VIDEO FOUNDATION MODELS ARE NATIVE SPATIO-TEMPORAL FIELD FORECASTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural videos and urban spatio-temporal fields exhibit similar patterns of spatial evolution, suggesting that video pretraining may provide useful forecasting representations. However, the benefits of video pretraining and the feasibility of sharing a backbone across heterogeneous urban domains remain insufficiently explored. We introduce VFM4ST, a framework that renders numerical fields as pseudo-videos and jointly encodes spatial and temporal structure using a pretrained video foundation model. Domain-specific three-port decoders separate forecast baselines, field evolution, and output geometry. Direct access to native observations retains local information for forecasting on both dense grids and sparse stations. Across traffic flow, radar nowcasting, crime, and air quality, one shared 87M-parameter backbone achieves better mean performance than the strongest evaluated baseline in every domain. The shared model remains within 2.2% of independently specialised models in each domain. Under the same joint training protocol, video pretraining outperforms image pretraining and random initialisation across all four domains. Frozen-backbone and low-data evaluations further support the value of pretrained video representations.

1 INTRODUCTION

Urban spatio-temporal forecasting supports traffic management, radar nowcasting, crime analysis, and air quality monitoring. Forecasting models must accommodate different variables, observation layouts, and temporal scales. Urban foundation models such as OpenCity (Li et al., 2025) and Uniflow (Yuan et al., 2026) seek to reuse representations across forecasting scenarios. An additional opportunity comes from natural videos, which exhibit evolving spatial patterns resembling those in urban fields. Large-scale video pretraining may therefore provide useful representations beyond those learned from urban observations alone.

From this perspective, video foundation models (VFMs) offer a suitable starting point by encoding spatial and temporal structure jointly (Tong et al., 2022; Wang et al., 2023a). We investigate VFMs as native field forecasters, where native refers to the compatibility of spatio-temporal field sequences with joint video encoding. Encoding each location as a separate time series omits explicit spatial relationships, while frame-wise image encoding provides no pretrained temporal representation (Chen et al., 2025; Zhao et al., 2026)(Figure 1). Prior work demonstrates the potential of video architectures for forecasting (He et al., 2023; Gao et al., 2022) and visual pretraining for time series (Chen et al., 2024), but neither alone establishes the value of video-pretrained representations across heterogeneous urban domains. Studies on fully observed synthetic fields also leave sparse urban observations and cross-domain sharing insufficiently tested (Nguyen et al., 2025; Perini et al., 2026). Two questions therefore motivate our study: **does video pretraining improve forecasting across such domains, and can one shared backbone support different observation structures and forecasting requirements?**

Answering these questions requires addressing domain differences beyond structural compatibility. Rendering and coarse tokenisation can obscure local information needed at native observation locations. Forecasting also requires different reference fields and evolution mechanisms: radar echoes exhibit spatial transport and intensity changes, whereas station observations require location-specific predictions. Dense grids and sparse stations further demand different output geometries. Joint training must also accommodate different history lengths, forecast horizons, and loss scales. Reusing

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.

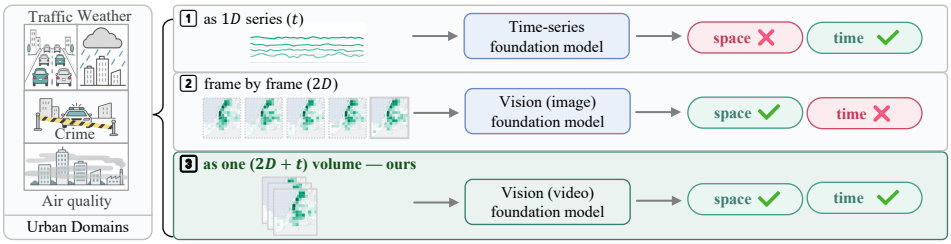


Figure 1: Three ways to encode a spatio-temporal field. Independent per-location series encode temporal variation without explicit spatial relationships; frame-wise image encoding supplies spatial features without a pretrained temporal representation; joint video encoding represents both dimensions within the backbone. VFM4ST adopts joint spatio-temporal encoding.

video representations therefore requires separating shared feature extraction from domain-specific prediction.

To address these requirements, we introduce VFM4ST, which renders numerical histories as pseudo-videos and encodes the complete spatio-temporal volume using a pretrained video backbone. Domain-specific three-port decoders separate forecast baselines, field evolution, and output geometry. The anchor supplies a reference field, dynamics combines optional advection with additive corrections, and readout maps predictions to dense grids or sparse stations. Direct access to native observations retains local information alongside backbone features. The shared variant jointly adapts one backbone using domain conditioning and balanced domain sampling; independently specialised variants adapt a separate model to each domain. We evaluate VFM4ST across traffic flow, radar nowcasting, crime, and air quality to assess both backbone sharing and the contribution of video pretraining. One shared 87M-parameter backbone achieves better mean performance than the strongest evaluated baseline in every domain. Shared performance remains within 2.2% of independently specialised models. Under the same joint training protocol, video pretraining outperforms image pretraining and random initialisation across all four domains. Frozen-backbone and low-data evaluations provide additional evidence for the value of pretrained video representations. Further analyses examine decoder components, training cost, and transfer to a held-out domain. Our contributions are threefold:

- We systematically investigate large-scale video pretraining across four real-world urban spatio-temporal domains with heterogeneous observation structures. Complementary initialisation, frozen-backbone, and low-data comparisons provide consistent evidence for the value of pretrained video representations.
- We introduce VFM4ST, a unified framework that processes numerical fields as complete spatio-temporal volumes through a pretrained video backbone. A three-port decoder separates shared representations from domain-specific forecast baselines, field evolution, and output geometry, preserving access to native information across dense grids and sparse stations.
- We demonstrate effective cross-domain sharing: a single backbone outperforms the strongest evaluated baseline in every domain and remains within 2.2% of independently specialised models. Comprehensive analyses of optimisation dynamics, training cost, and held-out-domain transfer characterise the benefits and limits of sharing video-pretrained representations.

2 Related Work

Urban spatio-temporal forecasting. Task-specific models capture road-network dependencies, radar evolution, regional crime patterns, and station-level air quality (Wu et al., 2019; Gao et al., 2022; Huang et al., 2018; Liang et al., 2023). Urban foundation models extend representation reuse across cities and tasks through shared encoders and domain conditioning (Li et al., 2025; 2024; Yuan et al., 2026; 2024). Complementing representation learning from urban observations, VFM4ST investigates natural-video pretraining for heterogeneous forecasting domains. Domain-specific decoders accommodate different forecast baselines and output geometries while retaining a shared video backbone.

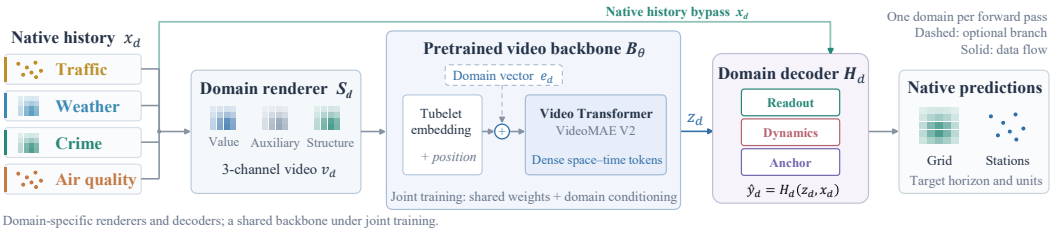


Figure 2: **Framework overview.** A domain-specific renderer converts native observations into a three-channel pseudo-video while preserving their temporal order. A shared pretrained video backbone extracts spatio-temporal tokens, which a domain-specific decoder maps to predictions at the target horizon and in native units.

Transfer to numerical sequences and fields. Video-style forecasting architectures demonstrate the relevance of joint spatial and temporal modelling, but training from scratch does not establish the value of video pretraining (He et al., 2023; Harris & Chen, 2025). Time-series foundation models demonstrate transfer across numerical sequences (Ansari et al., 2024; Das et al., 2023; Woo et al., 2024; Rasul et al., 2023; Liu et al., 2024), while visual representations also support sequence forecasting (Chen et al., 2024; Pavlyuk, 2020). Sequence-level transfer alone does not establish the benefit of jointly encoding spatial fields over time. ST-VFM and ST-VLM bring image or vision-language pretraining to traffic forecasting (Chen et al., 2025; Zhao et al., 2026), but do not transfer a jointly pretrained video representation. PhysiX and WinDiNet extend video pretraining to physical simulations (Nguyen et al., 2025; Perini et al., 2026), providing evidence of transfer beyond natural videos. VFM4ST examines transfer to real urban observations, covering both dense grids and sparse stations, and evaluates joint backbone adaptation across domains.

Video foundation models. Video pretraining includes masked reconstruction (Tong et al., 2022; Wang et al., 2023a;b), latent prediction (Assran et al., 2025), and multimodal alignment (Wang et al., 2024; Zhao et al., 2024). SciVid evaluates video models across scientific tasks, primarily through frozen backbones with task-specific readouts, and demonstrates the potential of pretrained representations for scientific applications (Hasson et al., 2025). Beyond transfer to individual tasks, VFM4ST investigates whether joint adaptation can support heterogeneous urban forecasting with one backbone. Using VideoMAE V2 (Wang et al., 2023a), the study evaluates video pretraining and cross-domain sharing through initialisation, frozen-backbone, low-data, and shared-versus-specialised comparisons.

3 Methodology

Heterogeneous spatio-temporal field forecasting. For a domain d, let $x_d \in \mathbb{R}^{T_{\text{in}}^d \times N_d \times C_d}$ denote the observed history and y_d the target field over the next T_{out}^d steps, where N_d indexes grid cells or observation stations. Domains differ in observation layout, variable semantics, temporal span, and target dimension. The goal is to reuse the spatio-temporal representation obtained from video pretraining while preserving these differences.

