is about two points on the four-dimension mean, mostly motion smoothness and subject consistency. In a blind rating of the four products of Table 3 (Appendix O), raters grade the four alike on whether the motion follows the prompt, so the extra motion comes at no cost in compliance; they separate the products on stability. Raters mark clear instability on one clip in twenty

DMD product, best within 2000 it.	at it.	stage-2 cost	dyn. degree ↑	consistency ↑	motion compliance ↑
Drift Forcing (two-stage)	1000	400 H800-h	90.0	93.15	.77
mean-pooled, 2 positives	1875	130 H800-h	91.7	92.85	.79
TF-dCM	1000	unpublished (10k it., batch 32)	75.8	94.17	.78
no initialization (stage-1 model)	1000	0 (no stage 2)	79.7	93.83	.79

Table 3: Cost and return at each initialization's best checkpoint within 2000 DMD iterations (Table 1); motion compliance: share of clips that blind raters graded as fully following the prompt's motion (100 prompts stratified by motion demand, Appendix O).

for the TF-dCM product; one in seven for the mean-pooled, 2 positives product; one in five for the two-stage product; and one in four for the no-initialization product.

5 Discussion

5.1 What the evidence supports

The field vanishes wherever the student's pooled features match the library's (Appendix L). Under a teacher-forced placement the Causal-rCM SF-DMD student's FVD rises monotonically, while under self-rollout the same student's FVD stays flat and a run from the stage-1 model descends to a minimum and stays there. We interpret the teacher-forced behavior as movement toward a fixed point off the manifold and the self-rollout behavior as convergence to one on it (Figure 6a, Appendix J). The two-stage product's lead is not teacher-fit, on which the consistency-distilled initializations lead; it is what DMD adds on top: motion and prompt-conditioned semantics.

5.2 Which instrument: a measured metric hierarchy

The instruments are ordered by what we observe (Appendices M and N). FVD and VBench invert rankings across our runs (the stiller the model, the more FVD flatters it), so FVD serves only to monitor trajectories. Per-dimension differences are compared with per-dimension noise. The local judge (Appendix N) agrees with VBench on motion, but on visual quality it and VisionReward (Xu et al., 2026) diverge from VBench's aesthetic and imaging-quality dimensions, so neither selects checkpoints; blind raters (Appendix O) order the products on stability as VisionReward does.

5.3 Limitations

Our scope is chunk-wise four-step students on a single 1.3B base. Whether other extractors can carry the field is open (§3.2). The structural account of the teacher-forced failure (§4.4) rests on the within-prompt spread measurement of Appendix G. The field needs a positive clip for every training prompt (Appendix K).

6 Conclusion

Drift Forcing trains the few-step initialization of an autoregressive video distillation stack on the student's own samples, with a drift field in causal-teacher feature space under self-rollout. Under one DMD trainer and one budget it yields the highest VBench Total among all products compared, above the products of the released consistency-distilled initializations and of the stage-1 model, and above the Causal-rCM SF-DMD student. Against the products of the consistency-distilled initializations, the gain is motion and prompt-conditioned semantics; the cost is consistency and teacher-fit. Against the no-initialization product the gain, on VBench, is motion alone: DMD refines what its initialization renders, and the stage-1 model renders clips that move less than the two-stage initialization's. The ablations bound these results (§4.4, §4.5): the field trains under self-rollout and not under teacher forcing; over the two-stage recipe it needs its repulsive term; and a motion field raises dynamic degree before and after DMD, though not the VBench Total after it. Open questions are whether a drift field can replace the consistency-distillation term in joint DMD objectives and whether representation-autoencoder latents (Zheng et al., 2026a; Xie et al., 2026) can replace the extractor.

REPRODUCIBILITY STATEMENT

Appendix A documents the local re-measurement of the released checkpoints on which every external number rests: the two released distilled students with published scores reproduce within 0.06 VBench Total, and their published ordering is preserved. Evaluation protocols are frozen: VBench with 944 prompts $\times$ 5 samples (seeds 0–4) over 16 dimensions with fixed script versions; the precision, coverage and Vendi protocol's prompt grid, seed table, encoder and pooling; and FVD of 128 generated clips against a 2,048-clip library reference. The complete stage-2 and stage-3 configurations, including every difference between our DMD runs and the released recipe, are in Appendix D; the provenance of every training run (initialization, positive library, field configuration, forcing mode, learning rate, iterations) is in Appendix F; full FVD trajectories are in Appendix B. Code, configurations and evaluation scripts are available at `https://anonymous.4open.science/r/DF0E51`.

ETHICS STATEMENT

This work distills existing open text-to-video models and does not introduce new generative capability beyond faster, cheaper inference of what the teachers already produce; it inherits their risks, including misuse for misleading content and training-data biases reflected in outputs. All training and evaluation data are synthetic clips generated from public models with public prompt sets; no personal data was collected. Faster autoregressive generation lowers the cost of producing convincing video.

AI USE STATEMENT

In this work we used generative AI tools to assist in analyzing and interpreting the experimental results, to edit the manuscript for readability, and to write part of the code. We reviewed every output of these tools and take responsibility for the content of the paper.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Hmrishav Bandyopadhyay, Xuanchi Ren, Zijian Huang, Jay Zhangjie Wu, Tianshi Cao, Ruilong Li, Bryan Chu, Sanja Fidler, Yi-Zhe Song, and Zian Wang. Context-matched distillation: Teacher causality for autoregressive video distillation. *arXiv preprint arXiv:2608.13391*, 2026.

Drake Brown, Yuhao Huang, Shih-Hsin Wang, and Bao Wang. Second order drifting models. *arXiv preprint arXiv:2608.07924*, 2026.

Shengqu Cai, Weili Nie, Chao Liu, Julius Berner, Lvmin Zhang, Nanye Ma, Hansheng Chen, Maneesh Agrawala, Leonidas Guibas, Gordon Wetzstein, and Arash Vahdat. Mode seeking meets mean seeking for fast long video generation. In *ICML*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024.

Shuo Chen, Cong Wei, Sun Sun, Tiancheng Shen, Ping Nie, Kai Zou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. In *ICML*, 2026a.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026b.

Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, and Yingcong Chen. ReWorld: An interactive world model with long-horizon memory. *arXiv preprint arXiv:2608.23565*, 2026c.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *ICLR*, 2026.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. DOLLAR: Few-step video generation via distillation and latent reward optimization. In *ICCV*, 2025.

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Haoyu Wang, Zeyue Xue, Songchun Zhang, and Junhao Zhuang. Long-horizon audio-visual generation for persistent stories and interactive worlds. *arXiv preprint arXiv:2608.23383*, 2026.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Javier Gamazo Tejero, Lukas Zbinden, Keyur Sheth, Raghavendra K M, Nadim Daher, Diego Granero Maraña, Filip Binkiewicz, Patrick Thornycroft, Mahdi Azizian, and Sean D. Huver. NVIDIA Cosmos-H-Dreams: Real-time generative physics simulation for surgical robotics. *arXiv preprint arXiv:2608.24199*, 2026.

Yuxuan Gao, Yedong Shen, Shiqi Zhang, Wenhao Yu, Yifan Duan, Jia Pan, Jiajia Wu, Jiajun Deng, and Yanyong Zhang. Drift-based policy optimization: Native one-step policy learning for online robot control. In *ACM International Conference on Multimedia*, 2026a.

Yuxuan Gao, Shiqi Zhang, Yedong Shen, Yifan Duan, Wenhao Yu, Xin Zhang, Siyuan Cao, Jiajun Deng, and Yanyong Zhang. DriftingVLA: Native one-step vision-language-action generation via per-dimension temporal drifting. *arXiv preprint arXiv:2608.29749*, 2026b.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *CVPR*, 2024.

Xingtong Ge, Yi Zhang, Yushi Huang, Dailan He, Xiahong Wang, Bingqi Ma, Guanglu Song, Yu Liu, and Jun Zhang. Salt: Self-consistent distribution matching with cache-aware training for fast video generation. In *ECCV*, 2026.

Arthur Gretton, Li Kevin Wenliang, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the Wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *NeurIPS*, 2025.

Yubo Huang, Sirui Zhao, Xinchen Yao, Zhengye Zhang, Jinyang Huang, Fengqi Cui, Shiwei Wu, and Enhong Chen. DynaForcing: Overcoming dynamic collapse in self-forcing distillation for streaming avatar generation. In *ACM International Conference on Multimedia*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Arkadii Kazanskii, Tatiana Petrova, Andrey Ustyuzhanin, Konstantin Bagrianskii, Aleksandr Puzikov, and Radu State. Friction-augmented drifting models for resource-efficient domain translation, 2026. URL `https://arxiv.org/abs/2604.18194`.

Denis Korzhenkov, Adil Karjauv, Animesh Karnewar, Mohsen Ghafoorian, and Amirhossein Habibian. PyramidalWan: On making pretrained video model pyramidal for efficient inference. In *CVPR*, 2026.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Haobo Li, Yanhong Zeng, Yunhong Lu, Jiapeng Zhu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yujun Shen, and Zhipeng Zhang. AAD-1: Asymmetric adversarial distillation for one-step autoregressive video generation. In *International Conference on Machine Learning*, 2026a.

Jiaxing Li, Kai Zou, Cindy Zhou, Kaichen Huang, Junyao Gao, Zile Wang, Yang Liu, Bin Liu, Bo An, and Yangguang Li. Distillalign: Coordinating mode covering and mode seeking in autoregressive video distillation. *arXiv preprint arXiv:2607.26811*, 2026b.

Zhiqi Li and Bo Zhu. A long-short flow-map perspective for drifting models. *arXiv preprint arXiv:2602.20463*, 2026.

Zian Li, Litong Gong, Borui Liao, Pengfei Liu, Xinyu Wang, Xinyuan Wei, Yifan Gao, Tiezheng Ge, and Muhan Zhang. DUET: A diversity–quality duet of distillation experts for two-step video generation. *arXiv preprint arXiv:2608.09637*, 2026c.

Jiesong Lian, Zixiang Zhou, Ruizhe Zhong, Yuan Zhou, Qinglin Lu, Rui Wang, Long Hu, Yixue Hao, and Baoru Huang. Sara: Semantically adaptive relational alignment for video diffusion models. *arXiv preprint arXiv:2605.07800*, 2026.

Jiuzhou Lin, Fei Zuo, Huan Ouyang, Junlong Wu, Dewen Fan, Boheng Zhang, Huaiqing Wang, Jia Sun, Fan Yang, Houde Liu, Kehai Chen, Min Zhang, Tingting Gao, and Han Li. Joint alignment and distillation for video generation via sample-guided distribution matching. In *ECCV*, 2026.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. Opsd-v: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026a.

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-Hsuan Yang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation. In *International Conference on Learning Representations*, 2026b.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *ICLR*, 2026c.

Susie Lu, Haonan Chen, Weirui Ye, and Yilun Du. Driftworld: Fast world modeling through drifting. *arXiv preprint arXiv:2607.15065*, 2026a.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J. Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *ICCV*, 2025.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *CVPR*, 2026b.

