Pretrained Video Models as Differentiable Physics Simulators for Urban Wind Flows

Janne Perini*[1], Rafael Bischof*[1], Moab Arar[2], Ayça Duran[1], Michael A. Kraus[3], Siddhartha Mishra[1], and Bernd Bickel[1]

[1] ETH Zürich, Switzerland
[2] Tel Aviv University, Israel
[3] TU Darmstadt, Germany

Abstract. Designing urban spaces that provide pedestrian wind comfort and safety requires time-resolved Computational Fluid Dynamics (CFD) simulations, but their current computational cost makes extensive design exploration impractical. We introduce WinDiNet (Wind Diffusion Network), a pretrained video diffusion model that is repurposed as a fast, differentiable surrogate for this task. Starting from LTX-Video, a 2B-parameter latent video transformer, we fine-tune on a dataset of 13,000 2D incompressible CFD simulations over procedurally generated building layouts. A systematic study of training regimes, conditioning mechanisms, and VAE adaptation strategies, including a physics-informed decoder loss, identifies a configuration that outperforms purpose-built neural PDE solvers. The resulting model generates full 112-frame rollouts in under a second. As the surrogate is end-to-end differentiable, it doubles as a physics simulator for gradient-based inverse optimization: given an urban footprint layout, we optimize building positions directly through backpropagation to improve wind safety as well as pedestrian wind comfort. Experiments on single- and multi-inlet layouts show that the optimizer discovers effective layouts even under challenging multi-objective configurations, with all improvements confirmed by ground-truth CFD simulations.

Keywords: Urban wind simulation · Video diffusion models · Neural surrogate · Computational fluid dynamics

1 Introduction

When urban designers and engineers plan modern, livable public spaces, wind is a key constraint. It affects pedestrian comfort, structural loads on buildings, and natural ventilation. Buildings that channel airflow can produce gusts that endanger pedestrians and stress building facades, while overly sheltered configurations create stagnant zones where heat and pollutants accumulate. In cities increasingly affected by air pollution and rising temperatures, these aspects carry real consequences for structural safety, public health, and quality of life. Exploring

* Equal contribution.

urban layouts that navigate this trade-off requires fast, accurate flow predictions and, ideally, guidance on how to update a given design to improve wind comfort and safety simultaneously. Computational fluid dynamics (CFD) simulations, the traditional method for this task, are computationally expensive and require tedious modeling and domain expertise in addition to taking minutes to hours at the resolution needed for a reliable comfort evaluation [31]. Furthermore, they are typically non-differentiable, meaning they offer no signal about which geometric changes would improve a given design.

Deep-learning surrogates mitigate these limitations by offering fast inference and differentiable predictions. While existing approaches based on convolutional and graph neural networks can predict mean velocity or wind-factor fields [27, 32], temporally averaged outputs cannot evaluate comfort criteria defined as exceedance probabilities over threshold speeds [21], nor capture transient gusts relevant to pedestrian safety. Temporal extensions based on Fourier neural operators and autoregressive transformers [3, 38] can in principle produce time-resolved predictions, but they struggle to maintain accuracy over long rollouts when trained from scratch on limited domain-specific data.

Meanwhile, video diffusion models have advanced rapidly in capabilities relevant to wind flow prediction. Recent latent diffusion transformers generate high-resolution, temporally coherent sequences with multi-scale motion and long-range spatial dependencies [1, 12, 51]. These models already encode physical priors, such as gravity, collisions, and fluid-like motion, that can be adapted to generate visually plausible physical phenomena [9, 52, 54, 58]. Predicting urban wind flows fits a similar framing: projected onto a horizontal plane, a velocity field evolves as a frame sequence whose channels encode physical quantities rather than pixel colors. We show that fine-tuning a pretrained video model on synthetic urban CFD data yields a differentiable surrogate accurate enough for both forward prediction and gradient-based inverse optimization of building layouts for pedestrian wind comfort, and that the physical priors transfer to quantitatively accurate simulations, not just visually plausible ones.

In this work, we fine-tune LTX-Video [12], a transformer-based latent video diffusion model, on a dataset of 13,000 2D CFD simulations over procedurally generated urban footprint layouts and systematically study the design choices needed to obtain physically accurate predictions from a pretrained video backbone (Figure 1). We compare (i) the training regime, such as Low-Rank Adaptation (LoRA) [15] *vs.* full fine-tuning, (ii) the conditioning mechanism (text prompts *vs.* scalar-conditioned latent modulation with the text encoder removed), and (iii) the Variational Autoencoder (VAE) adaptation strategy, including color adapters, decoder fine-tuning, and physics-informed objectives. Our main contributions are:

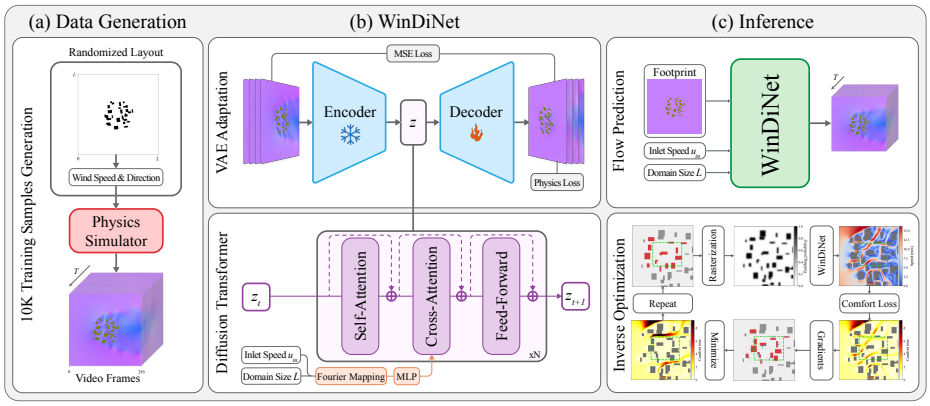


Fig. 1: Overview of the proposed framework. **(a)** Procedurally generated urban layouts are simulated with a 2D incompressible Euler solver to produce training data. **(b)** A latent diffusion model with a physics-informed VAE is trained to generate wind field sequences conditioned on building footprint, inlet speed u_{in}, and domain size L. **(c)** At inference, the model generates horizontal and vertical velocity fields (u, v) and enables gradient-based inverse optimization of building layouts.

