PROMOTING DETERMINISM PHYSICS-CONDITIONED VIDEO GENERATION VIA CONVERGENT FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-conditioned video generation is emerging as a promising paradigm for learning world models that can predict physically plausible futures from initial observations and explicit physical constraints. A distinctive property of these tasks is that the conditioning signals—such as the initial scene, object states, velocities, masses, and interaction parameters—often strongly constrain the future evolution, leading to conditional target distributions that are substantially more concentrated than those in generic video generation. Nevertheless, existing generative approaches typically retain a condition-independent standard Gaussian source. We argue that this default choice is poorly matched to strongly conditioned generation in two respects: its scale can be substantially misaligned with the remaining conditional variation, and its unstructured randomness fails to exploit information already available in the conditioning signal, imposing unnecessary learning burden on the generative flow. To address these limitations, we introduce **SCFM**, a task-aware source construction for conditional flow matching that adapts the source to both the residual uncertainty and the structural information provided by the condition. By starting the transport from a more informative yet non-degenerate distribution, SCFM enables the flow model to focus its capacity on modeling the remaining conditional dynamics rather than reconstructing structure already specified by the input. Beyond controlled physical supervision, we further introduce **VPhysBench**, a real-world benchmark for physics-conditioned video prediction, to evaluate whether such source design principles transfer to realistic physical interactions. Experiments on VPhysBench demonstrate that SCFM achieves better performance across multiple distributional metrics with greater learning efficiency, highlighting source distribution design as an important and underexplored dimension of efficient conditional video generation.

1 INTRODUCTION

Physics-conditioned video generation aims to predict future observations from an initial scene and explicit physical information, including object states, velocities, masses, and interaction parameters. Such models could support world modeling by connecting visual predictions to the quantities that govern physical evolution. The task requires more than plausible appearance and smooth motion: a predicted interaction must respond to the specified conditions while representing uncertainty due to measurement error in the recorded physical quantities. Learning this relationship efficiently is challenging because videos are high-dimensional, and small physical differences can change an entire future trajectory.

Strong physical conditioning nevertheless provides substantial predictive structure. With exact physical inputs, deterministic dynamics and fixed rendering specify a unique simulated future. In the real world, measurement uncertainty can straddle a critical threshold, making sliding and sticking, falling and remaining supported, or maintaining and losing contact compatible with the same observation. Within each outcome, small measurement errors yield similar trajectories when the dynamics have limited sensitivity over the prediction horizon. Sufficiently separated outcomes can then form distinct narrow modes. The generator must therefore exploit concentration within a physical branch while preserving uncertainty across branches. Treating every condition as a single deterministic target would discard plausible outcomes, whereas ignoring local concentration would waste information supplied by the condition.

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.

Recent physical video generators improve controllability by introducing explicit physical inputs into generative models. Force Prompting incorporates local forces and global wind, PhyCo uses spatially aligned physical properties, and PhyParam introduces instance-level parameter control (Gillman et al., 2025; Narayanan et al., 2026; Li et al., 2026). These advances primarily concern conditioning interfaces and generative supervision. In a standard flow-matching formulation, including that used by PhyParam, transport still begins from condition-independent Gaussian noise. Although the velocity network receives the physical inputs, the source itself carries no instance-specific prediction. The generative flow must consequently recover predictable scene and motion structure as well as model the uncertainty that remains. This allocation of learning effort motivates revisiting the source when training time and model capacity are limited.

Two mismatches are particularly relevant. First, an unbranched source coupled independently to a multimodal target can produce overlapping linear CFM supervision paths with conflicting velocity labels. A field conditioned only on the observed input averages the corresponding velocity labels at a given intermediate state and time, without an explicit indication of the physical outcome. Second, unit Gaussian variation can be poorly matched to the much smaller variation within an outcome, even when separation between outcomes makes the full target broadly dispersed. Conditional-prior methods already demonstrate the value of moving beyond a fixed Gaussian (Issachar et al., 2025; Chen et al., 2025; Kim et al., 2026). Our focus is to use the physical outcome structure to address both the ambiguity across branches and the scale mismatch within them.

We introduce **SCFM**, a task-aware source construction for conditional flow matching. Using outcome labels during training, SCFM learns condition-dependent branch probabilities, predicts a center for each branch, and fits its source scale to observed within-branch variation. The resulting non-degenerate sources place the starting distribution around outcome-specific predictions while retaining local randomness. We fit and freeze these source predictors before training a shared velocity network conditioned on the observed input and branch. At inference, the model samples a branch from its predicted probabilities and transports a sample from that branch's source. This construction lets the flow refine conditional dynamics and distributional shape from an informative starting point, while retaining multiple physically plausible outcomes.