Ge Ya Luo, Gian Mario Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond fvd: An enhanced evaluation metrics for video generation distribution quality. In *ICLR*, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

Weili Nie, Julius Berner, Nanye Ma, Chao Liu, Saining Xie, and Arash Vahdat. Transition matching distillation for fast video generation. In *CVPR*, 2026.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-Game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

Zhuguanyu Wu, Ruihao Gong, Yang Yong, Yushi Huang, Xiangyu Fan, Lei Yang, Dahua Lin, and Xianglong Liu. Sgmd: Score gradient matching distillation for few-step video diffusion distillation. In *ICML*, 2026.

Zhihao Xie, Junfeng Wu, Xinting Hu, Junchao Huang, and Li Jiang. Videorae: Taming video foundation models for generative modeling via representation autoencoders. *arXiv preprint arXiv:2607.14088*, 2026.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. *arXiv preprint arXiv:2502.15681*, 2025.

Yibin Yan, Jilan Xu, Shangzhe Di, Yikun Liu, Yudi Shi, Qirui Chen, Zeqian Li, Yifei Huang, and Weidi Xie. Learning streaming video representation via multitask training. In *ICCV*, 2025.

Yibin Yan, Jilan Xu, Shangzhe Di, Haoning Wu, and Weidi Xie. Omnistream: Mastering perception, reconstruction and action in continuous streams. In *ECCV*, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *ICLR*, 2026a.

Xiaoda Yang, Yuxiang Liu, Kaiwen Zheng, Yuan Liu, Yibo Lai, Shengpeng Ji, Kai Jiang, Jianfei Chen, Shan Yang, Sen Liang, Xiaobin Hu, Shuicheng Yan, Jintao Zhang, Jun Zhu, and Zhou Zhao. TurboT2VA: Fast large-scale text-to-video-audio generation via score-regularized consistency distillation. *arXiv preprint arXiv:2608.24674*, 2026b.

Yongqi Yang, Huayang Huang, Xu Peng, Xiaobin Hu, Donghao Luo, Jiangning Zhang, Chengjie Wang, and Yu Wu. Towards one-step causal video generation via adversarial self-distillation. In *ICLR*, 2026c.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Yuyang You, Yongzhi Li, Jiahui Li, Yadong Mu, Quan Chen, and Peng Jiang. Adaptive video distillation: Mitigating oversaturation and temporal collapse in few-step generation. In *CVPR*, 2026.

Jiaru Zhang, Zeyun Deng, Juanwu Lu, Ziran Wang, and Ruqi Zhang. Analytical correction for subsampling bias in drifting models. *arXiv preprint arXiv:2604.27239*, 2026a.

Jiawei Zhang, Mengfei Xia, Gen Li, and Yuantao Gu. Distilling drifting transformers with representation autoencoders. *arXiv preprint arXiv:2606.15553*, 2026b.

Wenhu Zhang, Kun Cheng, Changyuan Wang, Shiyao Li, Yuechen Zhang, Wenbo Li, Jiajun Zha, Jingyi Zhang, Kang Zhao, and Jiaya Jia. CoDMD: Copula-aware distribution matching distillation for fast video generation. *arXiv preprint arXiv:2606.21982*, 2026c.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. In *NeurIPS*, 2025.

Yuan Zhang, Chenyi Li, Haodong Yu, Guoqing Ma, Jiajun Zha, Yuanming Yang, Bo Wang, Wei Tang, Wenbo Li, Haoyang Huang, and Nan Duan. Teacher-feature drifting: One-step diffusion distillation with pretrained diffusion representations. *arXiv preprint arXiv:2605.07327*, 2026d.

Chuyang Zhao, Yifei Song, Hongfa Wang, Jianlong Yuan, Yuan Zhang, Siming Fu, Zhineng Chen, Huilin Deng, Haoyang Huang, and Nan Duan. Perceptual flow matching for few-step generative modeling. *arXiv preprint arXiv:2607.03524*, 2026a.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026b.

Zhuoran Zhao, Shengju Qian, Tongtong Liang, Xianghao Kong, Songchun Zhang, Junchao Huang, Guian Fang, Xin Wang, Pan Hui, and Anyi Rao. Mask forcing: Improving autoregressive video diffusion distillation via dual-noise masking rollout. *arXiv preprint arXiv:2609.09123*, 2026c.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026a.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *arXiv preprint arXiv:2606.25473*, 2026b.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *ICLR*, 2026c.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *ICML*, 2026. arXiv:2602.02214.

Junhao Zhuang, Shiyi Zhang, Yuxuan Bian, Yaowei Li, Yawen Luo, Yijun Liu, Weiyang Jin, Songchun Zhang, Xianglong He, Xuying Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Self gradient forcing: Native long video extrapolation. *arXiv preprint arXiv:2607.20368*, 2026.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Nenghai Yu. Hiar: Efficient autoregressive long video generation via hierarchical denoising. In *ECCV*, 2026a.

Zhen Zou, Xiaoxiao Ma, Mingde Yao, Jie Huang, LinJiang Huang, and Feng Zhao. Drift-AR: Single-step visual autoregressive generation via anti-symmetric drifting. In *ECCV*, 2026b.

A LOCAL RE-MEASUREMENT OF THE RELEASED CHECKPOINTS

We re-measure seven released checkpoints (six of Causal-rCM (Zheng et al., 2026b) and the Wan2.1-1.3B base) on the frozen VBench protocol of Appendix E (944 prompts $\times$ 5 samples, seeds 0–4, 16 dimensions; the two 50-step models at 944 $\times$ 1). Table 4 compares our Totals with the published values where they exist (Δ: ours minus published). The Causal-rCM four-step SF-DMD student is the dCM-initialized chunk-wise model of Causal-rCM Table 5 (published 84.33). The bidirectional 1.3B row is the Wan2.1-1.3B base model sampled in 50 steps, included for reference only.

Checkpoint	Total	Quality	Semantic	published / Δ
SF-DMD student, 4-step (dCM init)	84.39	85.27	80.85	84.33 / +0.06
SF-DMD student, 2-step	84.26	85.02	81.23	84.30 / -0.04
SF-DMD student, 2-step, noisy-context variant	84.20	84.95	81.21	n/a
TF-sCM (stage 2)	83.14	84.11	79.27	n/a
stage-1 model (1.3B causal, 50-step)	83.09	84.12	78.93	n/a
Wan2.1-1.3B bidirectional base (50-step)	82.06	82.48	80.41	82.78 / -0.72
TF-dCM (stage 2)	82.05	83.24	77.29	n/a

Table 4: Local re-measurement of six released Causal-rCM checkpoints and the Wan2.1-1.3B base on the frozen VBench protocol. The published ordering of the three rows with published values is preserved.

B FULL FVD AND VBENCH TRAJECTORIES

FVD on the frozen 128-prompt protocol is tabulated at the DMD iterations of Tables 5 and 6, a 125-iteration spacing for all but two runs, and plotted in Figure 5. The trajectories add the following to §4.3. The mean-pooled, 2 positives run is the only few-step initialization whose series returns to its pre-DMD baseline (iterations 1750 to 2000), and it rises again by iteration 3000. Among the other runs extended to iteration 3000, each series minimum lies above its baseline, the chunk-pooled field only run's by under 4. By iteration 3000 no VBench Total exceeds its best value from the first 2000 DMD iterations (Table 7); three runs (chunk-pooled field only; chunk-pooled, 4 positives; chunk-pooled + motion, 4 positives) score higher at iteration 2500 than at any earlier evaluated checkpoint. Table 8 gives precision, coverage and Vendi at iteration 2000 for the seven runs carried past it: every product has more coverage at iteration 2000 than at 1000, by .05 to .09 for the drift products other than mean-pooled field only and by .01 for TF-dCM. At iteration 3000 six of the seven Totals lie within 0.65 of one another, with the mean-pooled field only product 0.69 below the lowest of them, while dynamic degree runs from 77.2 (chunk-pooled field only) to 91.4 (mean-pooled, 2 positives).

C THE SMALL-DATA VARIANT

Two initializations (mean-pooled, 2 positives; TF-dCM) were each carried through DMD at two data scales: the 4,096-prompt pool at global batch 8 and VidProM at global batch 64. Each run's FVD is compared with its own pre-DMD baseline (Table 9; Figure 5c,d). At small scale the mean-pooled, 2 positives run reaches 6.6 below its pre-DMD FVD baseline (iteration 4750), whereas the TF-dCM run stays at least 59.7 above its own. Both small-data runs end above their initializations on VBench Total (Table 9); for the mean-pooled, 2 positives run the gain comes with 25.3 points of dynamic degree, gains in the prompt-conditioned dimensions and costs in the appearance and temporal-consistency dimensions. The large-data runs are the mean-pooled, 2 positives and TF-dCM rows of Table 1.

D TRAINING CONFIGURATIONS

Table 10 lists the initializations compared in §4.2, Table 11 the stage-2 design space of §3; the two paragraphs after them give the stage-3 and stage-2 training configurations.

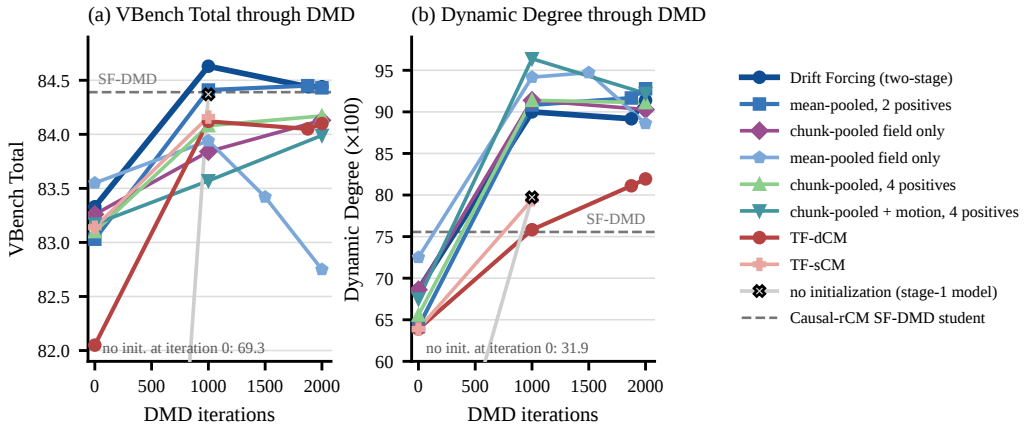

Figure 4: VBench Total (a) and dynamic degree (b) through stage-3 DMD under one trainer, from iteration 0 (the initialization's own VBench evaluation; the stage-1 model's four-step values fall below the plotted range of both plots) to iteration 2000. Names as in Table 1. The Causal-rCM SF-DMD student is the dashed line. Markers are VBench evaluations on the frozen protocol and lines only connect them; Figure 5 gives the FVD trajectories and Table 7 the extension to iteration 3000.