- We generate a dataset of 13,000 2D incompressible wind flow simulations over procedurally generated urban layouts, spanning diverse building configurations, inlet speeds, and domain sizes.
- We present a systematic study of adapting a pretrained video diffusion model for physics simulation, comparing training regimes, conditioning mechanisms, and VAE adaptation strategies, including physics-informed decoder losses.
- We demonstrate that the resulting differentiable surrogate enables gradient-based inverse optimization of building layouts for pedestrian wind comfort.

The code and model weights can be obtained via github.com/rbischof/windinet, and the dataset is shared upon request.

2 Related Work

Pedestrian wind comfort depends on transient flow phenomena (vortex shedding, recirculation zones, shear-layer instabilities) that are computationally expensive to resolve numerically [31]. A large body of work has therefore pursued data-driven surrogates trained on CFD datasets. Convolutional neural networks (CNNs) and CNN–generative adversarial network (GAN) hybrids map building footprints to mean velocity or wind-factor fields [5, 6, 19, 30, 32, 43]. Graph neural networks operate on unstructured meshes at city scale [10, 27], including physics-informed variants that embed Reynolds-Averaged Navier–Stokes (RANS) residuals in the loss [42]. Physics-informed neural networks have also been applied to reconstruct 3D wind fields around buildings from sparse measurements [40], demonstrating that embedding governing equations in the training

loss can compensate for limited data. These approaches predict temporally averaged or steady-state conditions. However, the wind comfort criteria applied in practice [17, 21] are formulated as exceedance probabilities over threshold speeds, which require time-resolved velocity data to evaluate. Although Fourier Neural Operator (FNO) [25] variants [3,38] can predict instantaneous states and generate temporal sequences through autoregressive rollout, they struggle with generating long rollouts because errors accumulate over successive steps.

Three recent threads suggest how to close this gap. First, diffusion models have been shown to generate accurate time-dependent physical fields [7, 33]. Second, partial differential equation (PDE) solving has been recast as video generation [23]. Third, foundation models pretrained on diverse equation families [13, 34, 44, 55], enabled by large-scale PDE benchmarks [36, 45], improve sample efficiency on downstream scientific tasks. These developments motivate formulating the unsteady urban wind flow simulation as a conditional video generation problem in which temporal coherence and long-range spatial dependencies are learned from the domain-specific data.

The video generation literature provides the architectural backbone for this idea. Early models were GAN-based [41, 48, 50] and confined to short, low-resolution clips. Diffusion models removed this limitation, first by adding temporal attention to 3D U-Nets [14], then by shifting denoising into a learned latent space [1], and most recently by replacing the U-Net with transformers over patchified latent tokens [12, 51]. These efforts have produced models that synthesize temporally coherent video of complex scenes with long-range dependencies across both space and time. These models also encode physical priors that enable zero-shot reasoning about gravity, collisions, and fluid-like behavior [54]. Several recent methods exploit this observation: Force Prompting [9], DiffPhy [58], PhysCtrl [52], and PhysVideoGenerator [35] condition video diffusion on forces, physical cues, or trajectories to improve the plausibility of generated motion. The evaluation of prediction performance in these approaches, however, remains perceptual (human preference or FID-type scores), and none of them targets quantitative agreement with a governing PDE.

Beyond forward prediction, a growing body of work uses surrogates to accelerate the optimization of urban layouts for wind comfort. Wu and Quan [57] survey the field and identify three dominant strategies: evolutionary methods coupled with CFD [18], GAN-based surrogates paired with genetic algorithms [16], and response-surface models fitted to parametric CFD sweeps. All of these treat the surrogate as a black-box evaluator and rely on derivative-free optimizers that require many candidate evaluations per iteration. In contrast, an end-to-end differentiable surrogate can offer gradient-based layout optimization when combined with a soft rasterizer inspired by differentiable rendering [26] to map continuous building coordinates to occupancy masks.

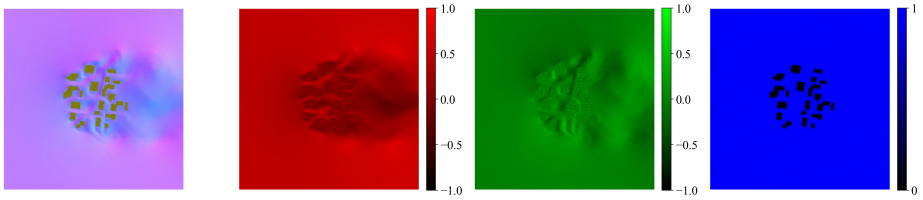


Fig. 2: Channel decomposition of a single simulation frame. From left to right: encoded RGB composite, red channel encoding horizontal velocity u, green channel encoding vertical velocity v, and blue channel encoding the fluid mask (1: fluid, 0: building). Velocity values are linearly rescaled to $[-1, 1]$ using the dataset-wide maximum speed.

3 Methodology

We consider 2D wind flow over a square urban domain of side length L (in meters). The built environment is described by a binary building footprint $B \in \{0, 1\}^{H \times W}$, where $B_{x,y} = 1$ indicates a solid structure and $B_{x,y} = 0$ indicates open space. The corresponding fluid mask is $F = 1 - B$. Without loss of generality, wind enters from the left boundary at a prescribed inlet speed u_{in} (in m/s) and exits through the right. Other wind directions are obtained by rotating the building footprint before inference and rotating the predicted field back. The flow evolves over time from the uniform inlet condition to a quasi-steady state shaped by the building geometry.

We cast wind field prediction as a conditional video generation task. Given (B, u_{in}, L), the goal is to generate the velocity field sequence $\{\mathbf{w}_t\}_{t=1}^{T}$, where each $\mathbf{w}_t = (u_t, v_t) \in \mathbb{R}^{H \times W \times 2}$ contains the horizontal and vertical velocity components at time step t.

To leverage a pretrained RGB video model, we encode the physical velocity fields as pixel values (Figure 2). Each frame maps the two velocity components (u, v) to the red and green channels of an RGB image by linearly rescaling with a dataset-wide maximum speed $u_{\max}$ so that values lie in $[-1, 1]$. The blue channel encodes the fluid mask F.