Joint source–flow training can also lead to vanishing source variance, as observed in prior work (Kim et al., 2026). Under the assumptions in Appendix **??**, we prove that the optimal population CFM risk decreases linearly to zero with the Gaussian source standard deviation for non-point-mass targets. This identifies an objective-level incentive for collapse and motivates SCFM's separate source fitting and subsequent freezing.

Evaluating this design requires repeated physical realizations under explicit, recorded conditions. Existing benchmarks address physical commonsense, real-world future prediction, and parameter recovery (Bansal et al., 2025; Motamed et al., 2025; Khanbayov et al., 2026), but public resources tailored to this conditional-distribution learning problem remain limited. We introduce **VPhysBench**, a real-world benchmark containing approximately 5,000 video examples across ten physical scenarios. For each recorded condition, at least ten physical realizations are collected, supporting evaluation of variation among repetitions rather than comparison with only one observed future. The benchmark pairs explicit physical conditions with recorded interactions and provides training and test data within each scenario for evaluating conditional prediction and learning efficiency.

Experiments on VPhysBench demonstrate more efficient learning across multiple distributional metrics, with comparisons and ablations supporting branch-aware source construction.

Our contributions are threefold:

- **SCFM.** We propose a conditional source construction that separates physical outcome branches and adapts their source means and scales, improving learning efficiency in physics-conditioned video generation.
- **VPhysBench.** We construct a real-world benchmark with explicit physical measurements and repeated recordings per condition to evaluate conditional video distributions and physical grounding.
- **Source-collapse theory.** We prove an exact source-scale law for the optimal population CFM objective under stated assumptions, providing a theoretical explanation for previously observed collapse (Appendix **??**).

2 Related Work

Physics-conditioned video generation. Force Prompting, PhyCo, and PhyParam incorporate explicit forces, physical properties, and object parameters into video generators (Gillman et al., 2025; Narayanan et al., 2026; Li et al., 2026). These advances focus on conditioning and physical controllability, while condition-independent Gaussian initialization remains common. For strongly conditioned prediction, such a source leaves the model to reconstruct structure already specified by the input and transport broad noise into narrow physical outcome distributions. This mismatch can impose unnecessary learning effort, motivating a source that exploits the predictive information in the condition.

Source distribution design. CPD constructs informative conditional priors, CAR-Flow introduces condition-aware reparameterization, and CSFM learns conditional Gaussian sources, demonstrating the importance of source design for flow matching (Issachar et al., 2025; Chen et al., 2025; Kim et al., 2026). These general approaches leave room to exploit the particular structure of physical prediction: several distinct outcomes may remain possible, while each outcome occupies a concentrated distribution. SCFM addresses both levels by separating physical branches and adapting source centers and scales within each branch. Source learning also raises the issue of collapse. CSFM reports source-variance collapse during joint training, while CAR-Flow analyzes related zero-loss degeneracies (Kim et al., 2026; Chen et al., 2025). We extend these findings with a rigorous quantitative explanation of the scale bias: under the stated assumptions, the optimal population CFM risk decreases linearly to zero as the Gaussian source scale vanishes, with a strictly positive slope for non-point-mass targets. This identifies an objective-level incentive for collapse; the exact characterization and proof are given in Appendix **??**.

3 Problem Formulation

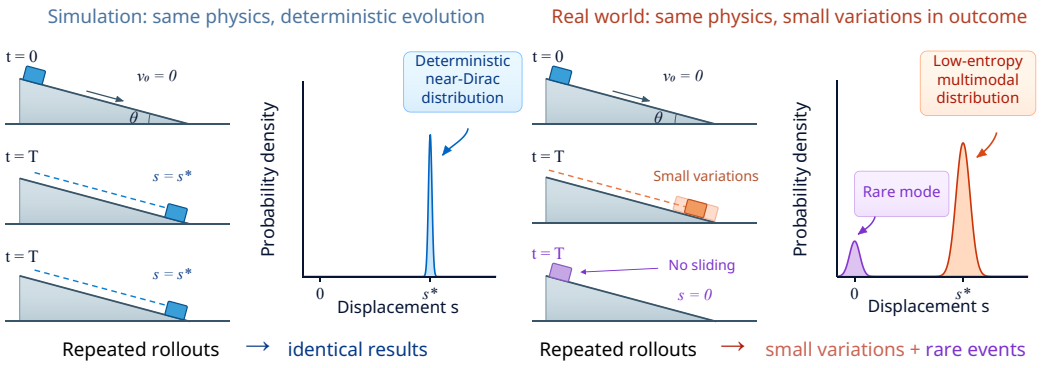


Figure 1: **Simulation with sufficient conditions.** Complete conditioning under deterministic dynamics and fixed rendering specifies a unique future video. The spike denotes a point mass in a schematic projection of video space.