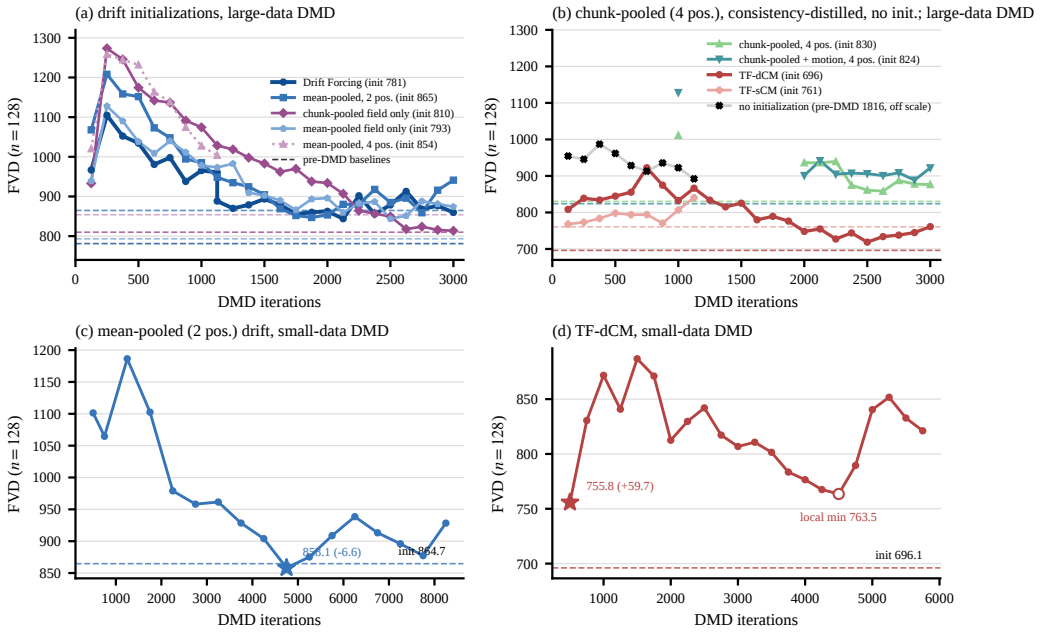

Figure 5: FVD through stage-3 DMD against each run's own pre-DMD baseline (dashed). (a) Drift Forcing (two-stage); mean-pooled field only; chunk-pooled field only; mean-pooled, 2 positives; mean-pooled, 4 positives. (b) chunk-pooled, 4 positives; chunk-pooled + motion, 4 positives; TF-dCM; TF-sCM; no initialization; all under the DMD recipe of §4.1, to iteration 3000 where extended (the stage-1 model's baseline of 1815.9 is off-scale). (c, d) The small-data runs of Appendix C.

Stage 3 (DMD). Every DMD run except the small-data runs of Appendix C uses the Causal Forcing++ trainer and its DMD loss on the filtered and extended VidProM prompt set (248K prompts), at global batch 64 (eight GPUs, batch 1 each, accumulation 8) in bf16, with the bidirectional Wan2.1-14B as the real score and the students' chunking (three-frame chunks, 21 latent frames, 832×480). Each iteration updates the fake score once, and every fifth iteration also updates the student; 1000

it.	Drift Forcing (two-stage)	mean-pooled, 2 pos.	chunk-pooled field only	mean-pooled field only	TF-dCM
init	781.1	864.7	809.9	793.3	696.1
125	966.7	1067.9	933.0	939.2	808.3
250	1104.7	1208.0	1273.5	1127.8	838.9
375	1052.3	1158.5	1245.6	1089.9	834.1
500	1034.5	1151.8	1174.4	1038.1	844.7
625	980.8	1073.0	1142.1	1008.7	855.5
750	997.9	1047.9	1136.8	1039.7	922.4
875	937.9	994.6	1091.4	1011.4	874.7
1000	965.3	985.1	1074.4	978.2	832.5
1125	960.9	948.4	1028.4	973.1	866.5
1250	869.9	934.7	1018.7	982.7	833.8
1375	878.8	924.4	997.8	909.5	815.3
1500	893.6	904.3	983.0	902.5	825.4
1625	878.8	869.0	962.2	890.0	779.9
1750	855.9	851.9	969.7	866.0	789.3
1875	860.0	**847.0**	937.7	893.4	776.3
2000	862.5	853.4	933.9	895.8	748.0
2125	**843.9**	880.1	906.9	859.4	754.8
2250	901.8	880.4	864.1	882.4	727.3
2375	855.6	917.7	856.8	886.6	743.9
2500	878.7	884.8	849.4	**843.8**	**718.7**
2625	912.8	895.7	817.7	851.4	733.9
2750	867.9	859.0	823.8	888.0	737.9
2875	876.7	915.7	815.7	881.2	744.9
3000	859.8	941.1	**813.5**	874.0	761.1

Table 5: FVD readings (n=128) of five of the seven initializations carried to iteration 3000 (the other two are in Table 6) under the DMD recipe of §4.1, by DMD iteration ("init": the pre-DMD baseline; bold: lowest value after DMD begins; names as in Table 1, "pos." = positives per prompt).

it.	mean-pooled, 4 pos.	TF-sCM	no initialization	it.	chunk-pooled, 4 pos.	chunk-pooled + motion, 4 pos.
init	853.9	760.6	1815.9	init	830.2	823.9
125	1020.4	**768.2**	954.5	1000	1011.1	1128.3
250	1258.5	773.1	945.6	2000	936.1	901.0
375	1244.6	783.5	987.0	2125	936.0	941.8
500	1231.9	798.0	961.6	2250	939.6	904.8
625	1163.4	793.8	928.5	2375	874.7	907.6
750	1138.4	794.1	912.9	2500	861.2	906.2
875	1074.5	770.7	935.6	2625	**858.2**	900.9
1000	1027.3	806.8	922.5	2750	887.5	908.7
1125	**1003.9**	840.3	**892.2**	2875	877.8	**888.4**
				3000	876.4	922.1

Table 6: FVD readings (n=128) of the remaining runs: three measured to iteration 1125 (left) and two measured at iteration 1000 and from iteration 2000 to 3000 (right: chunk-pooled, 4 positives; chunk-pooled + motion, 4 positives). Conventions as in Table 5.

DMD iterations process about 77K samples and 2000 about 154K, at about 200 H800-hours per 1000 iterations. Both optimizers are AdamW with β=$(0, .999)$ and weight decay 0.01, at learning rates 2×10^{-6} for the student and 4×10^{-7} for the fake score. DMD timesteps are drawn from $t \in [0.02, 0.98]$ under the shift-5 warp. The student samples in four steps on the same schedule as at inference, and the gradient enters through one randomly chosen step. The exponential moving average (EMA) rate is 0.99 from iteration 200, and gradients are clipped at norm 10. The Causal-rCM SF-DMD student comes from the rCM (score-regularized continuous-time consistency) trainer (Zheng et al., 2026c) as configured by Causal-rCM (Zheng et al., 2026b), run for 3,200 iterations with five fake-score updates per student update (about 2,670 and 530 updates respectively, 204.8K samples) on the Causal-rCM synthetic prompts, with uniform timesteps, backpropagation replayed through all four steps, a power EMA and no clipping. Its compute is unpublished. Batch, optimizers, real score and chunking are the same on both sides. Our runs start the fake-score network from the 1.3B bidirectional base (the released recipe: the 14B one) and use real-score guidance 3.0 (released: 5.0); the fake score runs without guidance.

Stage 2 (initialization). The released TF-dCM initialization is teacher-forced consistency distillation of the 14B causal teacher on the 48-step rectified-flow grid (shift 3, guidance 3), at global

Initialization	best within 2000 (it.)	it. 2000	it. 2500	it. 3000
Drift Forcing (two-stage)	84.63 (1000)	84.44	84.15	84.24
mean-pooled, 2 positives	84.45 (1875)	84.43	83.72	83.59
chunk-pooled field only	84.13 (2000)	84.13	84.17	83.82
mean-pooled field only	83.94 (1000)	82.75	82.65	82.90
chunk-pooled, 4 positives	84.17 (2000)	84.17	84.24	83.96
chunk-pooled + motion, 4 positives	83.99 (2000)	83.99	84.09	83.97
TF-dCM	84.12 (1000)	84.10	83.94	83.97

Table 7: VBench Total of the seven runs extended past iteration 2000: the best value within the first 2000 DMD iterations (iteration in parentheses) and the values at iterations 2000, 2500 and 3000.

Initialization	precision ↑	coverage ↑	Vendi ↑
Drift Forcing (two-stage)	.406	.209	1.291
chunk-pooled field only	.461	.180	1.295
mean-pooled field only	.342	.172	1.281
mean-pooled, 2 positives	.395	.191	1.292
chunk-pooled, 4 positives	.426	.178	1.293
chunk-pooled + motion, 4 positives	.473	.162	1.277
TF-dCM	.461	.234	1.302

Table 8: V-JEPA2 precision, coverage and Vendi at DMD iteration 2000 of the seven products carried past it (protocol of Table 2).

batch 32 for 10k iterations, about 320K processed samples of the Causal-rCM synthetic training set. TF-sCM, the continuous-time counterpart, is trained for 1k iterations at the same batch. Our stage 2 uses the same teacher only as the feature extractor ϕ (residual taps after blocks 1, 9, 19 and 29, no guidance), starts from the stage-1 model (from TF-dCM where stated) and trains on the positive library and the student's own four-step rollouts with four rollouts per prompt; the stage-2 runs behind Table 1 last 1,500 to 2,000 iterations (1,000 + 1,000 in the two-stage recipe) at global batch 8 (2 for the mean-pooled, 2 positives run). The field hyperparameters and the optimizer are in Appendix E, the schedule in §3.3, and the costs in Table 10.

E Reproducibility details

This appendix specifies what §3 and §4.1 summarize.

Feature pass. ϕ is the frozen 14B causal teacher (the Causal-rCM teacher-forced diffusion checkpoint), run once over the full 21-latent-frame clip with the block-causal chunk mask (chunk c attends to chunks $\leq c$) at a fixed rectified-flow time t_ϕ=0.1: the input is $0.9\,x + 0.1\,\epsilon$ with a fresh Gaussian ϵ per call (teacher timestep 100), conditioned on the umT5 embedding of the real prompt (no null or negative prompt). Activations are read from the residual stream after blocks 1, 9, 19 and 29 (zero-based; 1, 7, 15, 23 for the 1.3B extractor). Mean pooling averages all tokens of a tap into one vector. Chunk pooling averages the tokens of each three-frame latent chunk, giving seven vectors per tap. No centering or per-tap normalization is applied. Rollouts and positives take the identical path (same mask, time and conditioning); positives are encoded without gradient and re-encoded every iteration.

Order-one coordinates, kernel and floor. For one (tap, chunk position, prompt) group with n=4 rollout features $g \in \mathbb{R}^D$ and n_{pos} positives p, the scale is the mean Euclidean distance from each rollout feature to every other point of the group (positives and sibling rollouts), and $\tilde{g} = g\sqrt{D}/\text{scale}$, $\tilde{p} = p\sqrt{D}/\text{scale}$; pairwise distances are divided by the same scale and self-pairs are masked. The scale, like the distances and the pairing weights, is computed without gradients and enters Eq. 2 as a constant: the loss is differentiated with respect to g at fixed scale. Pairing weights use the Laplace kernel $\exp(-d/\tau)$ under the row and column softmaxes that define a (§3) and are clamped below at 10^{-6}. For each τ the field V_τ of Eq. 1 has RMS $r_\tau = (\text{mean}_{i,m}\, V_{\tau,im}^2)^{1/2}$ over rollouts i and feature dimensions m and contributes $V_\tau / \max(r_\tau, r_{\text{soft}})$ to V. With r_{soft}=0 this is the per-τ RMS

Initialization × data scale	FVD init	FVD min (it.)	Δ	VBench init → DMD (it.)	P/C init	P/C after DMD
mean-pooled, 2 pos. × small	864.7	858.1 (4750)	−6.6	83.03 → 83.30 (4750)	.518/.318	.283/.154
mean-pooled, 2 pos. × large	864.7	847.0 (1875)	−17.7	83.03 → 84.45 (1875)	.518/.318	.389/.170
TF-dCM × small	696.1	755.8 (500)	+59.7	82.05 → 83.41 (4500)	.789/.586	.449/.227
TF-dCM × large	696.1	748.0 (2000)	+51.9	82.05 → 84.12 (1000)	.789/.586	.596/.223

Table 9: The small-data variant: FVD, VBench Total and V-JEPA2 precision/coverage of two initializations under DMD at two data scales (small = 4,096-prompt pool, batch 8; large = VidProM, batch 64; large-data minima within 2000 DMD iterations, small-data minima over the full run). Δ = series minimum minus the run's own pre-DMD baseline; the large-data runs continue in Table 5 (the TF-dCM run reaches +22.6 at iteration 2500). P/C = V-JEPA2 precision/coverage.