Framework and separation principle. VFM4ST consists of a domain renderer S_d, a video backbone B_θ, and a domain decoder H_d:

$$z_d = B_\theta(S_d(x_d)), \qquad \hat{y}_d = H_d(z_d, x_d).$$

The renderer organises the numerical history as a video input, the backbone extracts spatio-temporal features, and the decoder produces the forecast through three ports: *anchor* supplies the forecast baseline, *dynamics* parameterises the evolution of the future field, and *readout* returns predictions at the native observation locations. The decoder receives the native history x_d directly, so information lost during rendering and tokenisation remains available to it (Figure 2).

3.1 Domain Renderer: From Numerical Fields to Video Inputs

The domain renderer S_d maps numerical histories to three-channel video inputs $v_d = S_d(x_d)$ through spatial embedding, channel encoding, and temporal organisation.

Spatial embedding. Each historical observation is mapped onto a two-dimensional canvas. Regular grids retain spatial layout, whereas sparse stations are embedded through a fixed mapping from sensor coordinates to canvas cells, with observation locations explicitly recorded. The canvas is resized to the backbone input resolution using bilinear interpolation for continuous channels and nearest-neighbour resampling for masks and categorical channels.

Channel encoding. Each frame contains three channels: a value channel carrying the normalised target variable, an auxiliary channel providing an observation mask or a complementary transformation of observed values, and a structure channel encoding spatial support, positional identity, or another domain-specific cue. Channel definitions vary across domains while maintaining a common three-channel format. Normalisation statistics are estimated exclusively from the training set, domain-specific mappings are provided in Appendix A.

Temporal organisation. Frames retain the original observation order without temporal interpolation. If the history length is not divisible by the backbone's temporal tubelet size, the final frame is repeated to complete the last tubelet, allowing different history lengths across domains.

3.2 Pretrained Video Backbone and Adaptation

Dense spatio-temporal feature extraction. A pretrained video Transformer B_θ encodes the rendered history v_d into a dense token grid,

$$z_d = B_\theta(v_d) \in \mathbb{R}^{T'_d \times H' \times W' \times D},$$

where T'_d, H', and W' denote the temporal and spatial token dimensions, and D is the feature dimension. The backbone preserves the spatio-temporal token structure without global pooling, providing local features for field reconstruction.

Per-domain adaptation. Each domain independently adapts the backbone from the same pretrained parameters θ_0 and trains a corresponding decoder. Backbone adaptation is expressed as $\theta_d = \theta_0 + \Delta\theta_d$. The update $\Delta\theta_d$ covers freezing, low-rank adaptation, and partial or full fine-tuning. These schemes differ in the scope and parameterisation of trainable parameters while preserving the same feature interface to the decoder. All domains reuse the same framework design but share no adapted parameters. The adaptation scheme for each domain is specified in the experimental setup.

Multi-domain sharing and conditioning. Joint training uses a single set of backbone parameters θ, while retaining domain-specific renderers and decoders. A learnable domain vector $e_d \in \mathbb{R}^D$ is added to each input token embedding to provide domain identity within the shared backbone. Domain conditioning distinguishes variable and task contexts, while the corresponding decoder handles domain-specific forecasting requirements and output geometry. Independently trained domain models omit the domain vector.

Joint optimisation. The shared backbone, domain vectors, and domain decoders are jointly optimised with

$$\mathcal{L}_{\text{joint}} = \mathbb{E}_{d\sim p(d)} \mathbb{E}_{(x_d, y_d)\sim \mathcal{D}_d} \left[\mathcal{L}_d\right],$$

where $\mathcal{D}_d$ is the training set for domain d, $p(d)$ is uniform over domains, and $\mathcal{L}_d$ is the domain loss defined in Section 3.4. Uniform domain sampling prevents dataset size from determining update frequency, but does not equalise loss or gradient magnitudes. Per-domain and shared training assess independent adaptation and cross-domain parameter sharing, respectively. Because adaptation schemes and training budgets differ between the two (Appendix B), comparisons reflect the complete training configurations; attributing differences specifically to parameter sharing requires matching these factors.

3.3 Three-Port Domain Decoder

To accommodate heterogeneous forecasting requirements, VFM4ST separates forecast baselines, field evolution, and output geometry through a three-port domain decoder. The decoder combines backbone features with direct access to the native history, preserving information that may be lost during rendering and tokenisation. The three ports, *Anchor*, *Dynamics*, and *Readout*, construct the prediction as

$$a_d = A_d(x_d), \qquad u_d = \mathcal{W}_d(a_d, f_d) + \Delta_d, \qquad \hat{y}_d = R_d(u_d; x_d, z_d),$$

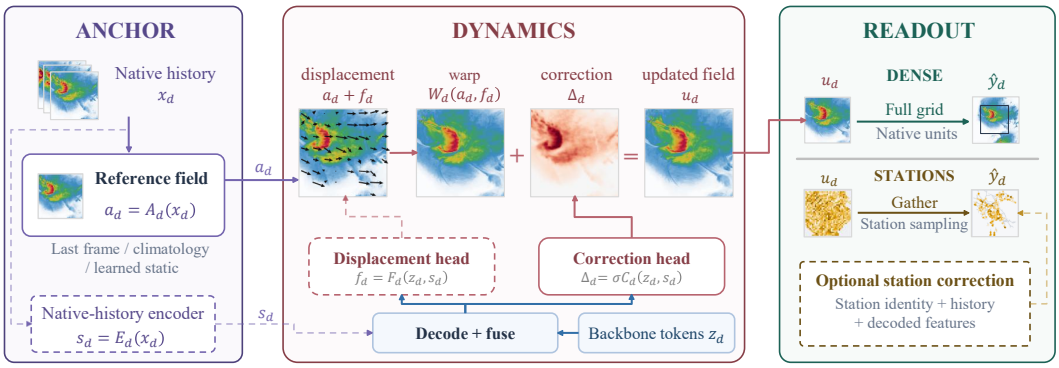


Figure 3: **Three-port decoder.** The anchor port constructs a reference field from persistence, climatology, or a learned static field. The dynamics port combines an optional displacement-based warp with an additive correction. The readout port maps the resulting field to dense grids or sparse observation stations, with optional station-specific correction.

where a_d is the reference field, f_d is the displacement field, Δ_d is the additive correction, and u_d is the intermediate predicted field as shown in Figure 3. The transformation $\mathcal{W}_d$ warps the reference when displacement modelling is enabled and otherwise reduces to the identity. Across domains, a common feature decoding structure concatenates temporal tokens along the feature dimension and applies spatial projection and progressive upsampling to obtain features for future prediction. Domain-specific heads accommodate the target spatial layout, forecast horizon, and variable dimensions.

3.3.1 Anchor: Reference Field

The anchor supplies a reference field, allowing future predictions to be parameterised relative to an explicit baseline. The reference is constructed from the latest observation, a conditional training-set mean, or a learnable static field, and is defined over the forecast horizon in coordinates compatible with field reconstruction. By entering the prediction directly, the reference bypasses the backbone's spatial compression and preserves the baseline at the reconstruction resolution.

3.3.2 Dynamics: Displacement and Additive Correction

The dynamics port models changes relative to the reference through displacement and additive correction. Backbone features are optionally supplemented with native-history features $s_d = E_d(x_d)$ to retain local detail, yielding

$$f_d = F_d(z_d, s_d), \qquad \tilde{\Delta}_d = C_d(z_d, s_d),$$

with s_d omitted when the native-history branch is disabled. For fields with spatial transport, f_d specifies a two-dimensional displacement at each forecast step and warps the reference through differentiable backward sampling. The additive branch models intensity changes and the formation or dissipation of structure, complementing the positional changes represented by displacement.

The correction is parameterised as $\Delta_d = \sigma_d \tilde{\Delta}_d$, where σ_d is estimated from the training set so that the correction matches the reference's units or normalised coordinates. Zero-initialising both prediction outputs gives an identity warp and zero correction, so the intermediate field initially equals the reference. Tasks without displacement modelling use the same formulation with an identity transform, reducing to $u_d = a_d + \Delta_d$.

3.3.3 Readout: Native Output Geometry

The readout maps the intermediate field to the target observation layout. Dense tasks return predictions on the native grid, while sparse tasks use a fixed station mapping to sample the intermediate field and decoded features. Inverse normalisation applies the corresponding spatial and variable statistics to recover predictions in native units.

To distinguish nearby stations with different temporal behaviour, an optional station-conditioned head combines sampled features with station identity and, when needed, each station's original history. Direct access to station histories supplies local information without rasterisation-induced mixing. The correction output is zero-initialised, so the head initially preserves the base coordinate readout.

3.4 Training Objective

The model learns to forecast future fields by minimising the L1 error over valid target observations:

$$\mathcal{L}_d = \frac{\sum_{\tau,i} m^d_{\tau i}\, w^d_{\tau i}\, \left|\hat{y}^d_{\tau i} - y^d_{\tau i}\right|}{\sum_{\tau,i} m^d_{\tau i}\, w^d_{\tau i}}.$$

Here, τ indexes forecast steps, i indexes target entries across spatial locations and variables, m^d is the valid-observation mask, and w^d denotes optional weights over forecast steps or targets.

Losses are computed after inverse normalisation, in the native units of each target. Joint training uses uniform domain sampling without additional domain weights, balancing update frequency while retaining differences in loss scale.

4 Experiments

4.1 Experimental Setup

Domains and protocols. We evaluate four urban forecasting domains: traffic flow (UTD19-Zurich (Loder et al., 2019)), radar vertically integrated liquid (SEVIR (Veillette et al., 2020)), daily complaint counts (NYPD), and hourly PM2.5 (Beijing Multi-Site (Chen, 2017)). These cover dense grids (weather and crime) and sparse stations (traffic and air quality), with domain-specific input lengths and forecast horizons following the corresponding baseline literature (Table 1).