A directional conditioning image is prepended as frame t=0: the red channel is set to $u_{\text{in}}/u_{\max}$ everywhere in fluid cells and zero inside buildings, the green channel is zero, and the blue channel carries the fluid mask. This conditioning signals both the geometry and the inlet condition to the model.

While this RGB encoding is convenient for training and inference, the per-channel differences are visually subtle. For all result figures, we therefore convert to wind speed magnitude $\|\mathbf{w}\| = \sqrt{u^2 + v^2}$ and apply a `coolwarm` colormap to better highlight flow structures.

3.1 Base Model

We build upon LTX-Video [12], a transformer-based latent video diffusion model that generates all frames jointly in a single denoising pass, rather than autore-

gressively one frame at a time. Specifically, we use the 2B-parameter text-and-image-to-video (TI2V) variant, which comprises three components: (i) a causal 3D VAE that compresses video into a compact latent representation, (ii) a T5-based text encoder for conditioning, and (iii) a diffusion transformer (DiT) [37] that performs denoising in latent space by flow matching. We consider both full fine-tuning and LoRA fine-tuning for the DiT in our experiments.

3.2 VAE Adaptation

The pretrained LTX-Video VAE was designed for natural RGB video. When applied directly to our wind-field encoding, it introduces reconstruction artifacts due to the domain gap between photorealistic content and pseudo-colored velocity fields (cf. Figure 5). Modifying the encoder would shift the latent distribution and corrupt the DiT's pre-trained denoising dynamics. We therefore keep the encoder frozen and instead learn a color mapping around the frozen VAE, or fine-tune only the decoder [49]. Both strategies are applied in a dedicated stage before training the diffusion model.

Color adapter. Our mapping from physical quantities (u, v, building footprint) to RGB channels is ultimately arbitrary (cf. supplement for an ablation). Rather than hand-picking a mapping, we can let the network learn one: a shallow Multilayer Perceptron (MLP) ($3 \to 32 \to 3$ with SiLU activation and tanh output) before the encoder transforms the wind-field encoding into a new, learned color space, and an equivalent MLP after the decoder maps back to physical channels. Both adapters are trained end-to-end with a frozen VAE.

Decoder fine-tuning. A different approach is to fine-tune the decoder weights while keeping only the encoder frozen [49]. The reconstruction loss can optionally be augmented with physics-based regularizers: a *divergence penalty* enforcing incompressibility ($\nabla \cdot \mathbf{w} = 0$), a *no-penetration penalty* at building walls, and distance-weighted Mean Squared Error (MSE) that upweights fluid cells near building boundaries where velocity gradients are steepest (full formulations in the supplement). Physics-informed losses are rarely combined with latent diffusion models because they operate in pixel space, and every training step would require decoding the full output and backpropagating through the decoder. Our preliminary decoder fine-tuning removes this obstacle, since these physics-informed losses are applied in a short stage before the diffusion model is trained. Approximately 5k optimization steps suffice for a significant improvement in reconstruction quality. Table 1 summarizes the resulting VAE configurations.

3.3 Conditioning Strategies

The original LTX-Video model is conditioned on text prompts via cross-attention. While text can describe simulation parameters (e.g., *"inlet speed 18 m/s, domain size 1300 m"*), natural language is an imprecise interface for continuous physical quantities. We therefore introduce an alternative *scalar conditioning* mechanism.

Table 1: VAE adaptation variants. Color Adapter learns a nonlinear channel transformation around the frozen VAE. Dec. FT fine-tunes the decoder weights. Dec. FT Physics adds physics-informed losses during decoder fine-tuning.

Variant	Color adaptation	Decoder	Physics losses
Base	–	frozen	–
Color Adapter	✓	frozen	–
Dec. FT	–	fine-tuned	–
Dec. FT Physics	–	fine-tuned	✓

A learnable embedding module maps u_{in} and L directly into the transformer's conditioning space, bypassing the text encoder entirely. Each scalar is normalized to $[0, 1]$, encoded via Fourier features [46] with log-spaced frequencies, and projected by a small MLP into embedding tokens that replace the text encoder output in the transformer's cross-attention.

4 Experiments

4.1 Dataset

We generate a dataset of 2D urban wind simulations using an incompressible Euler solver and procedural building footprint generator [8]. The incompressibility assumption is standard for pedestrian comfort and building aerodynamics studies, where typical wind speeds keep Mach numbers well below compressible regimes [22]. Building footprints consist of randomly placed rectangular blocks (10 to 50 m, min. 10 m alley width) within a circular city region whose diameter is sampled from $\{300, 400, \ldots, 800\}$ m, with block counts scaling proportionally with area. A 300 m buffer surrounds each city to allow flow development and reduce boundary effects. Together, the city region and buffer zone thus form the full domain with side length L in $[900, 1400]$ m. The inlet wind speeds u_{in} are sampled from $[0.1, 20]$ m/s, covering the range of conditions relevant to pedestrian comfort and safety assessments [2, 4]. Finally, inlet wind direction is sampled uniformly from $[0, 360]^\circ$, and each simulation is subsequently rotated so that wind flows left-to-right. This canonicalization removes wind direction as a degree of freedom, leaving only u_{in}, domain size L, and the domain geometry as varying parameters. Each simulation covers 112 s of physical time and is stored as T=112 velocity snapshots at 256×256 resolution plus an initial conditioning frame (Sec. 3). The dataset comprises 10,000 training, 1,000 validation, and 2,000 test simulations. Randomly selected samples are shown in Figure 3.