Initialization	VBench Total	Stage-2 cost	Stage-2 machinery
no initialization (stage-1 model), 4-step	69.29	0	none
TF-dCM (Zheng et al., 2026b)	82.05	10k it., b32†	trajectory regression
TF-sCM (Zheng et al., 2026b)	83.14	1k it., b32†	trajectory regression
mean-pooled, 2 positives (2,000 it.)	83.03	130 H800-h	teacher samples only
chunk-pooled + motion, 4 positives (1,500 it.)	83.17	348 H800-h	teacher samples only
Drift Forcing (two-stage), 1,000 + 1,000 it.	83.33	400 H800-h	teacher samples only
Causal-rCM SF-DMD student	84.39	n/a	n/a

Table 10: The initializations on their own: VBench Total on the frozen 944×5 protocol, stage-2 cost and stage-2 machinery. †Iterations and batch as released by Causal-rCM, GPU-hours unpublished. The first row is the stage-1 model through the students' four-step sampler, the start of the no-initialization run. The SF-DMD row is the Causal-rCM DMD product of TF-dCM, shown as a reference.

normalization of §3, whose gain $1/r_\tau$ is unbounded as $r_\tau \to 0$. The floor clamps the denominator, so a τ term with $r_\tau < r_{\text{soft}}$ keeps its direction and has RMS r_τ / r_{soft}. The per-group loss is Eq. 2 averaged over its $n \times D$ entries, i.e. the mean of $(\tilde{g} - \text{sg}[\tilde{g} + V])^2$. The training loss averages the per-group losses over the four taps and, under chunk pooling, the seven chunk positions (4 or 28 terms), then multiplies by a loss scale of 100. The bandwidth set is $\mathcal{T} = \{0.1, 0.2, 0.4, 0.8\}$ with w_{rep}=1; the motion field, where used, has weight w_m=0.09 and bandwidths $\mathcal{T}_m = \{0.05, 0.1, 0.2\}$.

Self-rollout and its gradient. Each stage-2 rollout is the student's own inference: seven chunks of three latent frames, four denoising calls per chunk on the rectified-flow grid $t \in \{1, 15/16, 5/6, 5/8\}$ (the timesteps 1000, 937.5, 833.3, 625 of the DMD stage after its shift-5 warp), each call predicting x_0 and re-noising to the next time with fresh noise. Only the last call of each chunk (t=5/8) carries gradient into the student. The first three run without gradient. During denoising the key-value (KV) cache is read-only. Once a chunk's x_0 is fixed, a no-gradient forward at t=0 appends its keys and values, stored detached, so no gradient crosses chunk boundaries and earlier chunks enter later ones only as fixed context. The four rollouts of a prompt share the prompt embedding and differ only in their Gaussian draws at each step. Unless Appendix F states otherwise, stage-2 runs use one prompt per data-parallel rank, context parallelism 4 over eight GPUs and gradient accumulation 4, i.e. 8 prompts (32 rollouts) per optimizer step; AdamW runs at learning rate 2×10^{-6} (4×10^{-6} in stage 2a and in the runs Appendix F lists), with β=(0, .999) and weight decay 0.01, a 100-step linear warm-up, no gradient clipping and bf16. Every checkpoint, and every hand-off between stages, uses the released recipe's power EMA of the student weights (rate 0.1).

One stage-2 step, in order. (i) For each prompt in the batch, roll out n=4 clips with the student's four-step chunk-wise sampler, keeping gradient only through the last call of each chunk. (ii) Encode the rollouts and the prompt's positives with ϕ and pool the activations at the four taps per chunk (chunk pooling) or per clip (mean pooling). (iii) For each group of one tap, one chunk position and one prompt: rescale the features to order one; form the distance matrix from each rollout to the positives and to its siblings, self-pairs masked; compute a by the row and column softmaxes at each $\tau \in \mathcal{T}$; compute V_τ by Eq. 1, divide it by $\max(r_\tau, r_{\text{soft}})$ and sum over τ. (iv) Take the mean of

Stage-2 objective	Matching	Extra network	Curation	Regime	Cost (GPU-h)
Causal ODE	per-sample	none	1,900 GiB	teacher-forced	11,600
Causal CD	per-sample	none	none	teacher-forced	2,900
Causal DMD-as-init	distributional	critic	none	teacher-forced	n/a*
Drift Forcing	distributional	none	none	self-rollout	400 (two-stage); 130–348 (other configurations)

Table 11: The stage-2 design space. Causal ODE, Causal CD and Causal DMD-as-init are the ODE-regression, consistency-distillation and DMD initializations of Causal Forcing++; "Curation" is the stored teacher-trajectory set the objective requires. "Cost" is stage-2 training compute as reported by Causal Forcing++ (A800-h) or measured by us (H800-h, Table 10); *Cost not reported; Causal Forcing++ reports that this initialization degrades under rollout.

$(\tilde{g} - \mathrm{sg}[\tilde{g} + V])^2$ over the group, average over groups, multiply by 100, backpropagate through ϕ into the student and take one AdamW step.

Hand-offs. Hand-offs between stages (2a to 2b, stage 2 to DMD) load the EMA weights and start a fresh optimizer. Every start of DMD training carries over only the student's weights: the fake-score network is initialized from the bidirectional Wan2.1-1.3B base weights, both optimizers start fresh, and the EMA begins after 200 iterations. Appendix D compares our DMD configuration with the released Causal-rCM recipe.

Checkpoint selection and evaluated points. The "best" column is the highest Total among the checkpoints evaluated at or before iteration 2000. Every DMD run was evaluated on VBench and the V-JEPA2 instruments at iteration 1000; the seven runs extended past iteration 2000 were evaluated on VBench at iterations 2000, 2500 and 3000 and on the V-JEPA2 instruments at iteration 2000, and the tables name every further point where it appears. FVD was measured on every DMD run at the iterations tabulated in Appendix B.

VBench protocol. Prompts are the 944 entries of the Causal-rCM VBench prompt file (943 distinct strings up to case). Generation is conditioned on the file's augmented prompt; scoring uses the original prompt. After lowercasing and removing punctuation, no VBench prompt and no augmented prompt occurs in the 4,096-prompt stage-2 pool or in the VidProM stage-3 set. Each prompt is sampled five times with seeds 0 to 4, the same five seeds for every prompt, with the four-step chunk-wise sampler (mid-times $15/16$, $5/6$, $5/8$; chunks of three latent frames; 81 frames at 832×480, 16 fps; guidance 1, no negative prompt). The 16 dimensions are scored with vbench 0.1.5 and aggregated as in Causal-rCM: per-dimension min-max normalization, weight 0.5 for dynamic degree and 1 otherwise, Quality as the weighted mean of the seven quality dimensions, Semantic as the mean of the nine semantic dimensions, and Total $= (4\,\mathrm{Quality} + \mathrm{Semantic})/5$.

Precision, coverage and Vendi. Features are V-JEPA2 ViT-H (`vjepa2-vith-fpc64-256`, frozen, bf16) on eight frames at frame indices 0, 11, 23, 34, 46, 57, 69 and 80 of the 81-frame clip, resized to 256 pixels by the model's processor; the 1,024 tokens of a clip are pooled into the concatenation of their mean and standard deviation (2,560 dimensions) and L2-normalized, and distances are Euclidean on the unit sphere. The reference set is 512 clips of the bidirectional Wan2.1-14B teacher on 32 prompts drawn at random from VidProM with 16 fixed seeds per prompt, Euler 25 steps, guidance 5, shift 3; each model contributes 512 clips on the same 32 prompts with seeds 0 to 15, four-step sampling at guidance 1. With $r_k(v)$ the distance from reference clip v to its fifth-nearest reference neighbor, precision is the fraction of the model's clips within $r_k(v)$ of some reference clip v (Kynkäänniemi et al., 2019) and coverage the fraction of reference clips with a model clip within their own r_k (Naeem et al., 2020); Vendi is $\exp(-\sum_i \lambda_i \log \lambda_i)$ over the eigenvalues of S/N with the shifted cosine kernel $S = (UU^\top + 1)/2$, where the N rows of U are the model's unit features. The noise floors, the change in each measure when the same checkpoint's clips are re-generated with seeds 100 to 115, are given in the caption of Table 2.

FVD. The 128 prompts are those of library clips 0 to 127 (a *library clip* is a prompt's clip in the Causal-rCM synthetic training set, §3.3) of the 4,096-prompt pool, truncated to 180 characters; each is sampled once at seed 0 with the four-step sampler. The reference statistics are those of library clips 0 to 2,047 (teacher samples at Euler 100 steps, guidance 5, shift 3), decoded to 81 frames at

832×480; features are the 400-dimensional Inflated 3D ConvNet (I3D) features of the StyleGAN-V evaluation code over all 81 frames, and the score is the Fréchet distance between the Gaussian fits. The prompts and reference clips come from the training library; the score is compared between runs scored on the same prompts and used to follow each run's trajectory.

F PROVENANCE OF THE STAGE-2 RUNS

The positive library is drawn from the bidirectional 14B teacher: *stock* positives are each prompt's clip in the Causal-rCM synthetic training set (the positive of §3.3); *re-generated* positives are new teacher samples drawn with a different seed; *adapted* positives come from a 720p teacher library of the same prompts, downsampled to 480p; a *self-positive* is the prompt's own library clip serving both as the ground-truth clip the teacher-forced rollout is anchored to and as the positive.

Every stage-2 run trains under self-rollout at global batch 8 and learning rate 2×10^{-6} with the 14B causal teacher as ϕ and no floor ($r_{\text{soft}}=0$), except where stated below.

The main recipe and its controls run from the stage-1 model on the 4,096-prompt pool with one stock positive per prompt. Stage 2a trains the mean-pooled field at learning rate 4×10^{-6} to iteration 1,000 for the hand-off and to 2,000 as the mean-pooled field only initialization. Stage 2b continues the 1,000-iteration checkpoint with the chunk-pooled field and the floor $r_{\text{soft}}=0.10$ for another 1,000. The chunk-pooled field only initialization is trained with the stage-2b field (chunk pooling with the floor $r_{\text{soft}}=0.10$) from the stage-1 model for 2,000 iterations, and the two-stage, attraction only initialization (Appendix G.3) repeats both stages with $w_{\text{rep}}=0$ in the translation-invariant form of Appendix G.