Table 1: Datasets and forecasting protocols across four domains.

	Traffic	Weather	Crime	Air Quality
Source	UTD19-Zurich	SEVIR VIL	NYPD Complaint	Beijing Multi-Site
Target	flow, veh/h/lane	VIL, 8-bit code	next-day complaint count	$PM_{2.5}$, $\mu g/m^3$
Spatial unit	1,020 detectors (sparse)	128×128 grid	18×17 cells	12 stations (sparse)
Cadence	3 min	5 min	1 day	1 h
$T_{\text{in}} \to T_{\text{out}}$	$20 \to 20$	$13 \to 12$	$30 \to 1$	$24 \to 24$
Anchor	learnable static	last frame	climatology	learnable static
Primary metric	Masked MAE@60m ↓	CSI-M ↑	MAE ↓	Masked MAE@24h ↓
Train / Val / Test	1,820 / 614 / 614	32,146 / 8,150 / 10,944	580 / 30 / 90	24,817 / 4,345 / 5,809

Baselines and evaluation. We compare against persistence, periodic/climatological means, and linear autoregression, alongside 18 learned baselines across five groups: domain-specific models, urban foundation models, video-style architectures trained from scratch, time-series foundation models, and image foundation models (Appendix B). For domain-specific models, we select the most recent state-of-the-art method(s) in each domain rather than earlier standard baselines; each of these methods has already been shown to outperform widely used classical models in its own domain. Learned baselines are trained separately per domain under the same domain-specific protocol. Primary metrics are masked MAE at +60 min for traffic and +24 h for air quality, CSI averaged over six VIL thresholds (CSI-M) for weather, and all-cell MAE for crime. For learned models, we report test-set mean $\pm$ standard deviation over three seeds.

VFM4ST variants and implementation. Both main variants start from the same pretrained VideoMAE V2 ViT-B backbone. The *shared* variant jointly fine-tunes one backbone across all four domains using balanced domain sampling, a domain token, and domain-specific decoders. The *specialised* variant trains a separate model per domain, with backbone adaptation selected on the validation set: LoRA (Hu et al., 2021) for traffic and air quality, full fine-tuning for weather, and partial unfreezing for crime. All models use AdamW (Loshchilov & Hutter, 2017) with a cosine learning-rate schedule. Per-domain anchor types (Table 1) are likewise selected on the validation set under specialised training and are shared by both variants (Appendix E). Table B1 summarises all variants. Backbone and decoder configurations, training budgets, baseline descriptions, and representative baseline costs are provided in Appendix B.

4.2 Main Results: One Backbone across Four Domains

VFM4ST supports all four domains with one shared 87M-parameter backbone and domain-specific decoders, achieving better mean performance than the strongest evaluated baseline in each domain (Table 2). On traffic, the shared variant reduces MAE by 2.5% relative to RAGC; on weather, CSI-M increases by 7.4% relative to FlowCast. On crime and air quality, MAE decreases by 2.6% and 2.3% relative to UniFlow and Air-DualODE, respectively. Although the strongest baseline varies across domains, the same backbone weights support both dense-grid and sparse-station forecasting under different history lengths and forecast horizons.

The specialised variant achieves the best mean performance in all four domains, while the shared variant remains within 2.2% of the specialised result in each domain. These results demonstrate that a shared video backbone can retain accuracy close to independently adapted models across heterogeneous forecasting tasks. The comparison reflects complete training configurations, including different adaptation strategies and budgets. Section 4.3 examines the contribution of video pretraining, and Section 4.4 analyses the trade-off between sharing and specialisation.

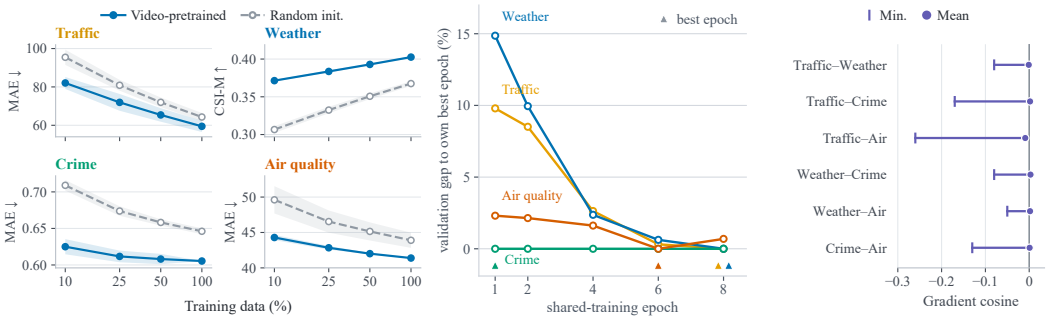
Table 2: Main results: one shared backbone across four domains.

Method	Pretraining	Traffic MAE ↓	Weather CSI-M ↑	Crime MAE ↓	Air MAE ↓
Statistical baselines					
Persistence	–	80.90	0.2427	0.8395	48.52
Periodic / clim. mean	–	109.69	0.0451	0.6577	56.13
Linear AR	–	68.01	0.2168	0.6673	48.29
Domain-specific models					
MetaDG (AAAI 2026)	none	63.53±1.22	–	–	–
RAGC (ICDE 2026)	none	62.29±0.77	–	–	–
FlowCast (ICLR 2026)	radar images (VAE)	–	0.3734±0.005	–	–
SDIR (ICML 2026)	none	–	0.3201±0.011	–	–
ST-HHOL (ICLR 2026)	text (LLM)	–	–	0.6306±0.013	–
AirDDE (AAAI 2026)	none	–	–	–	45.28±1.82
Air-DualODE (ICLR 2025)	none	–	–	–	42.99±1.27
Urban foundation models					
UniST (KDD 2024)	urban ST	66.84±2.30	0.2544±0.006	0.6294±0.011	44.64±0.98
OpenCity (ACM TIST 2025)	urban ST	63.86±1.14	0.2382±0.017	0.6387±0.013	43.58±1.60
UrbanGPT (KDD 2024)	urban ST + LLM	65.90±1.80	0.2050±0.013	0.6341±0.016	44.93±0.73
UniFlow (IEEE TMC 2026)	urban ST	63.61±1.53	0.2712±0.010	0.6218±0.009	43.16±1.89
Cross-modal transfer: video-style architectures, time-series and image foundation models					
ST-3DGMR (Inf. Sci. 2023)	none, video architecture	71.52±4.30	0.3156±0.011	0.6493±0.004	45.60±1.91
SaTformer (NeurIPS 2025)	none, video architecture	72.34±4.11	0.2960±0.018	0.6611±0.005	45.93±1.63
Chronos (TMLR 2024)	1-D time series	66.95±1.93	0.2297±0.013	0.6482±0.014	46.37±0.92
TimesFM (ICML 2024)	1-D time series	66.27±2.01	0.2214±0.005	0.6628±0.008	46.86±1.83
Moirai (ICML 2024)	1-D time series	65.47±2.90	0.2259±0.007	0.6557±0.003	45.61±0.74
VisionTS (ICML 2025)	images, applied to 1-D series	67.33±2.36	0.2183±0.008	0.6693±0.010	47.23±0.57
ST-VLM (AAAI 2026)	image–text (ViLT), correlation images	66.52±2.16	0.2348±0.005	0.6417±0.010	46.03±1.26
Ours					
VFM4ST shared	natural video	60.75±0.36	0.4010±0.006	0.6059±0.000	42.02±0.17
VFM4ST specialised	natural video	**59.46±0.95**	**0.4026±0.001**	**0.6054±0.000**	**41.39±0.04**

Mean ± std over three seeds. Bold, underlining, and dashed underlining denote the best, second-best, and third-best results in each column; ties share bold. Every baseline and the specialised variant train one model per domain. The shared variant uses a single backbone across all four domains. A displayed standard deviation of 0.000 reflects rounding, not zero seed variation.

4.3 Does Video Pretraining Account for the Gain?

Pretraining and adaptation. To examine the contribution of video pretraining, we compare specialised, frozen, image-pretrained, and randomly initialised variants, each trained on its own domain (Table C1, Appendix C). Relative to the specialised video-pretrained models, training from scratch increases MAE by 8.2% on traffic, 6.7% on crime, and 6.1% on air quality, and reduces CSI-M by 8.7% on weather. Freezing the backbone costs 6.8%, 4.2%, 1.8%, and 3.6%, yet the frozen model still exceeds both the scratch and image-pretrained variants in every domain. Specialised models use domain-selected adaptation strategies; scratch and image-pretrained variants use full fine-tuning.



(a) Data efficiency (b) Training dynamics (c) Gradient alignment

Figure 4: Data efficiency and optimisation dynamics across four forecasting domains. (a) Shading indicates standard deviation over three seeds. (b) Coloured triangles mark the best reported epochs. (c) Circles and vertical caps indicate mean and minimum gradient cosine similarities, respectively.

Contribution of the video pathway. We additionally compare the randomly initialised video model with a parameter-matched numeric MLP operating on native inputs with the same anchor and residual parameterisation. Relative to the specialised model, the MLP loses 22.9% on traffic and 27.4% on weather, against 8.2% and 8.7% for the randomly initialised video model, so the video pathway accounts for 14.8 and 18.7 percentage points of the gap; on crime and air quality the two deficits differ by only 0.7 and 0.1 percentage points, respectively. Moreover, the scratch variant remains below the strongest evaluated baseline in every domain.

Data efficiency. We compare the specialised and scratch variants using 10%, 25%, 50%, and 100% of each training set with matched optimisation steps. The benefit of video pretraining generally increases as training data become scarcer (Figure 4a). With only 25% of the training data, the specialised models outperform scratch models trained on the full dataset on weather, crime, and air quality.