4.2 Training

Training uses AdamW [29] (lr=10^{-4}, cosine schedule), batch size 64 for 2,000 steps. The model variants trained with full fine-tuning used AdamW (lr=10^{-5}, cosine schedule) and batch size 64 for 10,000 steps (156 steps/epoch, ∼64 epochs)

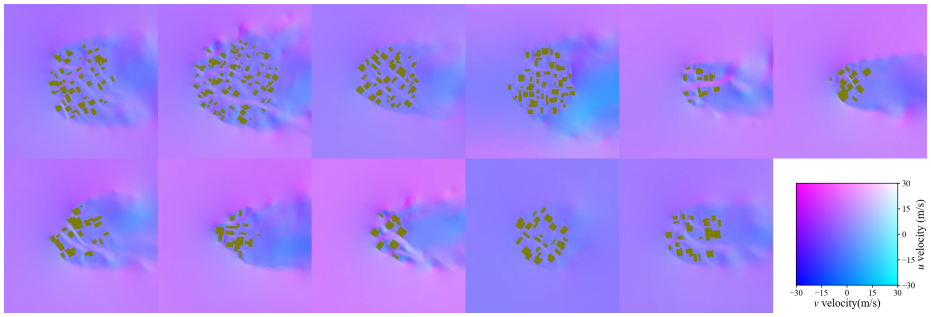


Fig. 3: Representative samples from the training dataset. Each tile shows the RGB-encoded velocity field at frame 100 for a distinct simulation.

for DiT transformer training. The variants with VAE decoder fine-tuning trained the decoder separately prior to DiT fine-tuning using the same configurations.

We evaluate generated wind fields against the ground truth using five metrics, computed on fluid pixels only: *MAE* (m/s), *MRE* (%), *VRMSE* (variance-normalized RMSE; primary ranking metric), *spectral divergence* (temporal frequency fidelity), and *Wasserstein-1 distance* W_1 (speed distribution accuracy). The full metric definitions are provided in the supplement.

4.3 Main Results

Table 2: Results of baseline models and all WinDiNet variants on the test set. All metrics computed on fluid pixels only.

	Model	Cond.	VRMSE ↓	MAE ↓	MRE ↓	Spectral ↓	W_1 ↓
Baselines	U-Net [39]	Scalar	1.467	4.73	15.46	2.74	7.06
	Poseidon [13]		0.781	2.17	7.05	2.40	2.79
	AFNO [11]		0.618	1.44	4.61	1.71	1.31
	FNO [25]		0.610	1.42	4.53	1.68	1.32
	OFormer [24]		0.585	1.36	4.34	**1.52**	1.22
	RNO [28]		0.563	1.19	3.91	1.72	1.14
WinDiNet (ours)	LoRA	Text	0.746	1.56	5.20	1.84	1.29
	Base		0.702	1.36	4.53	1.74	1.08
	Base	Scalar	0.616	1.12	3.75	1.61	0.95
	Color Adapter		0.577	1.11	3.72	1.63	0.91
	Dec. FT		0.531	1.02	3.37	1.55	**0.83**
	Dec. FT Physics		**0.520**	**1.01**	**3.33**	1.54	**0.83**

Table 2 compares WinDiNet against six neural operator baselines on the full inference pipeline, where the diffusion model generates latent sequences that

are decoded into velocity fields. The baselines fall into three performance tiers: autoregressive models (OFormer, RNO) perform best, followed by one-shot predictors (AFNO, FNO), with frame-to-frame models that must be rolled out autoregressively at inference (U-Net, Poseidon) trailing behind. RNO achieves the lowest baseline VRMSE at 0.563.

Full fine-tuning of the DiT transformer reduces VRMSE by 5.9% over LoRA (rank 512); therefore, all subsequent variants use full fine-tuning. Replacing text conditioning with scalar embeddings yields a further 12% reduction, bringing the scalar base model (VRMSE = 0.616) into the range of the stronger baselines despite operating through a lossy VAE bottleneck that none of the baselines require. Text conditioning tokenizes continuous physical quantities as strings, a representation that the pretrained text encoder was not designed to handle. The scalar embedding injects simulation parameters directly into the transformer's conditioning pathway, which improves both accuracy and efficiency.

With VAE adaptation, the color adapter brings WinDiNet close to the best baselines, and decoder fine-tuning pushes it beyond them. Dec. FT Physics (Fig. 4) outperforms the best baseline (RNO) by 7.6% in VRMSE and 15% in MAE. These are the lowest scores across all metrics except spectral divergence, where OFormer retains a marginal lead (1.52 vs. 1.54).

To isolate the contribution of the pretrained video prior, we retrain the model under the same compute budget while removing pretraining. Training the DiT from scratch on top of the fine-tuned VAE raises VRMSE from 0.52 to 0.57, and training both the DiT and the VAE from scratch raises it further to 0.92. Pretraining therefore accounts for a large share of the model's accuracy, consistent with what has been reported for physical reasoning [54].

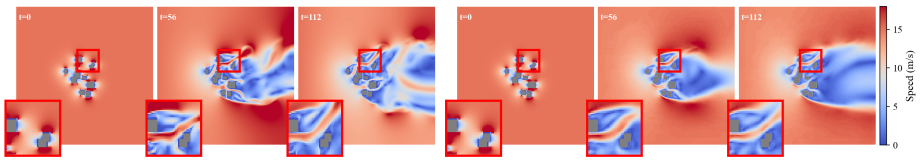
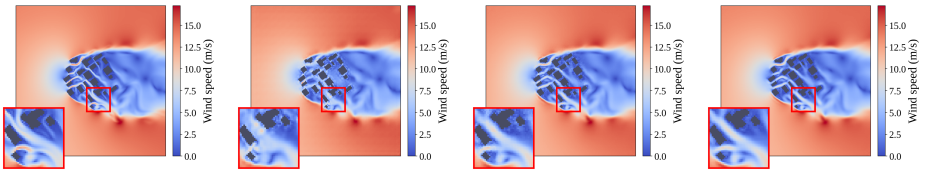


Fig. 4: Wind speed magnitude predicted by Dec. FT Physics for a procedurally generated urban layout at 15 m/s inlet velocity. Ground truth (left) and model prediction (right) at timesteps $t = 0$, 56, and 112.

4.4 VAE Adaptation

VAE adaptation is performed as a separate stage before diffusion model training. To evaluate its effect in isolation, we encode ground-truth velocity fields into the latent space and decode them back without involving the diffusion model (Table 3). Every adaptation method improves reconstruction fidelity over the frozen baseline, with the best decoder variant reducing VRMSE by over 62%.



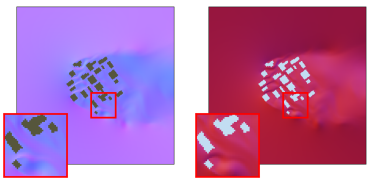
(a) Ground truth **(b)** Base **(c)** Dec. FT **(d)** Dec. FT Physics

Fig. 5: VAE reconstruction quality at t=90 for a sample from the test set.