The positive-library comparison of Appendix G.1 starts from the stage-1 model with mean pooling; all runs but the last use the 512-prompt pool. Four positive sets are compared. With a single stock positive the run lasts 2,500 iterations; with stock plus adapted, which gives the mean-pooled, 2 positives initialization, it lasts 2,000 (extended to 2,500 for the placement probe of Appendix J); both of these run at batch 2. The run with stock plus one re-generated positive and the run with stock, adapted and two re-generated positives each last 2,000 iterations; the iteration-1,500 checkpoint of the latter is the mean-pooled, 4 positives initialization. The wide-pool run, stock plus adapted over 4,096 prompts at 4×10^{-6}, runs for 1,500.

Four chunk-pooled runs start from the stage-1 model on the same pool with the four-positive library: chunk-pooled, 4 positives (2,000 iterations); chunk-pooled + motion, 4 positives (2,000); the chunk-pooled + motion field with the 1.3B causal teacher as ϕ (1,200); and the chunk-pooled + motion field under teacher forcing (400). The initializations of Table 1 are the iteration-1,500 checkpoints of the first two.

All runs from TF-dCM except the teacher-forced matched setup use the chunk-pooled + motion field with four positives on the 512-prompt pool. The base run, the reference for the ablations of Appendix G, lasts 1,500 iterations. Its variants are: the floor $r_{\text{soft}}=0.10$ (1,500 iterations); the true-prefix runs $K=1, 2, 3$, in which the field is withheld from the first K ground-truth chunks (1,500 each); and the teacher-forced matched setup, the chunk-pooled field without motion on one self-positive (500).

The placement probes of Appendix J train the Causal-rCM SF-DMD student and the stage-1 model for up to 1,000 iterations under the one-step placement, in which the field acts on the one-step prediction from a noised library clip; the Causal-rCM SF-DMD student is also trained under self-rollout. The probes from the stage-1 model use chunk pooling, a halved bandwidth, the field restricted to $t \leq 0.3$, or attraction only. The teacher-forced matched setup was also run with the repulsion weight varied (Appendix G).

The DMD runs start from the initializations of Table 1, from the mean-pooled, 4 positives initialization (the 512-prompt-pool initialization of Appendix I.1 that is not in Table 1) and from the two-stage, attraction only initialization (Appendix G.3); the small-data runs of Appendix C run DMD from the mean-pooled, 2 positives initialization and from TF-dCM at the smaller data scale; the checkpoints evaluated on each are listed in Appendix E.

G ABLATION RESULTS

The ablations of §4.4 vary five axes (field, rollout, library, schedule and start), all on the frozen protocols of §4.1. The paragraphs below summarize each axis; the subsections after them give the field, rollout and library results in full, and the schedule and start results are in Tables 1 and 7. The floor and true-prefix results are relative to the base run, the chunk-pooled + motion, 4 positives run from TF-dCM (Appendix F); the repulsion, extractor and teacher-forced results name their own reference runs.

Field. The repulsive term is not dispensable: over the two-stage recipe its removal costs teacher-fit before DMD and consistency after (Appendix G.3). The floor $r_{\text{soft}}{=}0.10$ lowers FVD by 24.2 at iteration 500 and 42.9 at iteration 1000, and at iteration 1000 leaves precision, coverage and VBench Total equal to the base run's, within .03 (precision and coverage) and 0.1 (VBench Total) of it at the other matched points; it is part of stage 2b of the main recipe. The 1.3B causal teacher as ϕ ties the 14B one on FVD by iteration 1200 (855.6 against 855.3) and is within a fifth of a VBench point of it.

Rollout. The teacher-forced variant is inert from the stage-1 model (FVD 1815.9 to 1715.4 over 400 iterations) and destructive from TF-dCM (696.1 to 2389.7). The true-prefix variants $K{=}1, 2, 3$ stay within 16 FVD of the base run at iteration 1000 while their precision and coverage stay at the initialization's values and their VBench Total lies below the base run's at all nine matched points: the prefix chunks freeze, and the recipe uses $K{=}0$.

Library. The 720p-adapted positive raises the FVD minimum relative to the single stock positive and changes VBench by +0.12 (83.03 against 82.91). The runs with re-generated positives at two and four per prompt reach the same minimum (852), at stage-2 iterations 1100 and 1950. The mean-pooled field only initialization (one positive per prompt over the 4,096-prompt pool) has higher coverage (.492) and stage-2 VBench (83.55) than any 512-prompt run with four positives from the stage-1 model (at most .441 and 83.31).

Schedule. At DMD iteration 1000 the product of the mean-pooled field only initialization (2,000 stage-2 iterations) scores 83.94 against the two-stage product's 84.63, with more motion (94.2 against 90.0) and less consistency (91.86 against 93.15). It falls to 82.75 by iteration 2000 and ends at 82.90 at iteration 3000, against 84.24, whereas every run that ends with a chunk-pooled field stays within 0.5 of its best Total through iteration 3000 (Table 7).

Start. Run from the stage-1 model and from TF-dCM, the chunk-pooled + motion, 4 positives configuration reaches VBench 83.17 and 83.31 at stage-2 iteration 1500, and from TF-dCM it ends 37.7 FVD above and .076 coverage below its own initialization.

G.1 DESIGN SPACE: POSITIVE LIBRARY

Table 12 collects the mean-pooled runs from the stage-1 model. The 720p-adapted positive reaches its FVD minimum earlier than the single stock positive (stage-2 iteration 2000 against 2500) at VBench Total 83.03 against 82.91, and the two runs with re-generated positives reach the same minimum (852.4 and 851.8) at iterations 1100 and 1950. The mean-pooled field only initialization (one positive per prompt over the 4,096-prompt pool) scores best in every column.

G.2 THE TRANSLATION-INVARIANT FORM OF THE REPULSION ABLATION

With $A_i = \sum_j a_{ij}^{\text{pos}}$, $B_i = \sum_k a_{ik}^{\text{sib}}$, and μ_i^+ and μ_i^- the pairing-weighted means of the positives $\tilde{p}_j$ and of the sibling features $\tilde{g}_k$, Eq. 1 becomes $V_\tau(g_i) = A_i B_i(\mu_i^+ - w_{\text{rep}}\mu_i^-)$: a common shift of all features cancels only at $w_{\text{rep}}{=}1$, so the repulsion ablation of Appendix G.3 uses the translation-invariant displacement form $A_i B_i[(\mu_i^+ - \tilde{g}_i) - w_{\text{rep}}(\mu_i^- - \tilde{g}_i)]$, which equals Eq. 1 at $w_{\text{rep}}{=}1$.

Configuration	FVD minimum (it.)	best VBench Total (it.)	P / C (it.)
single positive (512 pool)	811.4 (2500)	82.91 (2500)	.600 / .342 (2500)
mean-pooled, 2 positives	864.7 (2000)	83.03 (2000)	.518 / .318 (2000)
stock + re-generated (2 positives)	852.4 (1100)	83.11 (2000)	.566 / .404 (2000)
mean-pooled, 4 positives	851.8 (1950)	83.22 (2000)	.565 / .375 (1500)
stock + adapted (2 positives, 4,096 pool)	779.3 (1500)	83.39 (600)	.566 / .404 (1500)
mean-pooled field only (1 positive, 4,096 pool)	**756.8** (1000)	**83.55** (2000)	**.652 / .492** (2000)

Table 12: The mean-pooled design space (from the stage-1 model, no motion field): the FVD series minimum, the best VBench Total and V-JEPA2 precision/coverage, each with its stage-2 iteration. Bold: best in column.

G.3 Repulsion over the two-stage recipe (from the stage-1 model)

To test whether the main recipe needs the repulsive term, the two-stage, attraction only initialization repeats stage 2a and stage 2b from the stage-1 model with $w_{\text{rep}}=0$ in the translation-invariant form and everything else as in the main recipe (Appendix F); its stage-2b endpoint is then carried through the same DMD configuration to DMD iteration 1000. Table 13 pairs each value with the main recipe's and Table 14 gives the FVD series. Without repulsion, stage 2a is lower on FVD at iteration 250 (906.8 against 927.6) and then stalls while the main recipe descends, ending far higher (921.3 against 756.8 at iteration 1000), with a lower VBench Total (82.64 against 83.28) and less precision and coverage (.463/.297 against .648/.488). Stage 2b closes most of the FVD gap and all of the VBench gap, and dynamic degree and consistency are equal as well, so at the hand-off to DMD the two initializations differ only on the teacher-fit instruments. Through DMD the attraction-only FVD series stays above the main recipe's at every checkpoint. At iteration 1000 the product keeps less precision ($-.094$) and coverage ($-.028$; noise floors .0097 and .0078, Table 2), scores higher on dynamic degree, lower on each of the four consistency dimensions, and lower on Total (83.84 against 84.63). What the repulsive term buys before DMD is teacher-fit. What the product loses without it is consistency and Total.

Checkpoint		FVD	VBench Total	dyn. / consistency	P / C
stage 2a	it. 1000	921.3 vs. 756.8	82.64 vs. 83.28	70.3/94.00 vs. 71.7/94.01	.463/.297 vs. .648/.488
stage 2b	it. 1000	831.9 vs. 781.1	83.34 vs. 83.33	68.3/94.69 vs. 68.6/94.57	.629/.436 vs. .678/.473
DMD product	it. 1000	1065.6 vs. 965.3	83.84 vs. 84.63	91.4/91.70 vs. 90.0/93.15	.410/.109 vs. .504/.137

Table 13: The two-stage, attraction only initialization against Drift Forcing (two-stage) at the end of stage 2a and of stage 2b (stage-2 iteration 1000 of each) and after DMD (DMD iteration 1000). Each cell gives the attraction-only value first.

FVD	it. 250	500	750	1000
stage 2a, attraction only	906.8	843.6	826.5	921.3
stage 2a, main recipe	927.6	838.1	791.9	756.8
stage 2b, attraction only	832.7	842.5	825.0	831.9
stage 2b, main recipe	756.7	795.2	777.3	781.1

FVD through DMD (it.)	125	250	375	500	625	750	875	1000	1125
attraction only	994.3	1362.9	1304.1	1224.0	1157.4	1170.9	1138.4	1065.6	1093.3
main recipe	966.7	1104.7	1052.3	1034.5	980.8	997.9	937.9	965.3	960.9

Table 14: FVD series of the two-stage, attraction only initialization against Drift Forcing (two-stage) through stage 2a, stage 2b and DMD ($n=128$).

G.4 Floor and true prefix (from TF-dCM)

Tables 15 and 16 give the results of the floor and true-prefix variants from TF-dCM. K is the number of ground-truth prefix chunks (of seven) on which the field is not applied; $K=0$ is the base run, and the floor variant is the base run with $r_{\text{soft}}=0.10$ on the per-τ RMS normalization. The floor variant is ahead of the base run on FVD (-42.9 at iteration 1000) with precision/coverage equal to the base

run's to three digits and the same VBench Total. Every prefix variant is within 16 FVD of the base run at iteration 1000, and its precision and coverage stay at the initialization's values while the base run's move. On VBench no prefix variant exceeds the base run at any of the nine matched points: the prefix chunks are effectively frozen, the prefix variants forgo the spatial-relationship, object-class, imaging-quality and scene gains the base run makes from its initialization, and gain only dynamic degree.