Analysis. The frozen-backbone results demonstrate the usefulness of pretrained video features, while the larger low-data gains suggest reduced dependence on target-domain supervision. The video pathway's uneven benefits reflect differences in observable spatial evolution: traffic and radar contain propagation patterns, whereas daily crime counts and sparse air-quality observations may provide less coherent or less fully observed evolution. The small gain on the dense crime grid indicates that grid structure alone is insufficient.

4.4 Sharing vs. Specialisation

Accuracy and training cost. We compare four independently specialised models with one jointly trained backbone (Table 2). Specialised models achieve the best mean performance in every domain, while shared training remains within 2.2% of the specialised result in each domain. Under the budgets in Table B3, shared training requires 31.3 GPU-hours across all four domains, compared with 10.3 for specialisation.

Transfer to a held-out domain. To assess whether joint adaptation benefits another domain, we train the backbone on three domains, freeze it, and train the fourth domain's decoder. The reference trains the same decoder with the original video-pretrained backbone frozen. Across the four held-out domains, the jointly adapted backbone provides no improvement in mean performance: traffic and air-quality MAE increase by 1.5% and 5.2%, weather CSI-M decreases by 2.5%, and crime MAE remains nearly unchanged. Under this frozen-backbone protocol, adaptation to the other urban domains does not improve transfer beyond the original video-pretrained representation.

Optimisation dynamics. Shared training reveals different convergence patterns across domains (Figure 4b). Air quality reaches its best validation performance at epoch 6 and deteriorates by 0.7% by epoch 8, while traffic and weather continue improving and crime remains nearly unchanged. Mean pairwise gradient cosine similarities range from -0.009 to 0.003, indicating little average alignment, whereas minima reach -0.26 (Figure 4c). These near-zero means do not rule out intermittent gradient conflict.

Analysis. Sharing consolidates backbone weights while retaining near-specialised accuracy, but offers no training-cost advantage under these configurations. The held-out results suggest that adaptation may favour the observed domains' forecasting requirements without benefiting an unseen domain under the frozen-backbone protocol. Different convergence schedules also create a stopping-point compromise: further training benefits traffic and weather while degrading air quality. This may contribute to the sharing gap, although near-zero mean gradient cosines do not exclude intermittent conflict, and the differing adaptation strategies and budgets prevent attribution to sharing alone.

4.5 Ablation of Three-Port Decoder

Native detail and field evolution. Decoder ablations assess the contributions of native information and evolution modelling under the specialised training configurations and test protocols (Table 3). On weather, removing the native-history bypass reduces CSI-M by 3.8% with the last-frame anchor and by 1.6% with a static anchor, supporting a configuration-dependent benefit from native information alongside backbone features. The token-only configuration performs 11.8% below the full model, reflecting combined changes to the anchor, bypass, and warp. With the last-frame anchor and bypass fixed, removing additive correction or advection reduces CSI-M by 17.9% or 10.0%, respectively, supporting complementary contributions from both mechanisms. The pointwise regression variant reduces CSI@219 from 0.1691 to 0.0855 (49.4%), compared with a decrease from 0.7384 to 0.7203 (2.5%) at CSI@16. The larger relative loss at the high threshold indicates poorer performance on intense echoes under this configuration, but does not isolate the effect of removing advection or establish smoothing as the cause.

Station readout and domain configuration. On traffic and air quality, replacing the full readout with field interpolation increases MAE by 4.3% and 3.4%, respectively. Under the cumulative readout configuration, neighbouring histories yield the largest sequential MAE reduction on traffic, whereas own history provides the largest reduction on air quality. Coordinate queries yield only small improvements on both means, and adding own history yields a small further improvement on traffic. Without neighbouring histories, MAE remains 3.9% above the full model on traffic and 1.1% above it on air quality. The pattern supports station-conditioned readout and access to native histories, with a smaller contribution from neighbours on air quality. Additional anchor comparisons in Appendix E support domain-specific reference selection. Together, the ablations support configuring forecast baselines, field evolution, and readout to domain-specific forecasting requirements.

Table 3: Three-port decoder ablations. (A) Anchor and native-resolution bypass on weather; (B) dynamics on weather; (C) readout on station domains.

(A): Anchor	Native bypass	CSI-M↑	(B): Dynamics	CSI-M↑	CSI@16↑	CSI@219↑	(C): Readout	Traffic MAE ↓	Air MAE ↓
Last frame (full)	✓	**0.4026**	Advection + correction (full)	**0.4026**	**0.7384**	**0.1691**	Field interpolation	62.02 (+4.3%)	42.81 (+3.4%)
Last frame	–	0.3874 (-3.8%)	Advection only	0.3306 (-17.9%)	0.6978	0.1175	Coordinate query	61.83 (+4.0%)	42.65 (+3.0%)
Static	✓	0.3608 (-10.4%)	Correction only	0.3623 (-10.0%)	0.7251	0.1254	+ own history	61.75 (+3.9%)	41.85 (+1.1%)
Static (token-only)	–	0.3551 (-11.8%)	Pointwise regression	0.3408 (-15.4%)	0.7203	0.0855	+ neighbour history (full)	**59.46**	**41.39**

Percentages denote metric changes relative to the corresponding full model. Native bypass denotes the native-history feature path. Static anchors use an identity warp. Advection only sets $\Delta_d = 0$; correction only uses an identity warp. Readout additions are cumulative.

5 Conclusion

We introduced VFM4ST, a framework that transfers natural-video representations to heterogeneous urban spatio-temporal forecasting. By encoding numerical fields as joint spatio-temporal volumes and separating forecast baselines, field evolution, and output geometry through a three-port decoder, VFM4ST supports both dense grids and sparse observations. Across traffic, weather, crime, and air quality, one shared backbone outperforms the strongest evaluated baseline in each domain and remains within 2.2% of independently specialised models. Comparisons with random and image initialisation, together with frozen-backbone and low-data evaluations, support the value of video-pretrained representations. Decoder ablations further highlight the importance of native information and domain-specific prediction mechanisms. Under the frozen-backbone protocol, joint adaptation yields no additional benefit for held-out-domain transfer. Future work will explore domain-adaptive optimisation and broader transfer to unseen domains.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, translation, and code debugging. We did not use generative AI to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide key ingredients for or assist with mathematical proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. All AI-assisted text was carefully reviewed and revised by the authors, and all AI-assisted code was tested and verified for correctness. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to make this work reproducible. Our source code is available at `https://anonymous.4open.science/r/VFM4ST_ICLR_2027/`. The four datasets are publicly available. Table 1 gives the forecasting protocol for each domain. Appendix A and Table A1 cover data sources, licences, chronological splits, missing-value handling, normalisation and renderer mappings. Appendix B gives the model architecture, the adaptation schemes (Table B1) and the optimisation settings: learning rates, batch sizes, update counts and GPU-hours (Table B3). It also covers how the baselines were trained and what they cost (Table B4). For learned models, we report the mean and standard deviation of test results over three random seeds.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

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

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

Changlu Chen, Yanbin Liu, Chaoxi Niu, Ling Chen, and Tianqing Zhu. Reprogramming vision foundation models for spatio-temporal forecasting. *arXiv preprint arXiv:2507.11558*, 2025.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024.

Song Chen. Beijing Multi-Site Air Quality. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5RK5G.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Keqing Du, Yufan Kang, Xinyu Yang, and Wei Shao. St-hhol: Spatio-temporal hierarchical hypergraph online learning for crime prediction. In *International Conference on Learning Representations*, volume 2026, pp. 143229–143250, 2026.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

Levi Harris and Tianlong Chen. A space-time transformer for precipitation nowcasting. *arXiv preprint arXiv:2511.11090*, 2025.

Yana Hasson, Pauline Luc, Liliane Momeni, Maks Ovsjanikov, Guillaume Le Moing, Alina Kuznetsova, Ira Ktena, Jennifer J Sun, Skanda Koppula, Dilara Gokay, et al. Scivid: Cross-domain evaluation of video models in scientific applications. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21800–21811. IEEE, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Rui He, Yunpeng Xiao, Xingyu Lu, Song Zhang, and Yanbing Liu. St-3dgmr: Spatio-temporal 3d grouped multiscale resnet network for region-based urban traffic flow prediction. *Information Sciences*, 624:68–93, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V Chawla. Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pp. 1423–1432, 2018.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5351–5362, 2024.

Zhonghang Li, Long Xia, Lei Shi, Yong Xu, Dawei Yin, and Chao Huang. Open spatio-temporal foundation models for traffic prediction. *ACM Transactions on Intelligent Systems and Technology*, 2025.

Yuxuan Liang, Yutong Xia, Songyu Ke, Yiwei Wang, Qingsong Wen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. Airformer: Predicting nationwide air quality in china with transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 14329–14337, 2023.

Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, Bin Yang, et al. Air quality prediction with physics-guided dual neural odes in open systems. In *International Conference on Learning Representations*, volume 2025, pp. 55132–55154, 2025.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024.

Allister Loder, Lukas Ambühl, Monica Menendez, and Kay W Axhausen. Understanding traffic capacity of urban networks. *Scientific reports*, 9(1):16283, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Tung Nguyen, Arsh Koneru, Shufan Li, and Aditya Grover. Physix: A foundation model for physics simulations. *arXiv preprint arXiv:2506.17774*, 2025.

Dmitry Pavlyuk. Transfer learning: Video prediction and spatiotemporal urban traffic forecasting. *Algorithms*, 13(2):39, 2020.

Janne Perini, Rafael Bischof, Moab Arar, Ayça Duran, Michael A Kraus, Siddhartha Mishra, and Bernd Bickel. Pretrained video models as differentiable physics simulators for urban wind flows. *arXiv preprint arXiv:2603.21210*, 2026.