Table 3: VAE reconstruction quality (encode → decode) on the test set. Adapted variants trained for 5,000 steps. Base uses the pretrained VAE without modification.

Config	VRMSE↓	MAE↓	MRE↓	Spectral↓	W_1 ↓
Base	0.188	1.63×10^{-2}	8.9%	0.781	0.0157
Color Adapter	0.157	1.60×10^{-2}	8.4%	0.704	0.0167
Dec. FT	0.084	5.97×10^{-3}	3.2%	0.461	0.0054
Dec. FT Physics	**0.070**	$\mathbf{4.95\times10^{-3}}$	**2.6%**	**0.396**	**0.0045**

The color adapter learns a nonlinear channel mapping that transforms the velocity field into a visually distinct palette (Fig. 6), primarily increasing contrast between buildings and fluid regions rather than sharpening fine-scale flow structures such as vortices. This may explain why the color adapter slightly degrades speed distribution fidelity, despite improving spatial reconstruction metrics (cf. Table 3).



(a) Input **(b)** Color adapter

Fig. 6: Learned channel transformation by the color adapter.

Given the color adapter's limited impact on distributional fidelity, we instead fine-tune the decoder. This proves substantially more effective, reducing the reconstruction VRMSE by more than 55% relative to the base configuration and making the color adapter redundant. Fine-tuning allows the decoder to learn the domain-shifted distribution directly, rather than requiring latents to conform to pretrained expectations. As shown in Fig. 5, this produces significantly sharper vortex boundaries. Adding a physics-informed loss provides a complementary objective that penalizes aspects of the flow that pixel-level losses do not capture, such as near-wall gradients. These physics-informed gains are primarily reflected in the quantitative metrics rather than in qualitative reconstructions.

The ranking of VAE adaptation strategies in the isolated reconstruction experiment (Table 3) transfers directly to the full inference pipeline (Table 2), where each decoder variant produces velocity fields from the same diffusion-generated latents. VAE reconstruction quality is a reliable predictor of simulation accuracy: decoder-level improvements propagate through the generative

process and are not degraded by stochastic sampling. Since the underlying diffusion model weights are shared across all VAE variants, the gap between the Scalar Base and Scalar Dec. FT Physics (15.6% reduction in VRMSE) can be attributed to decoder adaptation.

4.5 Inference Configuration

We select the number of denoising steps and the classifier-free guidance (CFG) scale via grid search on the validation set (Fig. 7). The best configuration uses guidance scale 1.0 and 2 steps. At 2 denoising steps on a single NVIDIA H200 GPU, the model generates a full $(T+1)$-frame velocity field sequence at 256×256 resolution in approximately 0.32 s in FP16, roughly three orders of magnitude faster than the incompressible Euler solver used to generate the training data. At this setting inference fits within 24 GB of VRAM.

Two factors explain why so few steps are sufficient. First, LTX-Video uses rectified flow [12], which typically requires fewer integration steps than conventional DDPM schedulers. Second, 2D urban wind fields are structurally simpler than natural video: there are no occlusions, no texture variety, and the dynamics are governed by a single PDE. Regarding CFG, natural video generation typically uses CFG scales of 3 or higher [12]. We found that higher CFG values produced visually sharper vortices and building boundaries, but did not improve quantitative metrics. Because our conditioning signal encodes physical quantities rather than semantic descriptions, scaling its contribution distorts the learned mapping from simulation parameters to velocity fields. We conclude that CFG should be disabled in conditioned diffusion models for physics simulation.

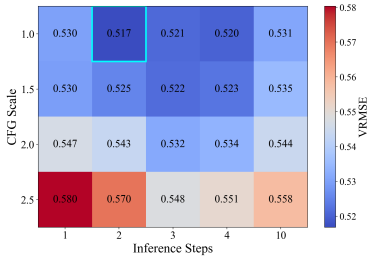
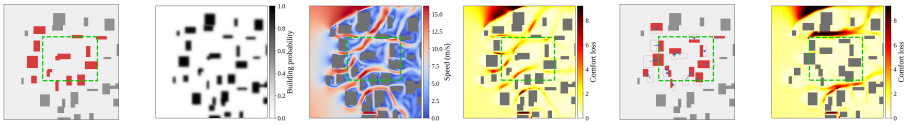


Fig. 7: Grid search over CFG scale and number of denoising steps on the validation set (Scalar conditioning, Dec. FT Physics).

Additional ablations on temporal extrapolation, domain size variation, inlet speed generalization, and RGB channel assignment are provided in the supplement.

5 Inverse Optimization of Building Layouts

Encouraged by the surrogate's accuracy and its ability to produce full velocity fields in under a second, we investigate whether it can serve as a differentiable physics simulator for inverse optimization of urban building layouts. Given an existing urban layout, we optimize building positions to minimize wind comfort violations in a target region, replacing high-fidelity CFD with the frozen surrogate during gradient computation. The coordinates of building footprint centers pass through a differentiable rasterizer into the frozen surrogate, which predicts the wind field for the current building layout.



(a) (b) (c) (d) (e) (f)

Fig. 8: Inverse optimization pipeline, zoomed in for visibility. (a) Movable buildings (red) and objective region (green rectangle) are defined. (b) A differentiable rasterizer produces a soft (blurred) occupancy mask. (c) WinDiNet predicts the wind field. (d) Comfort loss penalizes speeds outside the target band. (e) Gradient descent updates building positions. (f) Optimized layout concentrates speeds within the comfort band.

A composite loss then penalizes wind speeds outside the desired comfort range (Fig. 8). The loss takes the form

$$\mathcal{L} = \lambda_{\mathrm{d}}\, e_{\mathrm{danger}} + \lambda_{\mathrm{c}}\, e_{\mathrm{comfort}} + \lambda_{\mathrm{s}}\, e_{\mathrm{stag}}, \tag{1}$$

where e_{danger}, e_{comfort}, and e_{stag} measure the fraction of wind speeds above 15 m/s, above 5 m/s, and below 1 m/s, respectively, in the objective region. The danger term receives ten times the weight of the others ($\lambda_{\mathrm{d}} = 10$, $\lambda_{\mathrm{c}} = \lambda_{\mathrm{s}} = 1$).