	FVD at iteration						VBench Total at iteration		
Variant	250	500	750	1000	1250	1500	500	1000	1500
init (TF-dCM)			696.1					82.05	
base run (chunk-pooled + motion, 4 positives)	684.3	713.3	706.3	717.2	718.5	733.8	83.01	83.19	83.31
true prefix $K{=}1$	693.8	698.5	714.1	713.4	719.0	722.8	82.77	83.09	83.12
true prefix $K{=}2$	692.9	702.6	701.8	713.7	737.1	728.1	82.90	83.11	83.19
true prefix $K{=}3$	675.2	707.9	695.3	701.8	698.1	719.6	82.84	83.06	83.13
floor $r_{\text{soft}}{=}0.10$	690.5	689.1	689.0	**674.3**	706.4	698.8	82.91	83.19	83.26

Table 15: Normalization floor and true prefix from the TF-dCM start: FVD and VBench Total.

	iteration 500			iteration 1000			iteration 1500		
Variant	P	C	Vendi	P	C	Vendi	P	C	Vendi
init (TF-dCM)				.789 / .586 / 1.324					
base run (chunk-pooled + motion, 4 positives)	.691	.482	1.322	.697	.523	1.321	.705	.510	1.322
true prefix $K{=}1$	.777	.588	1.329	.770	.590	1.332	.771	.559	1.330
true prefix $K{=}2$	.785	.588	1.331	.783	.580	1.331	.777	.582	1.330
true prefix $K{=}3$	.797	.570	1.331	.770	.592	1.329	.758	.568	1.328
floor $r_{\text{soft}}{=}0.10$	.707	.508	1.322	.697	.523	1.323	.721	.531	1.325

Table 16: Normalization floor and true prefix from the TF-dCM start: V-JEPA2 precision, coverage and Vendi.

G.5 Extractor capacity

The chunk-pooled + motion, 4 positives initialization, trained with the 1.3B causal teacher as ϕ (taps after blocks 1, 7, 15, 23, the same bandwidth set), matches the 14B-extractor run on FVD by stage-2 iteration 1200 (855.6 against 855.3) ; at iteration 1200 it scores VBench Total 82.88 against the 14B-extractor run's 83.06 at 1250, and precision, coverage and Vendi of .631, .430 and 1.329 against the 14B run's .617, .441 and 1.333 at 1500.

G.6 Forcing regime

The teacher-forced variant of the chunk-pooled + motion, 4 positives run (teacher-forced rollouts, everything else identical) lowers FVD only from 1797.6 to 1715.4 over iterations 100–400, from an initialization at 1815.9, while the same run under self-rollout goes from 1765.5 to 1079.8. The matched teacher-forced construction from TF-dCM (single self-positive, no motion field) holds a plateau near 700 through iteration 150 and then climbs to 2389.7 by iteration 500 (precision/coverage .063/.020); its variants with other repulsion weights hold the same plateau, so the plateau does not depend on the repulsion weight. Under teacher forcing the field's rollout side is a conditional estimate anchored to a ground-truth prefix, and the batch statistics the field needs collapse: same-prompt rollouts share one anchor. The within-prompt spread statistic, the dispersion of a prompt's rollout features in the field's coordinates, is 0.08–0.37 under teacher forcing against 0.35–0.59 under self-rollout over the first 100 iterations of the two runs from the stage-1 model above.

H Cross-encoder measurement control

§3.2 excludes external encoders from the *training* loop. A separate question is whether the nearest-neighbor structure the field is built on, measured in the teacher's feature space, is an artifact of that feature family; a measurement-only control answers it. The control samples two students from the

TF-dCM start at stage-2 iteration 500, the base run and its true-prefix $K{=}2$ variant, four times per prompt on the FVD prompts, and compares each sample with the four positives the field trained on. It measures the rate at which a student token's nearest neighbor among positives and sibling rollouts is a positive, divided by chance ($4/7$), over 448 time-ordered tokens per clip, in six feature spaces: V-JEPA2 ViT-H, DINOv2 ViT-L/14, VideoMAEv2-B, VideoMAE-B, and the recipe's own extractor at its four taps (14B and 1.3B). Table 17 reports it with a teacher index (the same statistic computed for a fresh teacher sample of the prompt in place of the student) and the field's own chunk-level statistic. All twelve rows fall between 0.48 and 0.64 times chance, the 14B teacher band (0.53–0.61) lies inside the external band (0.48–0.61), and the two students agree row by row within 0.02; the field's own chunk-level rate, a raw rate against chance $4/7$, is 0.12–0.17 under the teacher and 0.15–0.21 under the external encoders. The result does not depend on the extractor family.

Extractor	D	base run		its true-prefix $K{=}2$ variant		chunk-level
		NN-to-positive / chance	index	NN-to-positive / chance	index	chunk-level rate
V-JEPA2 ViT-H	1280	0.55	0.88	0.53	0.88	0.19†
DINOv2 ViT-L/14	1024	0.52	0.95	0.51	0.95	0.15†
VideoMAEv2-B	768	0.48	0.91	0.48	0.93	0.20†
VideoMAE-B	768	0.60	0.81	0.61	0.84	0.21†
teacher-ϕ 14B block 1	5120	0.61	0.85	0.59	0.88	0.16
teacher-ϕ 14B block 9	5120	0.55	0.84	0.55	0.87	0.13
teacher-ϕ 14B block 19	5120	0.53	0.81	0.53	0.84	0.12
teacher-ϕ 14B block 29	5120	0.56	0.80	0.56	0.84	0.12
teacher-ϕ 1.3B block 1	1536	0.64	0.86	0.63	0.89	0.17
teacher-ϕ 1.3B block 7	1536	0.59	0.83	0.58	0.86	0.15
teacher-ϕ 1.3B block 15	1536	0.57	0.83	0.57	0.87	0.14
teacher-ϕ 1.3B block 23	1536	0.60	0.83	0.59	0.86	0.15

Table 17: Cross-encoder measurement control on two TF-dCM-start students at stage-2 iteration 500 ($n{=}128$ prompts $\times$ 4 rollouts per student): nearest-neighbor-to-positive rate over chance; index: the same rate for a fresh teacher sample of the prompt in place of the student; chunk-level: the chunk-vector rate the field itself uses (chance $4/7$; † = block-mean stand-in for the recipe's chunk vectors).

I THE MOTION FIELD AND THE THREE-WAY COMPARISON

The content field of §3 matches where the student's chunks sit in teacher-feature space but not how features move. The motion field is a second drift field on the temporal increments of the chunk-pooled features,

$$\delta_c \;=\; f_{c+1} - f_c, \qquad \mathcal{L}_{\text{motion}} = \big\|\tilde{\delta} - \text{sg}\big[\tilde{\delta} + V^{\mathcal{T}_m}(\delta)\big]\big\|_2^2, \tag{3}$$

where f_c is a clip's chunk-pooled feature at chunk position c, $\tilde{\delta}$ are the increments rescaled as in §3, and $V^{\mathcal{T}_m}$ is the field V of §3 computed on increments over the bandwidth set $\mathcal{T}_m$ (six increment positions for seven chunks) and the total objective $\mathcal{L} = \mathcal{L}_{\text{drift}} + w_m\mathcal{L}_{\text{motion}}$ ($w_m{=}0.09$). The motion field matches the *positives'* motion statistics rather than penalizing motion. A smoothness regularizer would instead recreate the static bias that the motion field is meant to fix. Increments are roughly an order of magnitude smaller than the chunk features themselves, so the motion field carries its own bandwidth set $\mathcal{T}_m$ (Eq. 3; bandwidths in Appendix E). It requires at least two positives per prompt. The main recipe uses one positive per prompt and therefore has no motion field.

The three-way comparison of §4.5 (Table 18) is three initializations from the stage-1 model at matched budget, data and a fixed four-positive library on the 512-prompt pool: mean-pooled, 4 positives; chunk-pooled, 4 positives; and chunk-pooled + motion, 4 positives, measured on VBench at stage-2 iterations 1000, 1250 and 1500 against the adjacent-checkpoint noise probe. The probe is the change between two checkpoints 50 iterations apart on a plateaued run, the mean-pooled, 4 positives run at iterations 1950 and 2000. On the distribution axis, chunk pooling still raises precision from .565 to .617 and coverage from .375 to .412 at stage-2 iteration 1500, while the motion field leaves precision unchanged by then (.617 for both) and keeps a coverage lead of .029 (Table 2). On VBench the chunk-pooled + motion, 4 positives run stays above the chunk-pooled, 4 positives run to iteration 2000 (83.31 against 83.21). Before DMD, the chunk-pooled + motion, 4 positives initialization exceeds the Causal-rCM sCM-initialized SF-DMD student on coverage (.441 against

.393) and Vendi (+0.025) while trailing it on precision (−0.043; noise floors .0097 / .0078 / .0001, Table 2).

I.1 THE THREE 512-PROMPT-POOL INITIALIZATIONS UNDER DMD

The three 512-prompt-pool initializations at stage-2 iteration 1500 were carried through the DMD recipe of §4.1 (Table 20). Both chunk-pooled products' FVD is lower at DMD iteration 2000 than at 1000 (Table 6), and their VBench Totals peak at iteration 2500. At iteration 1000 the products do not rank as their initializations did: the mean-pooled, 4 positives product and the chunk-pooled, 4 positives product tie on VBench once the dynamic-degree contribution to the Total is removed, the former keeps the most precision, and the chunk-pooled + motion, 4 positives product has the largest FVD rise (823.9 to 1128.3). None reaches the mean-pooled, 2 positives product or the two-stage product within 2000 DMD iterations. Table 19 repeats the decomposition of Table 18 at stage-2 iterations 1000 and 1250, before DMD.

stage-2 it.	mean	chunk	chunk + motion
1000	82.78	82.66	82.81
1250	83.02	82.99	83.06
1500	83.13	83.10	83.17

(a) VBench Total trajectories.

Effect at it. 1500	subj. cons.	dyn. degree	Total
pooling (chunk − mean)	+0.66	−2.50	−0.03
motion (+motion − chunk)	−0.13	+1.94	+0.07
probe	0.09	0.55	0.00

(b) Effect decomposition at stage-2 it. 1500 (normalized scores ×100, as in Figure 3; subj. cons.: subject consistency).

Table 18: The three 512-prompt-pool initializations before DMD: VBench Total (a) and its decomposition (b). Columns mean, chunk and chunk + motion: mean-pooled, 4 positives; chunk-pooled, 4 positives; chunk-pooled + motion, 4 positives. Probe: the change between two checkpoints 50 iterations apart on a plateaued run.

it.	Effect	subj. cons.	dyn. degree	Total
1000	pooling (chunk − mean)	+0.43	−2.50	−0.12
1000	motion (+motion − chunk)	−0.08	+2.22	+0.15
1250	pooling (chunk − mean)	+0.50	−2.78	−0.03
1250	motion (+motion − chunk)	−0.11	+1.39	+0.07

Table 19: The decomposition of Table 18 at stage-2 iterations 1000 and 1250, before DMD (normalized VBench scores ×100, as in Figure 3).