3.1 Physics-Conditioned Video Prediction

Let $C = (I_0, p, \ell)$ collect an initial observation I_0, recorded physical quantities p, and text conditioning ℓ. We consider conditions that record all task-relevant physical variables, with finite precision in their measured values. For a future video $\mathcal{V} = (I_1, \ldots, I_H)$, the task is to learn $P(\mathcal{V} \mid C)$ and generate futures consistent with C. We write c for a realization of C.

Simulation. With exact parameter values, deterministic dynamics, and fixed rendering, c specifies a unique video $f_{\text{sim}}(c)$. The conditional target is therefore $P(\mathcal{V} \in \cdot \mid C = c) = \delta_{f_{\text{sim}}(c)}$, where δ_y is the point mass at y.

Real-world multimodality. A measured parameter p_j constrains its true value to an uncertainty interval $[p_j^-, p_j^+]$. Even a small interval can span a physical threshold, such as the onset of sliding or

loss of contact. If values on both sides remain plausible given $C = c$, the same recorded condition can support different outcome branches: a block may stick or slide, or an object may remain supported or fall. Let $B \in \{1, \dots, K\}$ index these branches, with $\pi_k(c) = P(B = k \mid C = c)$ and $P_k(\cdot \mid c) = P(\mathcal{V} \in \cdot \mid C = c, B = k)$. Then

$$P(\mathcal{V} \in \cdot \mid C = c) = \sum_{k=1}^{K} \pi_k(c) P_k(\cdot \mid c), \qquad \sum_{k=1}^{K} \pi_k(c) = 1. \tag{1}$$

When these branches produce sufficiently separated videos, they can form distinct modes of the conditional target.

Concentration within each branch. Fixing a branch resolves the qualitative outcome, while the measured conditions constrain the remaining physical variation. We focus on finite prediction horizons for which the dynamics within a branch have sufficiently limited sensitivity: small measurement errors then induce only small trajectory changes. Repeated realizations within that branch consequently cluster around similar motions, yielding a concentrated component $P_k(\cdot \mid c)$. If the differences between branches exceed this internal spread, the target exhibits multiple narrow peaks. Strong conditioning thus limits variation within each outcome without necessarily resolving which outcome occurs; the total conditional variance can still be large. At inference, Section 4 samples a branch from probabilities predicted using C and models its remaining variation.

3.2 Conditional Flow Matching

A fixed encoder E maps the future video to a latent $Y = E(\mathcal{V}) \in \mathbb{R}^D$, with conditional target distribution $P(\cdot \mid C)$. Conditional flow matching (CFM) (Lipman et al., 2022; Tong et al., 2023) trains a velocity field from sampled source–target pairs. Given C, draw $X_0 \sim q_0(\cdot \mid C)$ and $Y \sim P(\cdot \mid C)$ independently, and sample $t \sim \mathcal{U}(0, 1)$ independently of the endpoints and condition. Here t is the flow time, distinct from the video frame index. The linear path and its sample-wise velocity are

$$X_t = (1 - t)X_0 + tY, \qquad U = Y - X_0. \tag{2}$$

The network learns to predict this velocity through

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{C,Y,X_0,t}\left[\left\| v_\theta(X_t, t, C) - (Y - X_0) \right\|_2^2\right]. \tag{3}$$

Training requires only endpoint and time sampling, without integrating an ODE. For fixed endpoint distributions with finite second moments, the optimal unconstrained predictor is the conditional mean $u_t(x, c) = \mathbb{E}[U \mid X_t = x, C = c]$, averaging the supervision velocities at each intermediate state and time.

At inference, a fresh draw $Z_0 \sim q_0(\cdot \mid C)$ is evolved by $\mathrm{d}Z_t/\mathrm{d}t = v_\theta(Z_t, t, C)$ and then decoded. For a Dirac target, the ideal endpoint is understood as the weak limit $t \uparrow 1$, without requiring a regular, invertible flow at $t = 1$. The standard choice is $q_0 = \mathcal{N}(0, I)$; a conditional source instead adapts the starting distribution to C. Section 4 applies this construction within each outcome branch z, using independent endpoints conditional on (C, z). Marginalizing the shared branch generally makes the endpoints dependent given C alone.

4 Method

SCFM models uncertainty across physical outcomes with a branch predictor, and variation within each outcome with an adapted source and a shared flow. It predicts branch probabilities, source means, and source scales from the condition, then fits the flow with these predictors frozen.