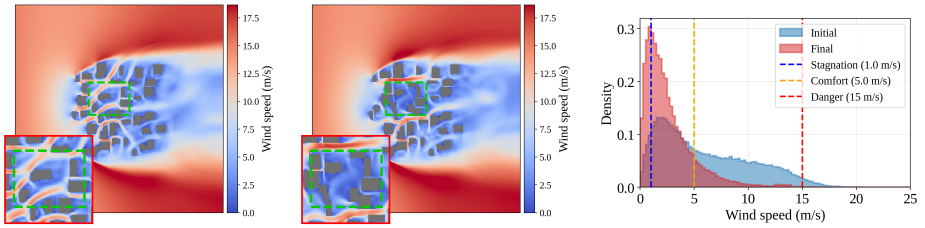
We explore two design modes. In *rigid* mode, each building translates as a unit with a fixed footprint geometry. However, relocating an entire building may not always be necessary. Often, small modifications to a building's footprint geometry can satisfy wind comfort requirements. To emulate this, *morph* mode subdivides each building into 2×2 sub-blocks that move independently, subject to a cohesion loss that prevents them from drifting apart and breaking the building into disconnected fragments. Optimization runs for 200 Adam [20] steps in both modes. Full details of the optimization, differentiable rasterizer, and regularization terms are provided in the supplement.

Table 4: Pedestrian-level wind speed distribution (%) before and after single-inlet layout optimization in rigid and morph mode.

	< 1 m/s	1–5 m/s	5–15 m/s	> 15 m/s
Initial	6.8	43.6	47.1	2.6
Rigid	23.7	63.5	12.6	0.2
Morph	19.3	62.0	18.3	0.4

Figure 9 shows results for a single inlet boundary condition (left-to-right wind at 15 m/s) in *rigid* mode. The buildings translate to form a windbreak upstream of the objective region, deflecting the incoming flow and reducing through flow. Table 4 quantifies the improvement: dangerous wind speeds above 15 m/s drop from 2.6% to 0.2%, and the fraction above 5 m/s falls from 49.7% to 12.8%.

The *morph* mode exploits its additional degrees of freedom to reshape building geometry rather than simply translating it (Fig. 10). The resulting layout

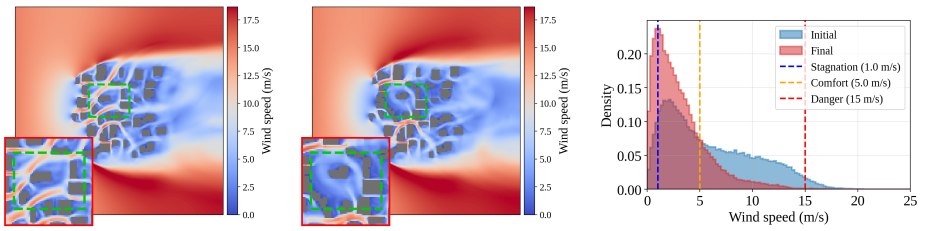


(a) Initial layout **(b)** Optimized layout **(c)** Speed distribution

Fig. 9: Single-inlet *rigid* optimization (15 m/s, left to right). Buildings translate to shelter the objective region (green rectangle), targeting wind speeds within the 1–5 m/s comfort band.

produces a different trade-off than *rigid* mode. The fraction of speeds above 5 m/s is somewhat higher (18.7% vs. 12.8%), but the stagnation fraction is substantially lower (19.3% vs. 23.7%). The *morph* optimizer appears to leave controlled openings in the shelter that maintain airflow through the objective region, channeling wind into a circular motion rather than blocking it entirely.

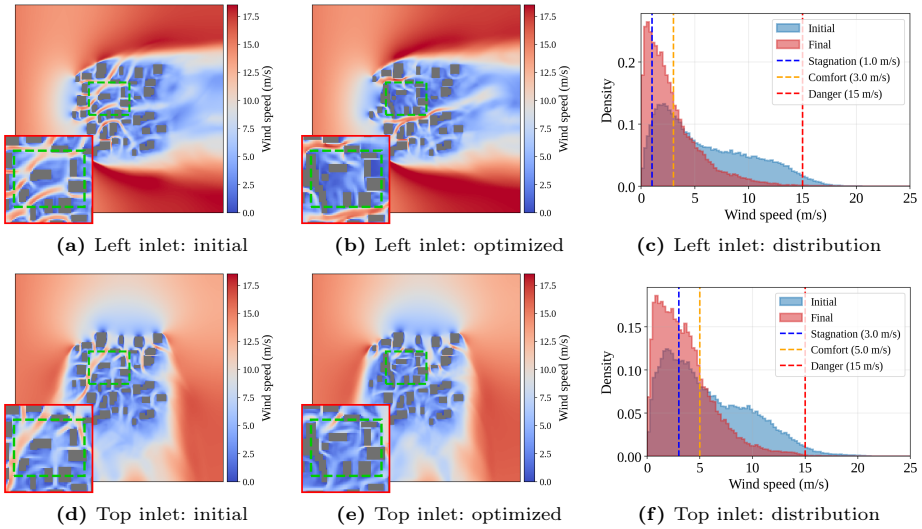
Both modes trade dangerous and uncomfortable winds for stagnation, as the density plots in Figs. 9c and 10c show. Most of the probability mass sits just above the 1 m/s threshold. This is inherent to the composite loss, since reducing high speeds inevitably concentrates mass at the low end. If stagnation is a primary concern for a given site, designers can increase λ_s to penalize it more aggressively.



(a) Initial layout **(b)** Optimized layout **(c)** Speed distribution

Fig. 10: Single-inlet *morph* optimization (15 m/s, left to right). Buildings are subdivided into independently movable sub-blocks that can deform the overall shape, targeting the 1–5 m/s band.