Initialization	point	FVD	P / C / Vendi	VBench Total (dyn.)
mean-pooled, 4 pos.	pre-DMD, it. 1500	853.9	.565 / .375 / 1.324	83.13 (68.1)
mean-pooled, 4 pos.	DMD, it. 1000	1027.3	.473 / .102 / 1.248	83.81 (86.7)
chunk-pooled, 4 pos.	pre-DMD, it. 1500	830.2	.617 / .412 / 1.334	83.10 (65.6)
chunk-pooled, 4 pos.	DMD, it. 1000	1011.1	.383 / .123 / 1.269	84.08 (91.4)
chunk-pooled, 4 pos.	DMD, it. 2000	936.1	.426 / .178 / 1.293	84.17 (91.1)
chunk-pooled, 4 pos.	DMD, it. 2500	861.2	.428 / .209 / 1.299	84.24 (81.1)
chunk-pooled, 4 pos.	DMD, it. 3000	876.4	.381 / .197 / 1.305	83.96 (85.8)
chunk-pooled + motion, 4 pos.	pre-DMD, it. 1500	823.9	.617 / .441 / 1.333	83.17 (67.5)
chunk-pooled + motion, 4 pos.	DMD, it. 1000	1128.3	.453 / .072 / 1.238	83.57 (96.4)
chunk-pooled + motion, 4 pos.	DMD, it. 2000	901.0	.473 / .162 / 1.277	83.99 (92.2)
chunk-pooled + motion, 4 pos.	DMD, it. 2500	906.2	.375 / .152 / 1.290	84.09 (91.7)
chunk-pooled + motion, 4 pos.	DMD, it. 3000	922.1	.375 / .164 / 1.294	83.97 (88.1)

Table 20: The three 512-prompt-pool initializations through DMD (pos.: positives per prompt).

J PLACEMENT PROBES

In the one-step placement the field acts on the student's one-step prediction from a noised library clip; this placement is teacher-forced but distinct from the teacher-forced rollouts of Appendix G. Under it, every probe run from the stage-1 model (chunk pooling, halved bandwidth, the field restricted to $t \leq 0.3$, attraction only) lowers FVD below the initialization's by iteration 250 and then raises it monotonically, and the FVD of the Causal-rCM SF-DMD student (800.6 before training) rises monotonically (from 819 at the first checkpoint to 1150 at iteration 1000). One interpretation is that under teacher forcing the field's fixed points lie off the data manifold and a start far from the data gains quality only while *crossing* it. On the student's own four-step rollouts, the placement used everywhere in this paper, the Causal-rCM SF-DMD student's FVD stays flat (from 817 at the first checkpoint to 796 at iteration 1000) and the run from the stage-1 model descends monotonically to 864.7 at iteration 2000 and stays within 17 of that minimum 500 iterations later; this run yields the mean-pooled, 2 positives initialization of Table 1. Placement, not kernel scale, timestep range or pooling, is what makes the field train (§4.4). Figure 6 summarizes the account and places the measured initializations and DMD products on the precision-coverage plane.

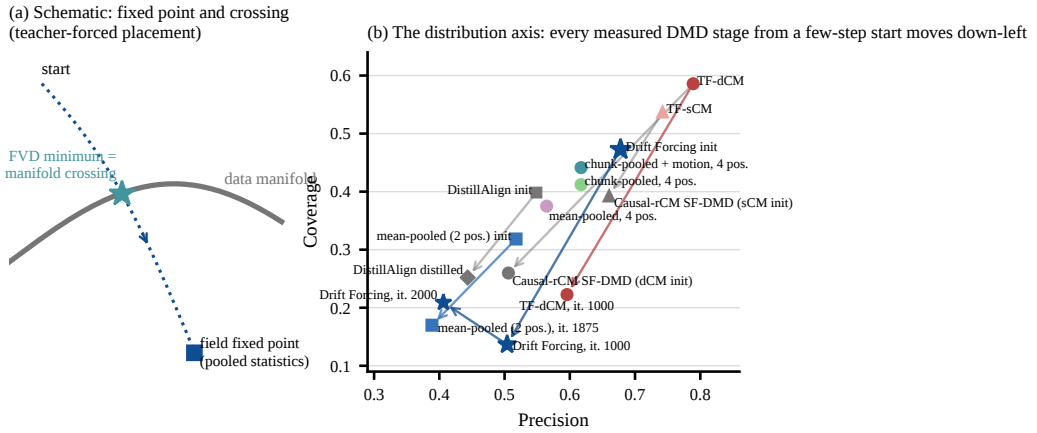

Figure 6: (a) The geometric account, an interpretation of the FVD trajectories (§5.1): under the teacher-forced placement quality peaks where the descent crosses the manifold; under the self-rollout placement the series minimum is a destination. (b) Precision-coverage plane (V-JEPA2, bidirectional-teacher reference): every DMD stage from a few-step initialization moves its student down-left (the stage-1 start, not few-step, moves up-right), and the chunk-pooled + motion, 4 positives initialization arrives pre-DMD with coverage above the Causal-rCM sCM-initialized SF-DMD student.

K IMAGE-DOMAIN PRECURSOR STUDY

The form of the field was chosen in an image-domain study on EDM ImageNet-64. Its question was whether the double-softmax attraction and repulsion of drifting should be replaced by score estimators (Stein, spectral-Stein, plug-in kernel) on either side. The answer was no on both sides. On the repulsion side every batch estimator degenerates, at four images per class in a few-thousand-dimensional feature space, to a push away from the mean and loses three properties of the plain construction: local collision avoidance, without which the run stagnates; distance gating, without which it diverges; and multi-scale decorrelation, without which the useful bandwidths are diluted by the others. On the attraction side, grafting the teacher's denoising target into the field fails because that target is a posterior mean outside the convex hull of realized sample features. The rule is that *the field's target must stay inside the convex hull of realized sample features and preserve local structure*; the video field keeps the plain construction for this reason. Its corollary, that a teacher score cannot free the attraction term from positives, is why the field depends on a positive library (§5.3). A dose study on the same setup (Fréchet inception distance, FID: 11.15, 9.43, 7.37 and

7.02 for 1, 2, 4 and 8 positives per class) shows diminishing returns past four, which is why no 512-prompt run uses more than four positives per prompt.

L PRECISION AND COVERAGE: WHAT THE OBJECTIVES IMPLY

Table 2 reports two facts about the distribution axis: every drift initialization enters DMD with less precision and coverage than the consistency-distilled ones, and every few-step initialization leaves DMD with less coverage than it entered. Both follow from what the two stage-2 objectives fix and what the DMD objective moves. Each step below is read against the rows of that table.

What the field constrains. Write $\Phi_{\ell,c}$ for the map from a clip to its rescaled pooled feature at tap ℓ and chunk position c, so that $\tilde{g}_i = \Phi_{\ell,c}(\text{rollout } i)$ and $\tilde{p}_j = \Phi_{\ell,c}(\text{positive } j)$. Write q for the student's distribution over clips of one prompt and p_{lib} for the empirical distribution of that prompt's positives. The field of Eq. 1 is anti-symmetric in its two sample sets, so it vanishes everywhere once the two feature clouds coincide, $(\Phi_{\ell,c})_{\#}q = (\Phi_{\ell,c})_{\#}p_{\text{lib}}$ (Deng et al., 2026). For a closely related variant of the double-softmax field (positives and negatives normalized separately, without self-pair masking) with a Gaussian kernel, the converse holds in the population limit under a finite-mass assumption (Gretton et al., 2026, Theorem 3); the recipe's joint normalization with self-masking and its Laplace kernel (Appendix E) lie outside that theorem, and Deng et al. (2026) note that the converse fails for general fields. Both statements are about the space in which the field is evaluated. Every student whose pushforward through $\Phi_{\ell,c}$ matches the library's at every tap and chunk position is therefore an equilibrium of stage 2, and since $\Phi_{\ell,c}$ averages the tokens of a chunk, two students that differ only inside a fiber of $\Phi_{\ell,c}$ are equally at equilibrium. Consistency distillation regresses the student along the teacher's ODE trajectories, and its minimizer is the teacher's own sample map. Precision and coverage are measured through a third map Ψ, the pooled V-JEPA2 clip feature of the precision and coverage protocol (Appendix E), against the teacher's own clips. The consistency-distilled minimizer is teacher-fit under any Ψ, whereas agreement under $\Phi_{\ell,c}$ constrains Ψ only through the part of Ψ that is a function of $\Phi_{\ell,c}$. Table 2 puts TF-dCM at .789/.586 and the two-stage initialization at .678/.473 before DMD, and changing Φ alone, from mean to chunk pooling on the four-positive library, moves the pair from .565/.375 to .617/.412 while VBench Total moves by 0.03 (Table 18).

What the field targets. The target is the library, not the teacher's distribution. Each prompt has $n_{\text{pos}} = 1$ positive in both stages of the main recipe and 1 to 4 in the 512-prompt runs, so the attraction term of Eq. 1 pulls each rollout toward one library feature or a convex combination of at most four, and the repulsion term is the only term that holds the four rollouts of a prompt apart. The reference holds 16 teacher clips of each prompt, and a student that renders a prompt as variations on at most four library clips reaches them only through the spread the repulsion term maintains. Removing that term over the two-stage recipe lowers coverage from .473 to .436 before DMD and from .137 to .109 after, with precision moving the same way (.678 to .629, .504 to .410). Deng et al. (2026) report the failure of attraction alone in the image domain, and in the positive-dose study of Appendix K FID improves from one to eight positives per class, as more positives widen the target. The mean-pooled field only initialization, with one positive per prompt over 4,096 prompts, has higher coverage than any four-positive 512-prompt run (.492 against at most .441, Table 2).

What DMD moves. Stage 3 minimizes the reverse KL divergence from the student to the teacher (Yin et al., 2024b), and its gradient is the difference of the two scores at the student's own noised samples (timestep weighting omitted):

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}_{t,\,x_t}\Big[\big(s_{\text{fake}}(x_t,t) - s_{\text{real}}(x_t,t)\big)\,\tfrac{\partial x_t}{\partial \theta}\Big]. \tag{4}$$

The expectation in Eq. 4 runs over what the student produces: the update moves those samples toward higher teacher density relative to the student's own, and teacher mass far from every student sample contributes nothing. Coverage of the reference is therefore inherited from the start rather than produced by the stage, and the reference clips in the teacher's low-density regions, which only a spreading student reaches, are the first to lose their nearest student clip. In Table 2 every few-step start loses coverage, and the two consistency-distilled starts, which enter with the most, leave with the most: .223 and .213 against .137 for the two-stage start and at most .170 for any drift start. The

stage-1 model's four-step rendering is the one start whose coverage rises (.049 to .170); it starts off the teacher's support, and the same transport that sharpens the other starts carries it onto the support.

The three statements give the direction of the differences that Table 2 reports between drift and consistency-distilled starts, between the full and the attraction-only field, and between an initialization and its product.