4.1 Predicting Branch Probabilities

We use the latent $Y = E(\mathcal{V}) \in \mathbb{R}^D$ from Section 3, with fixed encoder E and decoder Dec. The index $z \in \{1, \dots, K\}$ denotes a model branch representing outcomes under condition c. With a prescribed K and no provided branch labels, we first obtain branch pseudo-labels z by unsupervised clustering of the training video representations, using consistent cluster indices across conditions

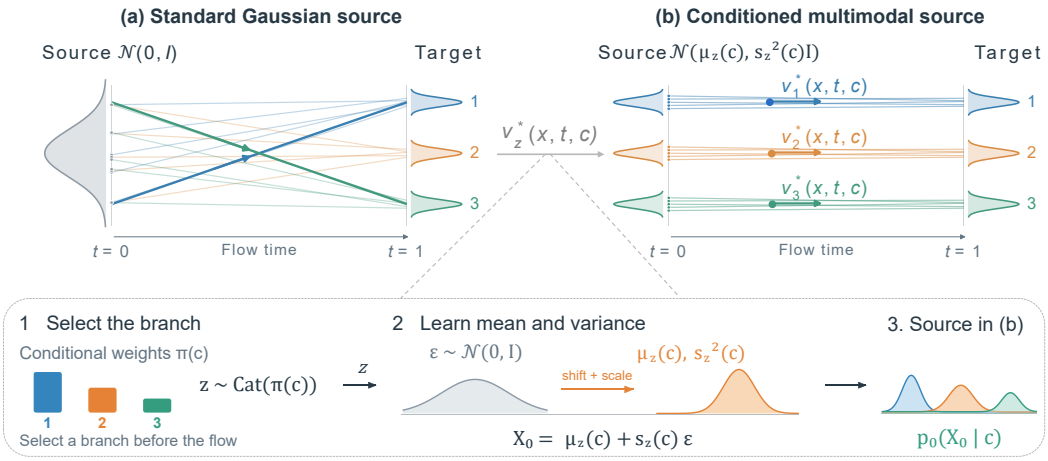


Figure 2: **Overview of SCFM.** For a given physical condition c, the target distribution can contain multiple narrow outcome modes. **(a)** Transport from a standard Gaussian source mixes supervision paths across modes, creating potentially conflicting velocity targets while requiring substantial contraction within each mode. **(b)** SCFM constructs a condition-dependent source for each branch and conditions the shared velocity network on the branch index, reducing cross-branch ambiguity and matching the scale of within-branch variation. **Bottom:** At inference, a branch z is sampled from the learned conditional weights $\pi(c)$, and Gaussian noise is shifted and scaled as $X_0 = \mu_z(c) + s_z(c)\varepsilon$. The branch remains fixed throughout integration. Source predictors are fitted before flow training and subsequently frozen. Paths are schematic illustrations of the source–target coupling.

and fixing these assignments during subsequent source fitting and flow training. Its generated distribution is

$$P_{\theta,\phi,\eta,\psi}(\mathrm{d}y \mid c) = \sum_{z=1}^{K} q_\phi(z \mid c) P_{\theta,\eta,\psi}(\mathrm{d}y \mid c, z), \tag{4}$$

where each component transports a branch-specific source through the shared flow.

For training conditions $\mathcal{C}_{\text{train}}$, group repeated futures by their complete condition c, including the initial observation, and cluster-assigned pseudo-label z. Let $\mathcal{I}_{cz}$ index a group, $n_{cz} = |\mathcal{I}_{cz}|$, and $n_c = \sum_z n_{cz}$. A softmax predictor learns the empirical branch frequencies:

$$\widehat{\pi}_{cz} = \frac{n_{cz}}{n_c}, \qquad \mathcal{L}_{\text{gate}} = -\frac{1}{|\mathcal{C}_{\text{train}}|} \sum_{c \in \mathcal{C}_{\text{train}}} \sum_{z=1}^{K} \widehat{\pi}_{cz} \log q_\phi(z \mid c). \tag{5}$$

Conditions receive equal weight. The moment objectives below instead average uniformly over valid groups $\mathcal{G}_{\text{train}} = \{(c, z) : n_{cz} \geq 2\}$. They require repeated observations to estimate within-branch variation. Branch selection represents differences between outcomes; source scales represent variation within them (Appendix **??**).

4.2 Predicting Source Means

We predict the conditional branch mean $m_z(c) = \mathbb{E}[Y \mid c, z]$ from empirical group means:

$$\overline{Y}_{cz} = \frac{1}{n_{cz}} \sum_{i \in \mathcal{I}_{cz}} Y_i, \qquad \mathcal{L}_{\text{center}} = \mathbb{E}_{(c,z) \sim \mathrm{Unif}(\mathcal{G}_{\text{train}})} \left[\frac{\|\mu_\eta(c, z) - \overline{Y}_{cz}\|_2^2}{D} \right]. \tag{6}$$

The mean predictor and gate may share condition features. For independent source–target endpoints and fixed source scale, the branch mean minimizes expected squared transport displacement, providing a predictive starting point; the identity and its interpretation are given in Appendix **??**.