Urban areas often experience wind from various directions throughout the year, and comfort requirements may differ between seasons and climate zones. In winter, shelter from cold wind is desirable (low target speeds), whereas in summer, some airflow is desired for cooling (moderate target speeds), provided it does not introduce dangerous gusts. Figure 11 demonstrates a climate-adaptive design scenario with two conflicting wind directions, a left inlet at 15 m/s tar-



(a) Left inlet: initial **(b)** Left inlet: optimized **(c)** Left inlet: distribution

(d) Top inlet: initial **(e)** Top inlet: optimized **(f)** Top inlet: distribution

Fig. 11: Multi-inlet *rigid* optimization with left inlet $(15, 0)$ m/s (target comfort band 1–3 m/s) and top inlet $(0, 15)$ m/s (target comfort band 3–5 m/s). A single layout is co-optimized for both inlets and evaluated under each direction separately.

geting a comfort band of 1–3 m/s and a top inlet at 15 m/s targeting 3–5 m/s. A single layout must satisfy both objectives simultaneously. The wind speed distribution for left inlet (Fig. 11c) concentrates mass toward the 1–3 m/s band, while the distribution for top inlet (Fig. 11f) shifts mass into the 3–5 m/s range. The corresponding *morph* and *rigid* results are shown in the supplement.

6 Conclusion

This work demonstrates that a video diffusion model pretrained on natural video can be fine-tuned into a quantitatively accurate surrogate for urban wind simulation, and that it can outperform neural operators designed specifically for PDE solving. The resulting model generates 112-frame rollouts in under a second, roughly three orders of magnitude faster than the CFD solver. The combination of speed and inherent differentiability enables a design workflow that is difficult to achieve with conventional CFD: candidate building layouts can be assessed in real time, and gradients obtained by backpropagation through the predicted flow field can be used to adjust geometry toward improved wind comfort. Our inverse optimization experiments show that this works for single and multiple inlet directions, with both *rigid* translation and continuous *morph*ing of buildings. The current sub-block parametrization could easily be extended to richer representations, such as per-face offsets or spline-based footprints, which would allow for realistic shape exploration. Furthermore, the objective itself could incorporate additional differentiable constraints, such as minimum passage widths

or pedestrian-flow norms. A complementary direction is to pair our gradient-based optimizer with generative inverse-design methods that impose a learned prior over candidate solutions, such as CinDM [56] and DiffPhyCon [53], which could further improve the discovered layouts.

The current formulation operates in 2D. An extension to 3D is relatively straightforward: instead of binary occupancy masks, building heights can be encoded as continuous values in the conditioning frame. The pretrained video models have substantial unused capacity, as shown by the fact that only two denoising steps are enough for our current flow fields. We expect the bottleneck to be the data, because generating 3D urban CFD at comparable diversity and resolution would require substantially more computational resources.

We selected LTX-Video for its open weights, active community, and focus on efficient generation. We expect our findings to transfer to other non-autoregressive diffusion models such as Wan 2.2 [51]. An interesting experiment would be to test autoregressive video models such as MAGI-1 [47], in particular, whether error accumulation across sequential generation steps would pose problems for inverse optimization, similar to what we observe for OFormer in our surrogate comparison (cf. supplement).

Acknowledgements

We thank Cornelia Kalender for guidance on evaluation metrics, Kai Marti for the real-world building footprints, and Spencer Folk for access to the fluid-solver codebase. This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a0120 on Alps.

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
2. CEN: Eurocode 1: Actions on structures – part 1-4: General actions – wind actions. Standard (2010)
3. Chen, C., Tian, G., Qin, S., Yang, S., Geng, D., Zhan, D., Yang, J., Vidal, D., Wang, L.L.: Generalization of urban wind environment using fourier neural operator across different wind directions and cities. arXiv preprint arXiv:2501.05499 (2025)
4. City of London Corporation: Wind microclimate guidelines for developments in the city of london. `https://www.cityoflondon.gov.uk/assets/Services-Environment/wind-microclimate-guidelines.pdf`, last accessed 2026-06-26 (2019)
5. Clarke, A., Giljarhus, K.E.T., Oggiano, L., Saddington, A., Depuru-Mohan, K.: Deep learning for urban wind prediction: An mlp-mixer approach with 3d encoding. Building and Environment p. 113495 (2025)
6. Clemente, A.V., Giljarhus, K.E.T., Oggiano, L., Ruocco, M.: Rapid pedestrian-level wind field prediction for early-stage design using pareto-optimized convolutional neural networks. Computer-Aided Civil and Infrastructure Engineering **39**(18), 2826–2839 (2024)
7. Du, P., Parikh, M.H., Fan, X., Liu, X.Y., Wang, J.X.: Conditional neural field latent diffusion model for generating spatiotemporal turbulence. Nature Communications **15**(1), 10416 (2024)
8. Folk, S., Melton, J., Margolis, B.W.L., Yim, M., Kumar, V.: Learning local urban wind flow fields from range sensing. IEEE Robotics and Automation Letters **9**(9), 7413–7420 (2024). `https://doi.org/10.1109/LRA.2024.3426209`
9. Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: Force prompting: Video generation models can learn and generalize physics-based control signals. arXiv preprint arXiv:2505.19386 (2025)
10. Giral, F., Manzano, Á., Gómez, I., Koumoutsakos, P., Clainche, S.L.: Generative urban flow modeling: From geometry to airflow with graph diffusion. arXiv preprint arXiv:2512.14725 (2025)
11. Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., Catanzaro, B.: Efficient token mixing for transformers via adaptive fourier neural operators. In: International conference on learning representations (2021)
12. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
13. Herde, M., Raonić, B., Rohner, T., Käppeli, R., Molinaro, R., De Bezenac, E., Mishra, S.: Poseidon: Efficient foundation models for pdes. Advances in Neural Information Processing Systems **37**, 72525–72624 (2024)
14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
15. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)