Bernardo Perrone Ribeiro and Jana Faganeli Pucer. Flowcast: Advancing precipitation nowcasting with conditional flow matching. In *International Conference on Learning Representations*, volume 2026, pp. 156863–156883, 2026.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in neural information processing systems*, 33:22009–22019, 2020.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560. IEEE, 2023a.

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6312–6322, 2023b.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Binqing Wu, Zongjiang Shang, Shiyu Liu, Jianlong Huang, Jiahui Xu, and Ling Chen. Airdde: Multifactor neural delay differential equations for air quality forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15941–15949, 2026a.

Kaiqi Wu, Weiyang Kong, Sen Zhang, Zitong Chen, and Yubao Liu. Efficient traffic forecasting on large-scale road network by regularized adaptive graph convolution. In *2026 IEEE 42nd International Conference on Data Engineering (ICDE)*, pp. 835–848. IEEE, 2026b.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4095–4106, 2024.

Yuan Yuan, Jingtao Ding, Chonghua Han, Zhi Sheng, Depeng Jin, and Yong Li. Uniflow: a foundation model for unified urban spatio-temporal flow prediction. *IEEE Transactions on Mobile Computing*, 2026.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*, 2024.

Tong Zhao, Junping Du, Zhe Xue, Meiyu Liang, Aijing Li, Xiaolong Meng, and Dandan Liu. St-vlm: A spatial-to-image multimodal spatial-temporal prediction framework with vision-language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16441–16449, 2026.

Yunlong Zhou, Chen Zhao, Danyang Peng, Fanfan Ji, and Xiao-Tong Yuan. Learning to refine: Spectral-decoupled iterative refinement framework for precipitation nowcasting. *arXiv preprint arXiv:2606.02661*, 2026.

Yiqing Zou, Hanning Yuan, Qianyu Yang, Ziqiang Yuan, Shuliang Wang, and Sijie Ruan. Meta dynamic graph for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16584–16592, 2026.

A Data and Domain Renderers

Data sources and protocols. Table A1 supplements Table 1 with dataset provenance, preprocessing, and renderer mappings. All domains use chronological splits rather than random window assignment. Traffic is partitioned by timestamp, and crime and air quality by target date or hour. Weather's primary metric pools all 12 forecast leads. Source identifiers, licence records, and data sizes are listed in Table A1.

Missing observations and normalisation. Traffic and air quality retain missing observations through Boolean masks without temporal interpolation. Values are standardised using per-detector or per-station training statistics; missing inputs are encoded as zero alongside the mask, and missing targets are excluded from loss and evaluation. Weather requires $\texttt{pct_missing}= 0$ and uses the fixed encoding $\mathrm{VIL}/255$. For crime, training-set 99th-percentile clipping and $\log(1+x)$ apply only to inputs; targets retain their original counts. All fitted preprocessing statistics use training data only, and losses and metrics are computed in native target units.

Domain renderer mappings. The renderer uses a common value / auxiliary / structure interface with domain-specific channel definitions (Table A1). Traffic and air quality combine standardised values with observation masks and graph or station-identity information. Weather combines scaled VIL, a logarithmic transformation, and a thresholded intensity channel; crime combines category-specific counts, total counts, and a training support mask. The air-quality 3×4 arrangement encodes station identity rather than geographic proximity. Canvases are resized to 224×224 using the channel-specific interpolation rules in the table. Frames retain native temporal sampling; the final weather frame is repeated once to complete a tubelet of size two. Figures A1 and A2 illustrate the channel mappings and spatial layouts.

Renderer ablations. Table A2 changes one renderer component at a time under the specialised training configurations, retaining the decoder and native evaluation protocol. Fixed RGB encoding and 2×2 linear projection worsen mean performance in all four domains. Increasing resolution to 384 pixels improves weather CSI-M by 0.8% but provides no mean improvement elsewhere. Temporal subsampling worsens all four scores, with the largest relative loss on weather (3.4%). Across the tested renderer variants, the specialised model retains better mean performance than the strongest baseline reported in each domain in Table 2, suggesting that its advantage is not specific to the default rendering scheme.

Table A1: Supplementary data protocols and renderer mappings. Task definitions, observation cadence, input/output lengths, sample counts, and anchor types are given in Table 1 and are omitted here. All fitted preprocessing statistics use training data only. Canvas sizes describe rendering before resizing to the backbone input and do not specify the evaluation grid.

	Traffic	Weather	Crime	Air quality
Provenance and temporal protocol				
Source reference	(Loder et al., 2019)	(Veillette et al., 2020)	NYC Open Data; dataset	(Chen, 2017)
Licence / access record	CC BY-NC-SA 4.0	None stated in the recorded AWS Open Data entry	NYC Open Data Law	CC BY 4.0
Recorded data size	6.99 GB	147.4 GB	32.8 MB	8.2 MB
Split basis	Chronological blocks over 3,359 timestamps	Event date; all windows from an event stay together	Target date	Target hour
Missing observations and input preprocessing				
Missing values	Retained with Boolean masks; no interpolation	Require `pct_missing`$=0$	No missing count entries	Retained with Boolean masks; no interpolation
Input transform	Per-detector training-set z-score	Fixed $\mathrm{VIL}/255$	Training-set q_{99} clipping and $\log(1+x)$; inputs only	Per-station training-set z-score
Renderer mapping				
Canvas	36×36; 34×34 interior	128×128	18×17, padded to 18×18	3×4 station-name grid; non-geographic
Channel 0: value	Standardised flow; zero at missing entries	$v/255$	Log-transformed count for the selected category	Standardised $\mathrm{PM}_{2.5}$; zero at missing entries
Channel 1: auxiliary	Observation / detector mask	$\log(1+v)/\log 256$	Log-transformed total count across four categories	Observation / station mask
Channel 2: structure	Detector-graph raster; eight nearest neighbours by haversine distance and self-loops	$\mathrm{clip}_{[0,1]}((v-74)/181)$	Training-set support mask	Fixed station-identity surface
Spatial resizing	Bilinear for values and graph; nearest neighbour for masks	Bilinear	Nearest neighbour	Bilinear for values and identity; nearest neighbour for masks
Backbone input	224×224	224×224	224×224	224×224
Runtime frames	20	14; repeat the last frame once	30	24

v denotes the original VIL code. $\mathrm{clip}_{[0,1]}$ truncates values to $[0, 1]$. Runtime frame counts include temporal padding for tubelet size two. Input transformations do not alter native target units.

Table A2: Sensitivity to renderer choices under specialised training. Results are test-set mean ± standard deviation over three seeds. Each non-default row changes one renderer component; the domain-specific adaptation recipe, decoder configuration, and native evaluation protocol are retained. Parentheses give relative improvement over the default mean: $100(m_0 - m)/m_0$ for MAE and $100(m - m_0)/m_0$ for CSI-M, where m_0 is the default score. Positive values indicate improvement. Bold marks the best mean in each column.

Renderer variant	Traffic MAE ↓	Weather CSI-M ↑	Crime MAE ↓	Air MAE ↓
Default channels, 224 px, native temporal sampling	**59.46** ± 0.95	0.4026 ± 0.001	**0.6054** ± 0.000	**41.39** ± 0.04
Fixed RGB colormap (WinDiNet-style)	61.30 ± 2.70 (−3.1%)	0.3941 ± 0.003 (−2.1%)	0.6090 ± 0.001 (−0.6%)	41.97 ± 0.12 (−1.4%)
2 × 2 linear projection (ST-VFM-style)	62.08 ± 3.10 (−4.4%)	0.3901 ± 0.004 (−3.1%)	0.6108 ± 0.002 (−0.9%)	42.22 ± 0.18 (−2.0%)
Resolution 384 px	59.70 ± 2.60 (−0.4%)	**0.4058** ± 0.002 (+0.8%)	**0.6054** ± 0.001 (0.0%)	41.51 ± 0.06 (−0.3%)
Temporal subsampling ×2	60.89 ± 2.80 (−2.4%)	0.3889 ± 0.003 (−3.4%)	0.6066 ± 0.001 (−0.2%)	42.05 ± 0.14 (−1.6%)

Default channels denote the value / auxiliary / structure mapping in Table A1. The default row reproduces the specialised results in Table 2. A displayed standard deviation of 0.000 reflects rounding and does not imply identical results across seeds.

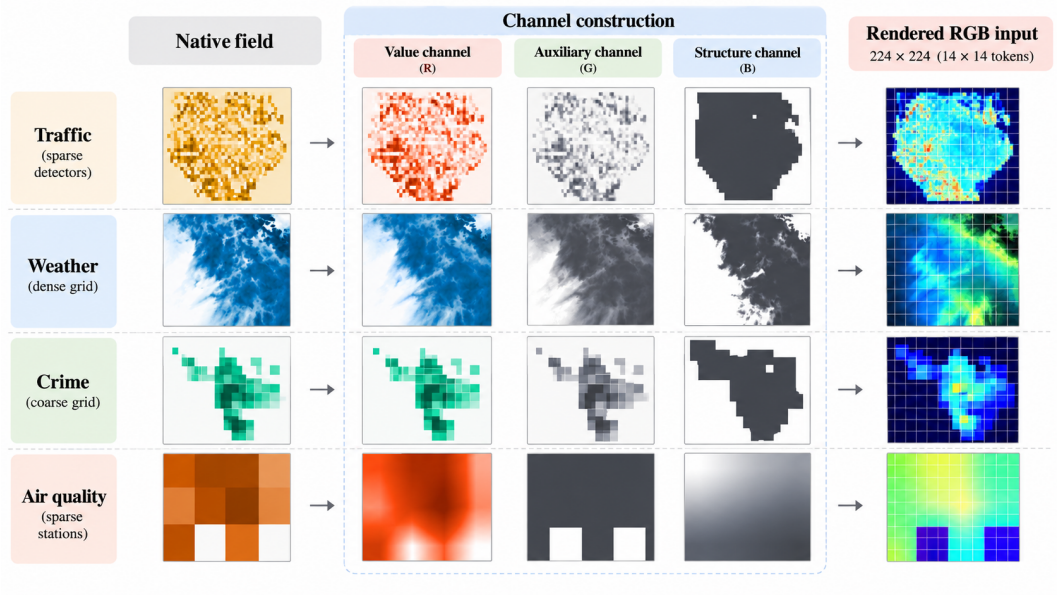


Figure A1: Domain renderer examples at the final observed history step. Each row shows the native input, the value, auxiliary, and structure channels, and their 224 × 224 composite. The overlaid 14 × 14 grid indicates spatial token boundaries for patch size 16.