4.3 Matching Source Variance

Each branch uses an isotropic source whose variance matches the coordinate-averaged target variance $\tau_z^2(c) = \operatorname{tr}\operatorname{Cov}(Y \mid c, z)/D$:

$$q_0(\cdot \mid c, z) = \mathcal{N}\big(\mu_\eta(c, z), s_\psi^2(c, z) I_D\big), \qquad s_\psi^2(c, z) \approx \tau_z^2(c). \tag{7}$$

This scalar matches variation in latent coordinates, not the full covariance; the transported distribution need not be Gaussian. We estimate the scale target around the empirical group mean,

$$\widehat{\tau}_{cz}^2 = \frac{1}{D(n_{cz} - 1)} \sum_{i \in \mathcal{I}_{cz}} \|Y_i - \overline{Y}_{cz}\|_2^2, \tag{8}$$

which isolates within-branch variation from mean-prediction error and differences between branches. A separate scale predictor minimizes

$$\mathcal{L}_{\text{scale}} = \mathbb{E}_{(c,z)\sim\text{Unif}(\mathcal{G}_{\text{train}})} \left[\log s_\psi(c, z) + \frac{\widehat{\tau}_{cz}^2}{2 s_\psi^2(c, z)} \right]. \tag{9}$$

We use $s_\psi(c, z) = \max\{s_{\min}, \exp(r_\psi(c, z))\}$ in fitting and sampling, with a positive numerical floor $s_{\min}$. This objective fits the scale to observed variation rather than minimizing CFM loss over source scales; Appendix **??** gives its scalar optimum and boundary cases.

4.4 Flow Training with Frozen Sources

We first fit the mean and gate using

$$\mathcal{L}_{\text{source-structure}} = \mathcal{L}_{\text{center}} + \lambda_q \mathcal{L}_{\text{gate}}, \qquad \lambda_q > 0, \tag{10}$$

then fit the scale predictor. All source parameters and shared source features are subsequently frozen, and only the velocity network is trained. This preserves the fitted variation, avoiding the population CFM objective's bias toward vanishing source scale for non-point-mass targets (Appendix **??**).

For each training future (c, Y, z), retain its cluster-assigned pseudo-label and independently draw $t \sim \mathcal{U}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, I_D)$. Construct

$$X_0 = \mu_\eta(c, z) + s_\psi(c, z)\varepsilon, \qquad X_t = (1 - t)X_0 + tY, \qquad U = Y - X_0, \tag{11}$$

and train one velocity network shared across branches:

$$\mathcal{L}_{\text{branch-CFM}}(\theta) = \mathbb{E}_{(c,Y,z)\sim\mathcal{D}_{\text{train}},\, t,\, \varepsilon} \left[\frac{\|v_\theta(X_t, t, c, z) - (Y - X_0)\|_2^2}{D} \right]. \tag{12}$$

The endpoints are independent given (c, z), but generally dependent given c after marginalizing the shared branch. Thus Section 3 applies within each branch, up to the constant normalization D^{-1}, and Appendix **??** applies under its stated assumptions. Flow updates sample future examples, in contrast to the condition and group averages used for source fitting; they do not pair a newly sampled branch with an unrelated training future. Algorithm **??** in Appendix **??** summarizes training.

4.5 Sampling Future Videos

Given a new condition c, sample an outcome and its source state:

$$z \sim \text{Categorical}(q_\phi(\cdot \mid c)), \qquad Z_0 = \mu_\eta(c, z) + s_\psi(c, z)\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I_D). \tag{13}$$

Hold c, z fixed throughout integration,

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = v_\theta(Z_t, t, c, z), \qquad t : 0 \longrightarrow 1, \qquad \widehat{Y} = Z_1, \tag{14}$$

and decode $\widehat{Y}$ with Dec, supplying the initial observation if required. Independent repetitions capture both outcome uncertainty and within-branch variation. Generation uses predicted source parameters and a sampled branch, without future videos or their labels. Dirac targets retain the terminal-limit interpretation in Section 3. Algorithm **??** in Appendix **??** gives the sampling procedure.

5 VPHYSBENCH: EVALUATING PHYSICALLY GROUNDED VIDEO GENERATION

From plausibility to physical grounding. Visually convincing motion does not necessarily reflect the physical conditions that produced it. We distinguish three levels of video generation: *visual plausibility*, which concerns perceptual quality; *physical plausibility*, which concerns qualitative consistency with physical intuition; and *physical grounding*, which further requires the generated dynamics to respond correctly to specified numerical physical conditions. For example, a collision may appear physically plausible while producing post-collision motion inconsistent with the supplied masses and initial velocities.