M METRIC HIERARCHY: THE EVIDENCE

Three observations underlie the protocol of §5.2. First, FVD and VBench invert rankings across our runs, and across the three 512-prompt-pool initializations the inversion is monotone in dynamic degree: the stiller the model, the more FVD flatters it relative to VBench. Static content has fewer temporal ways to be wrong under a distributional distance while VBench charges for the missing motion, so FVD serves as a trajectory monitor. Second, per-dimension VBench differences are read against per-dimension noise: the adjacent-checkpoint probe (two checkpoints 50 iterations apart on the mean-pooled, 4 positives run) ranges from below printing precision on the Total to 0.014 (1.4 points) on human action. Third, VBench and the V-JEPA2 instruments attribute the three-way comparison differently (pooling raises precision and coverage but leaves VBench flat; the motion field raises VBench, and by stage-2 iteration 1500 its precision gain is gone while a coverage lead of .029 remains), so both are reported and neither instrument chooses the checkpoint at which the other is reported.

N REFERENCE INSTRUMENTS: VISION-LANGUAGE-MODEL JUDGES AND VISIONREWARD

Two learned preference instruments were run on the frozen-protocol clips as references. The first is VisionReward-Video (Xu et al., 2026), the protocol of Causal Forcing and Causal Forcing++, which scores 29 yes/no questions with the released weights. The second is a Reward-Forcing-style judge (Lu et al., 2026b) that scores text alignment, dynamics and visual quality on a 1–5 scale from eight uniformly sampled frames at temperature 0. It was run locally with Qwen3.5-397B-A17B on every fourth VBench prompt (236 clips per model) and, on a 100-prompt high-motion set, with four further judges (GPT-5.6, Claude Opus 5, Gemini 3.8 Flash, Qwen3.8-Max). Across the products scored, the dynamics axis of four of the five judges tracks VBench dynamic degree (Spearman 0.84 to 0.94), and the visual-quality axis follows VisionReward rather than VBench aesthetic or imaging quality; VisionReward rewards shape preservation and stability. Table 21 gives paired per-clip differences against the TF-dCM product at iteration 1000. The local judge's dynamics scale agrees with VBench on the motion gain of the drift-initialized products and on the alignment gain of the mean-pooled, 2 positives product; VisionReward has no motion axis. On visual quality the local judge and VisionReward agree with each other and diverge from VBench's aesthetic and imaging dimensions, which score the two-stage and TF-dCM products equal.

Product − TF-dCM product (it. 1000)	VisionReward (×100)	alignment	dynamics	visual quality
Drift Forcing (two-stage), it. 1000	−1.00 [−1.48, −0.51]	+0.03 [−0.07, +0.14]	+0.10 [0.00, +0.20]	−0.11 [−0.21, −0.01]
mean-pooled, 2 pos., it. 1875	−0.11 [−0.55, +0.35]	+0.15 [+0.05, +0.25]	+0.29 [+0.18, +0.40]	−0.03 [−0.12, +0.06]
chunk-pooled, 4 pos., it. 2000	−0.34 [−0.82, +0.15]	+0.16 [+0.07, +0.26]	+0.35 [+0.25, +0.44]	−0.03 [−0.12, +0.06]
chunk-pooled + motion, 4 pos., it. 2000	−0.18 [−0.63, +0.26]	+0.06 [−0.03, +0.14]	+0.28 [+0.17, +0.38]	−0.03 [−0.11, +0.06]
no initialization, it. 1000	−1.03 [−1.56, −0.49]	−0.03 [−0.14, +0.07]	−0.09 [−0.20, +0.02]	−0.20 [−0.30, −0.11]
TF-sCM, it. 1000	−0.69 [−1.12, −0.23]	−0.02 [−0.11, +0.08]	−0.01 [−0.11, +0.09]	−0.01 [−0.09, +0.08]
Causal-rCM SF-DMD student	+0.10 [−0.27, +0.46]	+0.03 [−0.06, +0.12]	−0.03 [−0.12, +0.06]	+0.07 [−0.02, +0.16]

Table 21: Reference instruments on the VBench clips (every fourth prompt, n=236 per model): mean paired per-clip difference against the TF-dCM product at iteration 1000 with bootstrap 95% intervals, for VisionReward-Video (×100) and the three 1–5 scales of the local judge (pos.: positives per prompt).

O BLIND RATING OF THE FOUR DMD PRODUCTS

The motion-compliance column of Table 3 comes from a blind rating of the four DMD products of that table, each at its best checkpoint within 2000 DMD iterations. 100 VBench prompts were stratified by motion demand (20 asking for stillness, 20 for small, localized motion, 40 for large motion or locomotion, 20 carrying a camera or temporal instruction), with one sample per prompt, so that the four clips of a prompt share the initial noise. The four clips played side by side in a 2×2 grid, synchronized, with positions balanced across products and no product identified. Raters graded each clip on whether its motion follows the prompt (fully, partly, or not) and whether its subject and scene stay stable (stable, slightly unstable, or clearly unstable), then ranked the four overall, ties allowed. Six raters produced 397 ratings over the 100 prompts; rates are averaged within a prompt and then across prompts, with 95% bootstrap intervals over prompts (Table 22).

On motion compliance the four products' rates lie within one another's intervals: 0.77 to 0.79 of their clips are graded as fully following the prompt and 0.91 to 0.94 as at least partly, and raters grade the four clips of a prompt alike in three ratings out of four. The drift products' larger motion (dynamic degree in Table 3) is not accompanied by lower compliance. The rates spread only on the camera-instruction stratum, where the TF-dCM product follows the instruction least often (0.48 of its clips against 0.57 to 0.62).

Stability is where the products separate, and it separates them in the order of VisionReward in Appendix N: clear instability is graded on 0.05 of the TF-dCM product's clips; 0.15 of the mean-pooled, 2 positives product's; 0.19 of the two-stage product's; and 0.24 of the no-initialization product's. The TF-dCM product's interval lies below the other three, whose intervals overlap, and the gap in the share of stable clips to the TF-dCM product is widest on the prompts asking for small, localized motion (hands and faces). When a rater grades two clips differently on stability but equally on compliance, the overall rank favors the more stable clip in 0.94 of such cases. Across products the mean rank puts the TF-dCM product first, with the lowest mean rank on 49 of the 100 prompts (0.40 of first places once ties are split), and the mean-pooled, 2 positives product second (Table 22); behind the TF-dCM product, the intervals of neighboring products overlap. Raters' stability grades track VBench consistency only weakly: per clip, the VBench consistency dimensions that cover the rated prompts (subject consistency, motion smoothness, temporal flickering) correlate with the stability grade at Spearman 0.15 or below, and across products the two instruments order the two-stage product and the mean-pooled, 2 positives product oppositely. On the 32 rated prompts that lie in VBench's dynamic-degree prompt set, the clips graded stable still move more often for the drift products: VBench scores 0.85 of the two-stage product's stable clips as moving; 0.88 of the mean-pooled, 2 positives product's; 0.75 of the TF-dCM product's; and 0.69 of the no-initialization product's, so the extra motion is not carried by the unstable clips.

Raters agree with one another: when two raters separate the same pair of clips they name the same winner in 0.79 of cases on compliance, 0.92 on stability and 0.76 overall.

	motion follows the prompt			subject and scene stability			overall	
DMD product	fully	partly	not	stable	slightly unstable	clearly unstable	first	mean rank
Drift Forcing (two-stage)	.77 [.71, .83]	.14	.09	.54	.28	.19 [.13, .24]	.17	2.72 [2.62, 2.81]
mean-pooled, 2 positives	.79 [.74, .84]	.15	.06	.65	.21	.15 [.10, .20]	.24	2.48 [2.38, 2.59]
TF-dCM	.78 [.73, .84]	.15	.07	.82	.13	.05 [.02, .09]	.40	2.13 [2.02, 2.24]
no initialization (stage-1 model)	.79 [.73, .84]	.14	.07	.60	.16	.24 [.17, .30]	.19	2.67 [2.56, 2.80]

Table 22: Blind rating of the four DMD products of Table 3 (100 prompts, 397 ratings): share of clips at each grade of the two per-clip questions, share of first places (ties split) and mean overall rank (1 best, ties averaged), averaged within and then across prompts, with 95% bootstrap intervals over prompts for the values discussed in the text.

P THE INITIALIZATION SLOT AGAINST JOINT LOSSES AND NO-INITIALIZATION PIPELINES

Joint losses. DistillAlign anchors DMD with a consistency-distillation term (Li et al., 2026b). A drift field could serve as that anchor, so the two approaches compose; our contribution is the ini-

tialization family. *No-initialization pipelines.* For causal-teacher DMD without a few-step initialization stage (CMD (Bandyopadhyay et al., 2026), whose student starts from the causal teacher's weights; ReWorld and JoyAI-Echo (Chen et al., 2026c; Duan et al., 2026) likewise distill from their teacher-forced causal models), three facts bear on the conclusion that such a stage is unnecessary: the initialization cost does not vanish but moves into a Diffusion-Forcing teacher finetune (Chen et al., 2024) of about 8K iterations against 3–5K for the distillation run; the reported gain over bidirectional scoring comes from CMD's own ablation, whose bidirectional-scoring row moves quality and dynamic degree in opposite directions, a coupling that the hierarchy of §5.2 keeps apart; and the pipeline exists only as image-to-video on a different base model, so no number transfers to the text-to-video setting studied here. Where a bidirectional-teacher DMD recipe and a fixed compute budget apply, as in the Causal Forcing++ setting, a cheap samples-only initialization is the practical option; our results show that DMD builds on it.

Q What DMD adds: per-dimension decomposition and qualitative results

Figure 3 compares the two-stage product at iteration 1000 with the TF-dCM product at iteration 1000 and with the Causal-rCM SF-DMD student across the 16 VBench dimensions. The gains concentrate in dynamic degree (+14.2 against the TF-dCM product) and the prompt-conditioned dimensions; the costs are in temporal consistency and multiple objects. Against its own initialization the Total change outside dynamic degree is -0.02: DMD from this initialization gains motion and prompt-conditioned semantics and loses temporal consistency. Against the no-initialization product at iteration 1000 the two-stage product scores +10.3 dynamic degree, -1.17 motion smoothness, -0.67 temporal flickering, -0.49 subject consistency and -0.41 background consistency (normalized scores $\times 100$, the scale of Figure 3): the no-initialization product moves less and is smoother in time. Prompt-conditioned semantics are DMD's contribution: in stage 2 the prompt enters through the student's rollout, the teacher's feature extraction and the choice of positive clip, and no term scores the match between clip and prompt; the teacher's prompt-conditioned score enters the objective only in DMD. Figure 7 shows before/after DMD frames on frozen-protocol prompts.

Figure 8 places the four DMD products of Table 3 side by side on prompts that carry an explicit motion instruction, same prompt and same initial noise for the four rows: on these four prompts the two-stage product carries out the instruction (the two walks, the acceleration, the glide) where the other three hold the pose or move less. The prompts come from the VBench set.



Figure 7: Qualitative before/after DMD on frozen-protocol prompts (same prompt, same protocol): the mean-pooled, 2 positives initialization at stage-2 iteration 2,000 (rows 1, 3) renders cleanly but conservatively, while its small-data DMD product at iteration 4750 (rows 2, 4; the small-data run of Appendix C) commits to large, prompt-consistent motion. Frames 0, 26, 53, 80 of 81.



Figure 8: Frame-by-frame comparison of the four DMD products of Table 3, each at its best checkpoint within 2000 DMD iterations. Rows within a block share the prompt and the initial noise; frames 0, 16, 32, 48, 64, 80 of 81. The instructed motions are a walk in a snowstorm, a dog's walk, a truck's acceleration and a sailboat's glide.