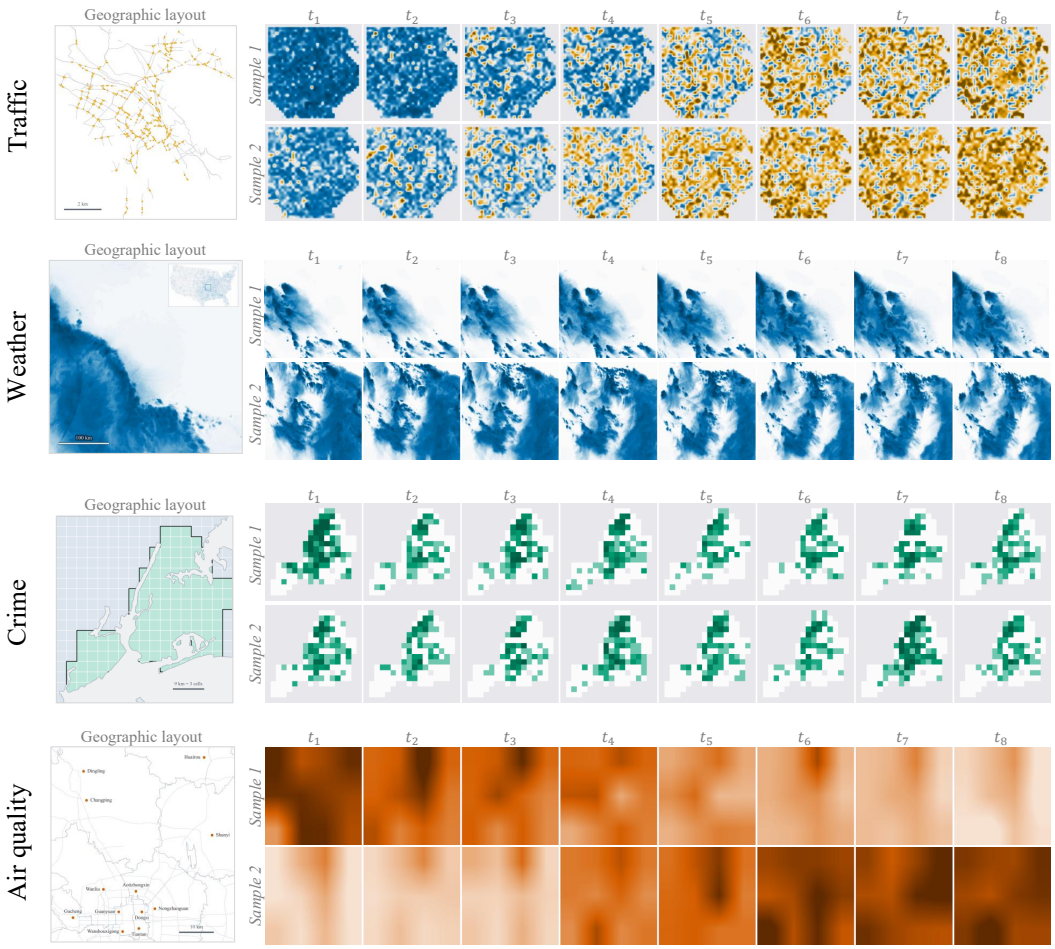


Figure A2: The four domains in space and time. Left: the geometry each renderer works on—1,020 Zurich detectors over the UTD19 road links; one 128×128 SEVIR VIL patch with its location among all training events; the 18×17 grid of 3 km cells clipped to the New York shoreline; the 12 Beijing stations at approximate public positions. Right: two training clips per domain, shown as the value channel of the rendered pseudo-video at eight steps of the history window, one fixed colour scale per strip; grey marks cells outside the support mask.

B Implementation Details and Computational Cost

Variant definitions. VFM4ST variants differ in backbone initialisation, adaptation, training domains, and parameter sharing (Table B1). Shared training updates one backbone across all four domains, whereas specialised training adapts a separate backbone to each domain. PEFT and frozen variants retain common pretrained weights with domain-specific adapters or decoders; leave-one-out training evaluates transfer from three domains to a fourth. Comparisons with different adaptation strategies or budgets assess complete configurations rather than a single experimental factor.

Backbone and domain decoders. All variants use the VideoMAE V2 ViT-B architecture, with fixed sinusoidal position encodings regenerated for the runtime temporal token count. Table B2 specifies tensor shapes, backbone adaptation, and the anchor, dynamics, and readout configurations. Weather combines advection with a native-history bypass; traffic and air quality augment station queries with local and neighbouring histories.Parameter counts include the backbone and domain-specific prediction modules.

Optimisation and training budget. Training uses AdamW with a cosine learning-rate schedule. Table B3 reports learning rates, batch sizes, update counts, and computational costs. Shared training balances domain update frequency, while leave-one-out training separates joint backbone adaptation from held-out decoder fitting. Frozen and held-out decoders use the domain-specific fitting schedules listed in the table. As in the main evaluation, learned-model results are summarised by the mean and standard deviation of test scores over three seeds.

Baseline methods. The 18 learned baselines in Table 2 fall into five groups. (i) *Domain-specific models* represent the most recent task-specific designs, published in 2025–2026. We choose these methods instead of earlier standard baselines because each of them reports consistent improvements over established classical models in its domain under its original evaluation protocol. Together with the statistical references (persistence, periodic/climatological means, and linear autoregression), they bound performance from both the classical and the state-of-the-art ends. For traffic, MetaDG (Zou et al., 2026) models time-varying road-network dependencies through meta-learned dynamic graphs, and RAGC (Wu et al., 2026b) regularises adaptive graph convolution for efficient forecasting on large-scale road networks. For radar nowcasting, FlowCast (Perrone Ribeiro & Faganeli Pucer, 2026) generates future radar fields through conditional flow matching in the latent space of a pretrained VAE, and SDIR (Zhou et al., 2026) iteratively refines forecasts with spectrally decoupled components. For crime, ST-HHOL (Du et al., 2026) performs online learning over hierarchical spatio-temporal hypergraphs with a frozen language model. For air quality, AirDDE (Wu et al., 2026a) models delayed multifactor pollutant propagation with neural delay differential equations, and Air-DualODE (Liang et al., 2025) couples physics-guided and data-driven neural ODEs for open atmospheric systems. (ii) *Urban foundation models*, UniST (Yuan et al., 2024), OpenCity (Li et al., 2025), UrbanGPT (Li et al., 2024), and UniFlow (Yuan et al., 2026), learn transferable representations from urban spatio-temporal data; UrbanGPT additionally aligns spatio-temporal encodings with a large language model. (iii) *Video-style architectures trained from scratch* treat field sequences as videos without pretraining: ST-3DGMR (He et al., 2023) applies grouped multiscale 3D convolutions to region-based flow maps, and SaTformer (Harris & Chen, 2025) applies space-time attention to precipitation nowcasting. (iv) *Time-series foundation models*, Chronos (Ansari et al., 2024), TimesFM (Das et al., 2023), and Moirai (Woo et al., 2024), are pretrained on large time-series corpora and applied to each location independently. (v) *Image foundation models* transfer visual pretraining to numerical forecasting: VisionTS (Chen et al., 2024) casts univariate forecasting as masked image reconstruction, and ST-VLM (Zhao et al., 2026) converts cross-location correlations into images for a pretrained vision-language model. Groups (iii)–(v) form the cross-modal transfer block of Table 2.

Baseline integration. The 18 learned baselines are trained separately under each evaluated domain's data and forecasting protocol, with hyperparameters selected on the validation set. Graph models use four-neighbour grid graphs on dense domains; grid models use the same station canvases as VFM4ST on sparse domains and read predictions at station cells. Time-series foundation baselines process each station or grid-cell variable as a univariate series and are fine-tuned in all four domains. VisionTS renders each univariate series as an image, whereas ST-VLM uses correlation images across observation locations. Table B4 reports model sizes and training costs for representative baselines. The parameter count of ST-HHOL excludes its frozen LLM.

Table B1: VFM4ST variants. Columns specify initialisation, backbone training, training data, backbone storage, and evaluation protocol. Domain-specific decoders are trained in every variant.

Variant	Backbone init.	Backbone training	Training data	Backbones	Test
Shared	Video pretrained	Full fine-tuning; domain token	Four domains; balanced sampling	1	Each domain; shared backbone, domain-specific decoder
Specialised	Video pretrained	T: LoRA $r = 16$; W: full FT; C: last 6 blocks; A: LoRA $r = 32$	Own domain	4	Own domain
PEFT	Video pretrained	Frozen base; domain-specific LoRA	Own domain	1 + 4 adapter sets	Own domain
Frozen	Video pretrained	Frozen; decoder only	Own domain	1	Own domain
Scratch	Random	Full fine-tuning	Own domain	4	Own domain
Image-pretrained	Image MAE, inflated to video	Full fine-tuning	Own domain	4	Own domain
Leave-one-out	Video pretrained	Full FT on three domains; freeze backbone for held-out decoder fitting	Three domains, then the fourth	4 folds	Held-out domain

T / W / C / A denote traffic / weather / crime / air quality throughout Tables B1–B4. FT denotes fine-tuning. Backbone counts exclude domain-specific decoders; PEFT stores one base checkpoint and four adapter sets, and leave-one-out training produces one adapted backbone per fold. Adapter ranks and training budgets are given in Table B3.