VPhysBench. We introduce VPhysBench, a benchmark for evaluating physically grounded video generation across ten classical physical scenarios, including collisions, free fall, projectile motion, pendulums, oscillation, rebound, and inclined-plane motion. Each example provides an initial frame I_0, physical parameters $\mathbf{p}$, and a reference video V. Given $(I_0, \mathbf{p})$, the model generates a continuation $\widehat{V}$. A key design principle is to include visually similar initial states with different physical parameters, so that generic scene-appropriate motion is insufficient: successful generation must reflect the specified conditions in the resulting dynamics. Dataset construction, parameterization, and evaluation protocols are detailed in Appendix **??**.

Cumulative Soft Tube IoU. To measure whether generated motion follows the reference dynamics, we introduce **Cumulative Soft Tube IoU (CSTI)**, a reference-conditioned spatiotemporal trajectory metric. Unlike spatial IoU, CSTI preserves temporal information; unlike hard tube overlap, it provides graded tolerance to small spatial and temporal offsets. It further evaluates cumulative trajectory prefixes, giving persistent influence to early deviations rather than allowing them to be diluted by later agreement. Formally,

$$\mathrm{CSTI}(R, P) = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{SoftIoU}\Big(R^{(q)}, P^{(q)}\Big), \qquad Q = \{0.25, 0.50, 0.75, 1.00\}, \tag{15}$$

where $R^{(q)}$ and $P^{(q)}$ are reference and predicted object-mask tubes truncated to the first q fraction of the sequence. The soft spatiotemporal matching function, tolerance choices, and implementation details are given in Appendix **??**. Together, VPhysBench and CSTI evaluate not merely whether generated motion looks plausible, but whether it quantitatively realizes the supplied physical conditions.

6 EXPERIMENTS

We evaluate whether an outcome-conditioned source improves physics-conditioned video prediction under a fixed training budget. External comparisons measure distributional accuracy and learning efficiency, while internal ablations examine outcome conditioning, source centers, local scales, and the training schedule.

6.1 EXPERIMENTAL SETUP

Data. We use VPhysBench (Section 5), which contains approximately 5,000 examples across ten physical scenarios, with 200–1,000 cases per scenario spanning diverse conditions. Each condition is collected at least ten times to capture repeated physical realizations. Videos are trimmed and processed to 832×480 pixels and 121 frames. Data within each scenario are divided into training and test sets, and predictions are evaluated against the corresponding test references.

Models and training. All experiments use a 70M-parameter DiT and the frozen VAE from Wan2.2-TI2V-5B. We evaluate PhyParam-style and PhyCo-style conditioning mechanisms (Li et al., 2026; Narayanan et al., 2026) on this common backbone, sharing the data and latent representation across source strategies. Each run uses a training budget of 1,000 optimization steps. We report means and sample standard deviations over three training seeds; these SDs quantify training variability. Sampling uses Heun with 16 steps (32 NFE) and 32 generated futures per condition.

Evaluation. We compute distributional scores within each condition and average them equally across conditions. Energy score (ES) uses latent Euclidean distances normalized by $\sqrt{D}$ and an

off-diagonal predictive self-distance estimate. We also report unbiased multikernel MMD^2 with training-derived bandwidths and sliced W_2 (SW_2) over 64 fixed directions. Tables report test scores; the learning curves use validation ES with the display adjustments specified in Figure 3. CSTI and SSIM measure physical and visual fidelity, respectively. The test split was inspected during method development, so these results should be regarded as exploratory.

6.2 External Comparisons

We compare with standard Gaussian sources at the base and twice the base learning rate, CPD (Issachar et al., 2025), and CSFM (Kim et al., 2026), adapting each source strategy to the common video backbone. CPD uses a fitted, frozen conditional mean and $\sigma = 0.7$, selected from $\{0.1, 0.3, 0.7\}$ by validation ES. CSFM jointly learns a diagonal Gaussian source and the flow. We report both its v2 equation-based objective (variance-KL and sample-cosine weights 5 and 1) and a sensitivity variant using the public code's normalized-mean L2 alignment. Our method additionally uses training outcome labels; the corresponding controls are examined in Section 6.4.

Table 1: **External comparison on VPhysBench after 1,000 training steps.** Values are test-set mean±SD over three seeds.