16. Huang, C., Zhang, G., Yao, J., Wang, X., Calautit, J.K., Zhao, C., An, N., Peng, X.: Accelerated environmental performance-driven urban design with generative adversarial network. Building and Environment **224**, 109575 (2022)
17. Isyumov, N.: The ground level wind environment in built-up areas. In: Proc. of the 4th Int. Conf. on Wind Effects on Buildings and Structures, London, 1975 (1975)
18. Kaseb, Z., Rahbar, M.: Towards cfd-based optimization of urban wind conditions: Comparison of genetic algorithm, particle swarm optimization, and a hybrid algorithm. Sustainable Cities and Society **77**, 103565 (2022)
19. Kastner, P., Dogan, T.: A gan-based surrogate model for instantaneous urban wind flow prediction. Building and Environment **242**, 110384 (2023)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Lawson, T.V.: The wind content of the built environment. Journal of Wind Engineering and Industrial Aerodynamics **3**(2–3), 93–105 (1978)
22. Letzel, M.O.: High resolution large-eddy simulation of turbulent flow around buildings (2007)
23. Li, E., Wang, Z., Huang, J., Park, J.J.: Videopde: Unified generative pde solving via video inpainting diffusion models. arXiv preprint arXiv:2506.13754 (2025)
24. Li, Z., Meidani, K., Farimani, A.B.: Transformer for partial differential equations' operator learning. arXiv preprint arXiv:2205.13671 (2022)
25. Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Fourier neural operator for parametric partial differential equations. arXiv preprint arXiv:2010.08895 (2020)
26. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7708–7717 (2019)
27. Liu, Z., Zhang, S., Shao, X., Wu, Z.: Accurate and efficient urban wind prediction at city-scale with memory-scalable graph neural network. Sustainable Cities and Society (2023)
28. Liu-Schiaffini, M., Singer, C.E., Kovachki, N., Schneider, T., Azizzadenesheli, K., Anandkumar, A.: Tipping point forecasting in non-stationary dynamics on function spaces. arXiv preprint arXiv:2308.08794 (2023)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
30. Lu, J., Li, W., Hobeichi, S., Azad, S.A., Nazarian, N.: Machine learning predicts pedestrian wind flow from urban morphology and prevailing wind direction. Environmental Research Letters **20**(5), 054006 (2025)
31. Mirzaei, P.A.: Cfd modeling of micro and urban climates: Problems to be solved in the new decade. Sustainable Cities and Society **69**, 102839 (2021)
32. Mokhtar, S., Beveridge, M., Cao, Y., Drori, I.: Pedestrian wind factor estimation in complex urban environments. In: Asian conference on machine learning. pp. 486–501. PMLR (2021)
33. Molinaro, R., Lanthaler, S., Raonić, B., Rohner, T., Armegioiu, V., Simonis, S., Grund, D., Ramic, Y., Wan, Z.Y., Sha, F., et al.: Generative ai for fast and accurate statistical computation of fluids. arXiv preprint arXiv:2409.18359 (2024)
34. Nguyen, T., Koneru, A., Li, S., Grover, A.: Physix: A foundation model for physics simulations. arXiv preprint arXiv:2506.17774 (2025)
35. Nilol Kundur Satish, S., Jaiswal, D., Chen, H., Bakshi, A.: PhysVideoGenerator: Towards Physically Aware Video Generation via Latent Physics Guidance. arXiv e-prints arXiv:2601.03665 (Jan 2026). https://doi.org/10.48550/arXiv.2601.03665

36. Ohana, R., McCabe, M., Meyer, L., Morel, R., Agocs, F.J., Beneitez, M., Berger, M., Burkhart, B., Dalziel, S.B., Fielding, D.B., et al.: The well: a large-scale collection of diverse physics simulations for machine learning. Advances in Neural Information Processing Systems **37**, 44989–45037 (2024)
37. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
38. Qin, S., Zhan, D., Geng, D., Peng, W., Tian, G., Shi, Y., Gao, N., Liu, X., Wang, L.L.: Modeling multivariable high-resolution 3d urban microclimate using localized fourier neural operator. Building and Environment **273**, 112668 (2025)
39. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
40. Rui, E.Z., Chen, Z.W., Ni, Y.Q., Yuan, L., Zeng, G.Z.: Reconstruction of 3d flow field around a building model in wind tunnel: a novel physics-informed neural network framework adopting dynamic prioritization self-adaptive loss balance strategy. Engineering Applications of Computational Fluid Mechanics **17**(1), 2238849 (2023)
41. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE international conference on computer vision. pp. 2830–2839 (2017)
42. Shao, X., Liu, Z., Zhang, S., Zhao, Z., Hu, C.: Pignn-cfd: A physics-informed graph neural network for rapid predicting urban wind field defined on unstructured mesh. Building and Environment **232**, 110056 (2023)
43. Snaiki, R., Lu, J., Li, S., Nazarian, N.: A hierarchical deep learning model for predicting pedestrian-level urban winds. Building and Environment p. 114354 (2026)
44. Soares, E., Brazil, E.V., Shirasuna, V., de Carvalho, B.W., Malossi, C.: Towards a foundation model for partial differential equations across physics domains. arXiv preprint arXiv:2511.21861 (2025)
45. Takamoto, M., Praditia, T., Leiteritz, R., MacKinlay, D., Alesiani, F., Pflüger, D., Niepert, M.: Pdebench: An extensive benchmark for scientific machine learning. Advances in neural information processing systems **35**, 1596–1611 (2022)
46. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. Advances in neural information processing systems **33**, 7537–7547 (2020)
47. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
48. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
49. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837 (2024)
50. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. Advances in neural information processing systems **29** (2016)
51. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
52. Wang, C., Chen, C., Huang, Y., Dou, Z., Liu, Y., Gu, J., Liu, L.: Physctrl: Generative physics for controllable and physics-grounded video generation. arXiv preprint arXiv:2509.20358 (2025)

53. Wei, L., Hu, P., Feng, R., Feng, H., Du, Y., Zhang, T., Wang, R., Wang, Y., Ma, Z.M., Wu, T.: Diffphycon: a generative approach to control complex physical systems. Advances in Neural Information Processing Systems **37**, 4090–4147 (2024)
54. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
55. Wiesner, F., Wessling, M., Baek, S.: Towards a physics foundation model. arXiv preprint arXiv:2509.13805 (2025)
56. Wu, T., Maruyama, T., Wei, L., Zhang, T., Du, Y., Iaccarino, G., Leskovec, J.: Compositional generative inverse design. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=wmX0CqFSd7`, last accessed 2026-06-26
57. Wu, Y., Quan, S.J.: A review of surrogate-assisted design optimization for improving urban wind environment. Building and Environment **253**, 111157 (2024)
58. Zhang, K., Xiao, C., Xu, J., Mei, Y., Patel, V.M.: Think before you diffuse: Infusing physical rules into video diffusion. arXiv preprint arXiv:2505.21653 (2025)