Table B2: Backbone, tensor shapes, and domain decoder configurations for the specialised variant unless stated otherwise. Batch dimensions are omitted; video inputs use channel–time–height–width order.

	Traffic	Weather	Crime	Air quality
Backbone and feature decoding				
Backbone	VideoMAE V2 ViT-B; 12 blocks, width 768, patch size 16, tubelet size 2; 224×224 inputs, 14×14 spatial tokens. Video-pretrained variants use UnlabeledHybrid-1M weights.			
Position encoding	Fixed sinusoidal encoding regenerated for the runtime temporal token count $T' = T_{\text{frames}}/2$.			
Domain conditioning	Shared only: a learned domain vector $e_d \in \mathbb{R}^{768}$ is added to every input token.			
Decoder trunk	Concatenate temporal features to width $768T'$; apply a 1×1 projection to width 256 and GroupNorm. Each upsampling stage uses $2\times$ bilinear upsampling and a 3×3 convolution, halving width to a minimum of 32.			
Upsampling / resize	1 stage: $14 \to 28$; resize to 36×36	4 stages: $14 \to 224$; resize to 128×128	0 stages; resize to 18×18	0 stages; resize to 3×4
Tensor shapes and adaptation				
Native history x_d	$20 \times 1{,}020$; mask	$13 \times 128 \times 128$	$30 \times 18 \times 17 \times 4$	24×12; mask
Video input v_d	$3 \times 20 \times 224 \times 224$	$3 \times 14 \times 224 \times 224$	$3 \times 30 \times 224 \times 224$	$3 \times 24 \times 224 \times 224$
Features z_d	$10 \times 14 \times 14 \times 768$	$7 \times 14 \times 14 \times 768$	$15 \times 14 \times 14 \times 768$	$12 \times 14 \times 14 \times 768$
Displacement f_d	–	$12 \times 2 \times 128 \times 128$	–	–
Prediction $\hat{y}_d$	$20 \times 1{,}020$	$12 \times 128 \times 128$	$1 \times 18 \times 17 \times 4$	24×12
Backbone adaptation	LoRA $r = 16$; attention QKV and MLP	Full fine-tuning	Unfreeze last 6 blocks	LoRA $r = 32$
Anchor, dynamics, and readout				
Anchor	Learned static field	Last frame	Climatology	Learned static field
Warp	Identity	Backward sampling	Identity	Identity
Native bypass s_d	–	Native 128×128 history	–	–
Correction scale	Training flow std.	VIL code units	Training count std.	Training $PM_{2.5}$ std.
Readout	Station query; rank 256	Dense 128×128 grid	Dense $18 \times 17 \times 4$ grid	Station query; rank 256
Readout history	Own + 8 correlated neighbours; $h = 256$, 20 steps	–	–	Own + 8 correlated neighbours; $h = 256$, 24 steps
Direct linear branch	Per-node; zero-initialised	–	–	Per-node; zero-initialised
Parameters (M), total / trainable	99.6 / 13.3	88.2 / 88.2	89.2 / 45.5	92.8 / 6.5

Table B3: Hyperparameters and training budgets. Slash-separated entries follow T / W / C / A order. Steps denote optimiser updates; GPU-h denotes training compute.

Variant	Domain	Adaptation	LR	Batch	Epochs	Steps	GPU-h
Specialised	T	LoRA $r = 16$	3×10^{-5}	4	10	4,550	0.51
Specialised	W	Full FT	3×10^{-5}	8	10	40,190	8.66
Specialised	C	Last 6 blocks	3×10^{-5}	4	10	1,450	0.24
Specialised	A	LoRA $r = 32$	6×10^{-5}	6	1	4,137	0.89
Shared	All	Full FT; domain token; balanced sampling	3×10^{-5}	4 / 8 / 4 / 6	8	142,880	31.3
PEFT	Each	LoRA $r = 16$ / 32 / 32 / 32	3×10^{-5}; A: 6×10^{-5}	4 / 8 / 4 / 6	10 / 10 / 10 / 1	4,550 / 40,190 / 1,450 / 4,137	0.5 / 6.1 / 0.6 / 0.9
Frozen	Each	Frozen backbone; decoder only	3×10^{-5}; A: 6×10^{-5}	4 / 8 / 4 / 6	64 / 16 / 1 / 1	29,120 / 71,440 / 145 / 4,137	1.0 / 9.1 / 0.2 / 0.4
Scratch / Image-pretrained	Each	Full FT	3×10^{-5}; A: 6×10^{-5}	4 / 8 / 4 / 6	10 / 10 / 10 / 1	4,550 / 40,190 / 1,450 / 4,137	1.0 / 8.7 / 0.3 / 0.9
Leave-one-out	Three to one	Full FT on three domains; frozen backbone for held-out decoder	3×10^{-5}	4 / 8 / 4 / 6	$8 + P^*$	107,160 + held-out fitting	37.4 / 31.4 / 31.6 / 32.6

Each leave-one-out fold uses $8 \times 3 \times 4{,}465 = 107{,}160$ backbone updates, followed by $P^* = 64/16/1/1$ held-out decoder epochs (29,120 / 71,440 / 145 / 4,137 updates).

Table B4: Representative baselines: evaluated domains, parameter counts, and one-seed training costs on an RTX 4090. Multi-domain costs follow T / W / C / A order, with the four-domain total in parentheses.

Baseline	Domain	Params (M)	Train GPU-h
Domain-specific models			
MetaDG	Traffic	0.674	4.2
FlowCast	Weather	160.4	26.1
ST-HHOL	Crime	0.15	0.15
Air-DualODE	Air quality	0.105	2.9
Representatives of other baseline groups			
UniST (urban foundation model)	All four	24.8	0.4 / 5.8 / 0.2 / 0.6 (7.0)
ST-3DGMR (video-style, from scratch)	All four	2.1	0.2 / 4.6 / 0.1 / 0.3 (5.2)
Chronos (time-series foundation model)	All four	212	2.1 / 9.6 / 0.3 / 0.8 (12.8)

C Backbone Scale, Pretraining, Modality, and Joint Initialisation

Backbone scale, pretraining, and modality.

We retain the specialised decoder and domain-specific adaptation recipes in Table B1 and the three-seed test protocol of Table 2, replacing the backbone architecture or pretrained weights (Table C1). VideoMAE V2 Base is the specialised model in Table 2; the frozen, Image MAE, Scratch, and Direct numeric rows provide the initialisation comparison in Section 4.3. Scaling from 87M to 304M and 1012M parameters yields no consistent improvement: weather CSI-M increases by 0.9% and 0.6%, traffic MAE worsens by 2.5% and 4.9%, and crime and air-quality changes remain below 0.8%. Macro changes are -0.5% and -1.4%. Base-selected hyperparameters are retained without retuning; Giant also changes the patch grid and pretraining duration. Different reliance on backbone features and optimisation under a fixed budget may explain the domain-dependent response, but the comparison does not isolate capacity alone. Among alternative video checkpoints, VideoMAE V1 (Tong et al., 2022), TimeSformer (Bertasius et al., 2021), and V-JEPA 2 (Assran et al., 2025) all underperform V2 Base in every domain. TimeSformer is closest on traffic, whereas V1 is closest on weather, crime, and air quality; the 325M-parameter V-JEPA 2 does not recover the gap. Inflated image-MAE weights (He et al., 2022) yield losses of 3.0–7.7%, and frame-wise use of V2 yields losses of 3.4–12.0%, supporting the value of joint spatio-temporal encoding. However, checkpoint comparisons also differ in pretraining data or architecture, and the reused image-pretrained and scratch references use full fine-tuning rather than domain-selected adaptation.

Initialisation under joint training. We compare video-pretrained, image-pretrained, and randomly initialised backbones under the shared recipe with domain conditioning, balanced sampling, and eight training epochs (Table C2). Relative to the video-pretrained shared model, scratch training degrades traffic, weather, crime, and air-quality performance by 10.6%, 12.1%, 7.6%, and 6.4%, respectively, compared with 8.2%, 8.7%, 6.7%, and 6.1% under own-domain training. Image pretraining reduces the macro deficit from 9.2% to 6.4%. The frozen reference reuses the frozen-backbone, domain-specific decoder results in Table C1, with the original decoder-fitting budgets rather than the eight-epoch joint schedule. Across all four domains, video-pretrained adaptation performs best, followed by frozen video features, image pretraining, and scratch training, matching the own-domain ordering. The larger scratch deficits are consistent with a greater dependence on pretrained representations when learning several domains together, although differences between the specialised and shared training configurations prevent attribution solely to joint learning.

D Leave-one-domain-out

Frozen-backbone protocol. We compare four backbone states: random initialisation (rand), the original video-pretrained weights (θ_0), joint adaptation on the three non-target domains (lodo3), and joint adaptation on all four domains (joint4). Each backbone is frozen before fitting a fresh target-domain decoder, using the same domain-specific configuration and fitting budget across arms. The decoder-fitting schedules are $P^* = 64/16/1/1$ epochs for traffic, weather, crime, and air quality, respectively (Table B3). The θ_0 scores reuse the frozen row of Table C1; the θ_0 and lodo3 arms provide the comparison in Section 4.4. Joint4 includes the target domain during backbone adaptation and serves as a target-exposed reference rather than a held-out transfer result.