Source strategy	ES↓	MMD^2 ↓	SW_2 ↓	CSTI↑	SSIM↑
PhyParam-style conditioning					
Standard Gaussian	0.1548±0.0017	0.1850±0.0021	0.1946±0.0011	0.1514	0.9423
Standard, 2×LR	0.1488±0.0055	0.1776±0.0064	0.1739±0.0106	0.1423	0.9516
CPD	0.1490±0.0021	0.1782±0.0027	0.1729±0.0051	0.0962	0.9591
CSFM	0.1592±0.0079	0.1905±0.0093	0.2016±0.0150	0.1736	0.9611
Ours	**0.1273±0.0012**	**0.1162±0.0049**	**0.1302±0.0032**	**0.2912**	**0.9773**
PhyCo-style conditioning					
Standard Gaussian	0.1493±0.0014	0.1788±0.0017	0.1833±0.0035	0.1251	0.9421
Standard, 2×LR	0.1481±0.0056	0.1765±0.0069	0.1732±0.0128	0.0914	0.9500
CPD	0.1540±0.0021	0.1833±0.0019	0.1872±0.0074	0.1262	0.9677
CSFM	0.1705±0.0215	0.2039±0.0254	0.2102±0.0383	0.1935	0.9651
Ours	**0.1270±0.0012**	**0.1241±0.0035**	**0.1198±0.0025**	**0.2841**	**0.9754**

Table 1 shows that our method achieves the lowest mean ES, MMD^2, and SW_2, together with the highest CSTI and SSIM, in both conditioning configurations. These gains across distributional, physical, and visual metrics demonstrate more effective learning under the same 1,000-step budget.

6.3 Learning Efficiency

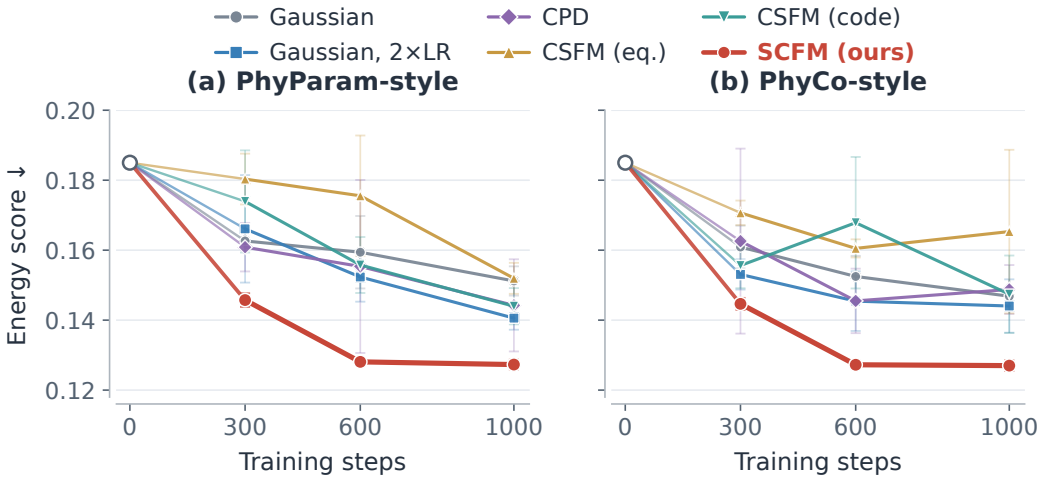


Figure 3: **Learning progress on VPhysBench.** Baseline points report validation ES at 300, 600, and 1,000 steps. All methods use 32 NFE.

Figure 3 summarizes learning progress at 300, 600, and 1,000 steps, with approximate initial values added for context. The curves illustrate progress across checkpoints, with the data corrections detailed in the caption. The fixed-budget results in Table 1 establish the gains across evaluation metrics.

Flow refinement contributes beyond source construction: on paired test samples, ES decreases from 0.1408 for source-only sampling to 0.1273/0.1270 after the flow, a further 9.57%/9.74% reduction. Directly returning the sampled branch center gives 0.1464. Thus the learned source provides an informative starting distribution, while transport further improves the prediction.

6.4 Internal Ablations

Table 2 examines branch conditioning, source construction, and scale freezing under the same training budget.

Table 2: **Internal ablations on VPhysBench.** Test ES↓ after 1,000 training steps, mean±SD over three seeds. The scale-unfreezing variants keep the branch means and gate fixed.

Configuration	PhyParam-style	PhyCo-style
Standard Gaussian (2×LR)	0.1488 ± 0.0055	0.1481 ± 0.0056
Gaussian + branch conditioning	0.1485 ± 0.0015	0.1453 ± 0.0021
Branch means + unit scale	0.1495 ± 0.0016	0.1451 ± 0.0011
Single Gaussian (no branches)	0.1825 ± 0.0170	0.1767 ± 0.0068
Unfreeze scales (no regularization)	0.1304 ± 0.0014	0.1317 ± 0.0024
Unfreeze scales + data-scale regularization	0.1268 ± 0.0012	0.1279 ± 0.0019
Ours (full model)	$\mathbf{0.1193 \pm 0.0012}$	$\mathbf{0.1170 \pm 0.0012}$

Branch information and separation. *Standard Gaussian (2×LR)* is the stronger unstructured-source reference. *Gaussian + branch conditioning* retains standard Gaussian noise while supplying our training branch labels and predicted inference gate. Its higher ES than the full model shows that access to branch information alone does not explain SCFM's gain. *Single Gaussian (no branches)* replaces the branch mixture with one conditional Gaussian, pooling distinct outcomes into a common mean and scale.

Branch means and scales. *Branch means + unit scale* uses the same predicted branch means and gate as SCFM but fixes the source standard deviation to one. Relative to Gaussian noise with branch conditioning, centering alone does not consistently improve ES. The full model replaces this unit noise with scales fitted to within-branch variation and achieves lower ES in both configurations. This comparison supports adapting the noise around each predicted branch center; it does not establish an independent benefit from mean prediction alone.

Freezing source scales. *Unfreeze scales (no regularization)* updates the source scales jointly with the flow while fixing the means and gate. *Unfreeze scales + data-scale regularization* adds a penalty anchoring scales to the fitted data variation. It improves ES over unregularized joint learning, but both variants have higher ES than freezing in this comparison. These results support fitting scales from repeated observations and then freezing them to preserve the intended source variation.

7 Conclusion

We studied source design for physics-conditioned video prediction when one observation admits several distinct, internally concentrated outcomes. Our method predicts outcome probabilities, branch centers, and local scales, then trains a shared flow with the source held fixed. The source-collapse analysis explains why the population CFM objective alone does not identify a useful noise scale. We also introduce VPhysBench for evaluating generation under explicit physical conditions. On its evaluated split, the complete pipeline achieves lower distributional error than standard Gaussian, CPD, and CSFM adaptations under the same training budget. Internal controls associate this gain with representing both outcome separation and within-branch variation, while source-only comparisons show an additional contribution from flow refinement. The evidence supports more efficient distributional learning in this setting; extending the result beyond supervised branches and establishing visual and physical fidelity remain important next steps.

AI USE STATEMENT

Generative AI tools were used to assist with research brainstorming, language polishing, and code development during the preparation of this work. All AI-assisted content, including methodological suggestions, code, and manuscript text, was reviewed and verified by the authors. The authors take full responsibility for the scientific claims, experimental results, and final content of this paper.

REPRODUCIBILITY STATEMENT

We provide implementation and experimental details in the main paper and appendix, including model architectures, training settings, datasets, and evaluation protocols. Code and additional materials necessary to reproduce the experiments will be released upon publication.

REFERENCES

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation, 2025. URL https://arxiv.org/abs/2503.06800.

Chen Chen, Pengsheng Guo, Liangchen Song, Jiasen Lu, Rui Qian, Xinze Wang, Tsu-Jui Fu, Wei Liu, Yinfei Yang, and Alex Schwing. CAR-Flow: Condition-Aware Reparameterization Aligns Source and Target for Better Flow Matching, 2025. URL https://arxiv.org/abs/2509.19300.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force Prompting: Video Generation Models Can Learn and Generalize Physics-based Control Signals, 2025. URL https://arxiv.org/abs/2505.19386.

Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. Designing a Conditional Prior Distribution for Flow-Based Generative Models, 2025. URL https://arxiv.org/abs/2502.09611.

Rasul Khanbayov, Mohamed Rayan Barhdadi, Erchin Serpedin, and Hasan Kurban. IRIS: A Real-World Benchmark for Inverse Recovery and Identification of Physical Dynamic Systems from Monocular Video, 2026. URL https://arxiv.org/abs/2603.16432.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better Source, Better Flow: Learning Condition-Dependent Source Distribution for Flow Matching, 2026. URL https://arxiv.org/abs/2602.05951.

Yanxun Li, Hao Wen, Bingze Song, Jiashu Zhu, Aiming Hao, Chubin Chen, Jintao Chen, Jiahong Wu, Xiangxiang Chu, and Miao Wang. Learning Explicit Physical Parameter Control and Benchmarking for Video Generation, 2026. URL https://arxiv.org/abs/2607.18924.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, 2022. URL https://arxiv.org/abs/2210.02747.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles?, 2025. URL https://arxiv.org/abs/2501.09038.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. PhyCo: Learning Controllable Physical Priors for Generative Motion, 2026. URL https://arxiv.org/abs/2604.28169.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport, 2023. URL https://arxiv.org/abs/2302.00482.