Transfer and reference comparisons. Adaptation on the other three domains provides no mean improvement over θ_0: traffic and air-quality MAE increase by 1.5% and 5.2%, weather CSI-M decreases by 2.5%, and crime MAE is unchanged at the reported precision (Table D1). Joint4 improves all four means by 2.8%, 2.6%, 1.1%, and 1.3%, respectively, with scores between θ_0 and the shared model in Table 2. For weather, the shared model improves CSI-M by 3.9% over θ_0. The positive joint4 differences show that the probe can detect benefits from target-exposed adaptation, although no benefit from held-out-domain transfer is observed. Relative to θ_0, rand performs worse by 4.7% on traffic and 10.6% on weather, compared with 1.0% on crime and 1.2% on air quality. The larger traffic and weather gaps are consistent with greater reliance on pretrained video features; the comparison concerns frozen feature initialisation, distinct from the video-pathway ablation in Section 4.3.

Table C1: Backbone initialisation, scale, pretraining, and modality under own-domain training. Results are test-set mean ± standard deviation over three seeds. Rows retain the specialised decoder and domain-specific adaptation recipes in Table B1 unless the notes state otherwise. Negative values indicate worse performance. Macro Δ averages the four unrounded relative changes. Bold marks the best mean in each domain.

Backbone	Params	Pretraining	Traffic MAE ↓	Weather CSI-M ↑	Crime MAE ↓	Air MAE ↓	Macro Δ
Video-pretrained reference and frozen backbone							
VideoMAE V2 Base	87M	Masked video; UH-1M	**59.46**±0.95	0.4026±0.001	**0.6054**±0.000	**41.39**±0.04	–
VideoMAE V2 Base, frozen	87M	Masked video; UH-1M	63.50±1.54 (-6.8%)	0.3858±0.003 (-4.2%)	0.6160±0.001 (-1.8%)	42.87±0.15 (-3.6%)	-4.1%
Scale: same family, objective, and corpus							
VideoMAE V2 Large	304M	Masked video; UH-1M	60.96±2.60 (-2.5%)	**0.4062**±0.002 (+0.9%)	0.6064±0.001 (-0.2%)	41.53±0.20 (-0.3%)	-0.5%
VideoMAE V2 Giant	1012M	Masked video; UH-1M	62.40±2.80 (-4.9%)	0.4051±0.004 (+0.6%)	0.6084±0.002 (-0.5%)	41.68±0.35 (-0.7%)	-1.4%
Video pretraining: objectives and checkpoints							
VideoMAE V1	87M	Masked video; K400	63.66±2.54 (-7.1%)	0.3925±0.004 (-2.5%)	0.6085±0.002 (-0.5%)	42.01±0.31 (-1.5%)	-2.9%
TimeSformer	121M	Supervised video; K400	63.03±2.30 (-6.0%)	0.3805±0.006 (-5.5%)	0.6126±0.004 (-1.2%)	42.47±0.57 (-2.6%)	-3.8%
V-JEPA 2 ViT-L	325M	Latent prediction; video	63.36±2.60 (-6.6%)	0.3862±0.005 (-4.1%)	0.6103±0.003 (-0.8%)	42.30±0.40 (-2.2%)	-3.4%
Modality and temporal encoding							
Image MAE, inflated	87M	Masked image; IN-1k	64.04±1.87 (-7.7%)	0.3741±0.008 (-7.1%)	0.6233±0.009 (-3.0%)	43.78±1.12 (-5.8%)	-5.9%
Frame-wise V2	87M	Masked video; UH-1M	66.58±2.70 (-12.0%)	0.3628±0.007 (-9.9%)	0.6261±0.008 (-3.4%)	44.03±1.00 (-6.4%)	-7.9%
Random initialisation and reference models							
Scratch	87M	None	64.31±2.12 (-8.2%)	0.3675±0.003 (-8.7%)	0.6461±0.004 (-6.7%)	43.90±1.07 (-6.1%)	-7.4%
Direct numeric (matched MLP)	matched	None; native input	73.09±2.98 (-22.9%)	0.2921±0.004 (-27.4%)	0.6504±0.006 (-7.4%)	43.93±0.58 (-6.1%)	-16.0%
Strongest baseline	–	Varies by domain	62.29±0.77 (-4.8%)	0.3734±0.005 (-7.3%)	0.6218±0.009 (-2.7%)	42.99±1.27 (-3.9%)	-4.6%

Parameter counts refer to backbones. UH-1M denotes UnlabeledHybrid-1M; K400 denotes Kinetics-400; IN-1k denotes ImageNet-1k. Scale variants retain Base-selected learning rates, LoRA ranks, and update budgets without layer-wise learning-rate decay. Base and Large use patch size 16 and a 14×14 spatial token grid; Giant uses patch size 14 and a 16×16 grid. Pretraining lasts 800 epochs for Base/Large and 1200 for Giant. V1 and TimeSformer use width 768 and 12 blocks; V-JEPA 2 uses ViT-L. Frame-wise V2 encodes frames independently without temporal attention. Frozen fixes the pretrained backbone and trains domain-specific decoders for the schedules in Table B3; Image MAE and Scratch use full fine-tuning rather than domain-selected adaptation. Direct numeric replaces the rendered video and backbone with a parameter-matched MLP on native inputs, retaining the anchor and residual parameterisation of the domain decoder.

Table C2: Initialisation under joint four-domain training. Changes are relative to the video-pretrained shared model in Table 2, using the convention in Table C1.

Domain / metric	Video-pretrained (shared)	Scratch	Frozen (reused)	Image-pretrained
Traffic MAE ↓	**60.75**±0.36	67.16±2.67 (-10.6%)	63.50±1.54 (-4.5%)	65.97±2.30 (-8.6%)
Weather CSI-M ↑	**0.4010**±0.006	0.3524±0.006 (-12.1%)	0.3858±0.003 (-3.8%)	0.3674±0.008 (-8.4%)
Crime MAE ↓	**0.6059**±0.000	0.6518±0.006 (-7.6%)	0.6160±0.001 (-1.7%)	0.6276±0.009 (-3.6%)
Air MAE ↓	**42.02**±0.17	44.71±1.13 (-6.4%)	42.87±0.15 (-2.0%)	44.18±1.02 (-5.1%)
Macro Δ	–	-9.2%	-3.0%	-6.4%

Trainable backbones use domain conditioning, balanced sampling, and eight epochs. Frozen reuses the frozen row of Table C1.

Table D1: Leave-one-domain-out comparison with four frozen-backbone arms. A fresh target-domain decoder is fitted for P^* epochs on each backbone. Changes are relative to θ_0: $100(m_0 - m)/m_0$ for MAE and $100(m - m_0)/m_0$ for CSI-M, where m_0 is the θ_0 mean. Positive values indicate improvement.

Domain / metric	P^* (epochs)	rand	θ_0 (Frozen)	lodo3	joint4	lodo3 $-\theta_0$	joint4 $-\theta_0$
Traffic MAE ↓	64	66.5	63.50	64.45	61.7	-1.5%	$+2.8\%$
Weather CSI-M ↑	16	0.345	0.3858	0.3762	0.396	-2.5%	$+2.6\%$
Crime MAE ↓	1	0.622	0.6160	0.6160	0.6093	0.0%	$+1.1\%$
Air MAE ↓	1	43.40	42.87	45.10	42.29	-5.2%	$+1.3\%$

Rand uses frozen random weights; θ_0 uses the original frozen VFM; lodo3 excludes the target domain during backbone adaptation; joint4 includes it. Changes are calculated from the displayed means, so the crime lodo3 difference is zero at the reported precision.

E ANCHOR CHOICE BY DOMAIN

Anchor comparison. We compare the last observation, a training-set climatology, and a learnable static field under the specialised domain-specific adaptation recipes (Table E1). The anchor for each domain is selected by validation performance; the table reports the corresponding test scores. The anchor is varied while retaining the remaining domain-specific decoder configuration. The final two rows evaluate the last-frame and climatological anchors as standalone forecasts, without learned dynamics or readout corrections. The selected anchors reproduce Table 2: the last frame for weather, climatology for crime, and a learnable static field for traffic and air quality.

Domain-specific effects. The selected anchor gives the best mean forecast in each domain. Replacing the last-frame anchor with climatology or a static field lowers weather CSI-M by 16.7% and 10.4%, respectively. On crime, the last-frame and static alternatives increase MAE by 3.8% and 1.3% relative to climatology. Last-frame anchoring also increases MAE by 4.2% on traffic and 1.1% on air quality relative to the learned static reference. The pattern is consistent with preserving evolving echo structure in weather and recurring spatial intensity in crime, while allowing history-conditioned station readout to recover temporal variation around a static reference in traffic and air quality.

Table E1: Anchor choice by domain under specialised training. Bold marks the anchor used in Table 2. Parentheses give relative improvement over the bold mean in each column: $100(m_0 - m)/m_0$ for MAE and $100(m - m_0)/m_0$ for CSI-M. Positive values indicate improvement.

Anchor	Traffic MAE ↓	Weather CSI-M ↑	Crime MAE ↓	Air MAE ↓
Last frame	61.97 (-4.2%)	**0.4026**$\pm$0.001	0.6282 (-3.8%)	41.86 (-1.1%)
Climatology	63.02 (-6.0%)	0.3354 (-16.7%)	**0.6054**$\pm$0.000	42.31 (-2.2%)
Learnable static	**59.46**$\pm$0.95	0.3608 (-10.4%)	0.6131 (-1.3%)	**41.39**$\pm$0.04
Reference: anchor-only forecasts				
Last frame (persistence)	80.90	0.2427	0.8395	48.52
Climatology	109.69	0.0451	0.6577	56.13

Anchor-only references are evaluated at the native target locations. Standard deviations are shown where available; the crime deviation of 0.000 reflects rounding, not zero seed variation.