Thermo-JEPA: Learning a Geometry-Grounded Thermal World Model via Cross-Modal Privileged Masking

Biwen Yang[1], Jin Zhang[1], Zhe Cao[1], and Ruiheng Zhang[1]*

Beijing Institute of Technology, Beijing, China

Abstract. World models have driven remarkable progress in physical perception and dynamic forecasting, yet current mainstream implementations predominantly rely on visible-light (RGB) observations, which limits their all-weather reliability. Developing a thermal world model is essential, but it is hindered by the unique distribution of thermal imagery and the extreme scarcity of continuous video datasets. In this paper, we identify a fundamental challenge in extending standard masked modeling to the thermal domain: intra-frame spatial homogenization. Due to thermal equilibrium, distinct physical entities often exhibit near-zero temperature gradients. Consequently, unconstrained self-attention blurs semantic boundaries and collapses object topologies, severely degrading downstream dynamics forecasting. To overcome this, we propose Thermo-JEPA, a geometry-grounded pre-training framework based on Learning Using Privileged Information (LUPI). We extract intra-frame spatial topologies from aligned RGB data and inject them as a confidence-gated structural prior into the thermal teacher's attention mechanism. This explicitly enforces semantic segregation without corrupting fine-grained intra-object thermal gradients. Concurrently, we release RGBT-World, the largest unified, spatiotemporally aligned RGB-Thermal video corpus tailored for generative pre-training. Extensive experiments demonstrate that Thermo-JEPA effectively prevents representation collapse, achieving state-of-the-art zero-shot dynamics alignment on pure thermal inputs and significantly outperforming massive generic video foundation models.

Keywords: World Models · Thermal Vision · Self-Supervised Learning · Cross-Modal Alignment · Zero-Shot Transfer

1 Introduction

Self-supervised Video Foundation Models (VFMs) and world models, particularly those based on the Joint-Embedding Predictive Architecture [2, 3, 15, 26], have demonstrated remarkable success in learning spatiotemporal dynamics from vast amounts of visible-light (RGB) data [1, 7, 8]. However, for safety-critical applications such as autonomous driving and robotic navigation, RGB sensors are

* Corresponding author. ruiheng.zhang@bit.edu.cn

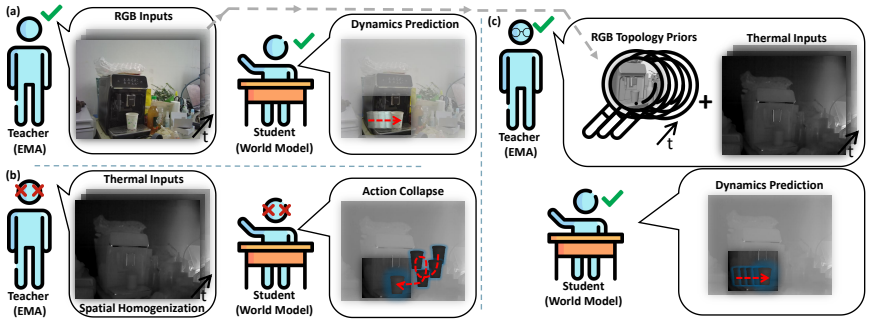

Fig. 1: Intra-frame spatial homogenization in thermal world modeling. While RGB provides clear boundaries (a), thermal equilibrium blends adjacent entities, causing standard models to suffer from action collapse (b). Thermo-JEPA overcomes this by injecting RGB spatial topology as a privileged prior into the teacher, enabling sharp pure-thermal kinematic predictions (c).

inherently brittle under adverse illumination. Extending generative world modeling to the thermal infrared (IR) domain—which captures emitted heat rather than reflected light—is an essential step toward robust [11], all-weather physical perception.

Despite the architecturally unified nature of Vision Transformers (ViTs), directly applying standard spatiotemporal masked modeling to thermal videos yields severely degraded representations. We formalize this core bottleneck as intra-frame spatial homogenization. As illustrated in Fig. 1, distinct physical entities (*e.g.*, a coffee machine, paper cups, water cups, and snack packages) often exist in near-thermal equilibrium. While RGB provides sharp structural boundaries, adjacent entities in an IR frame frequently exhibit near-zero pixel gradients. In unconstrained self-attention mechanisms, this lack of high-frequency spatial variation intrinsically drives the network to homogenize adjacent patches. When the geometric boundaries of physical entities collapse into continuous feature distributions, the model loses the structural prerequisites to track object trajectories, causing severe representation collapse in downstream temporal dynamics tasks.

Previous efforts predominantly rely on dense cross-modal alignment or early fusion architectures [13, 30, 35]. However, forcing the thermal latent space to directly mimic RGB distributions inevitably triggers modal pollution. The network is compelled to hallucinate high-frequency visible-light textures that physically do not exist in the thermal spectrum, thereby distorting the learning of intrinsic thermal kinematics [9, 46]. Moreover, fusion-based paradigms compromise the deployment autonomy of pure-IR systems.

To fundamentally resolve the spatial homogenization issue while preserving thermodynamic integrity, we propose **Thermo-JEPA**, a geometry-grounded world model pre-training framework strictly designed for thermal modalities. We formulate the pre-training process under the Learning Using Privileged In-

formation (LUPI) paradigm. Specifically, RGB data is utilized exclusively during pre-training to extract a structural spatial topology matrix. We inject this RGB-derived topological boundary as a gated structural prior into the Multi-Head Self-Attention (MHSA) mechanism of the exponential moving average (EMA) thermal teacher network. This mechanism explicitly segregates semantic entities without destroying intra-object thermal distribution. During inference, the student network relies entirely on pure thermal inputs, outputting dynamic features that can be decoded by action probes trained directly on thermal features.

Concurrently, to address the severe scarcity of continuous [20, 23], high-resolution thermal video data suitable for generative pre-training, we introduce **RGBT-World**, the largest unified, spatiotemporally aligned RGB-Thermal video corpus to date. RGBT-World aggregates 10 public benchmarks and manually realign visible light and infrared video in both time and space dimensions, providing robust multi-platform data diversity.

The primary contributions of this work are three-fold:

- We identify and formalize the Intra-frame Spatial Homogenization problem in thermal world modeling, demonstrating how thermal equilibrium collapses spatiotemporal attention and degrades temporal dynamic forecasting.
- We propose Thermo-JEPA, a geometry-grounded world model framework that leverages visible-light topology as a privileged soft prior. By injecting confidence-gated spatial boundaries into the EMA teacher's attention mechanism, our approach enforces semantic segregation within the thermal latent space while preserving intra-object thermodynamic gradients.
- We construct and release *RGBT-World*, the largest spatiotemporally aligned RGB-Thermal video corpus to date. By aggregating 10 public benchmarks and introducing a high-resolution self-collected subset, we provide the first unified platform tailored for large-scale generative thermal world modeling.

2 Related Work

2.1 World Models for Unsupervised Video Learning

World models aim to learn the predictive dynamics of the physical environment from high-dimensional sensory inputs [15, 26, 36]. Early approaches relied on pixel-level reconstruction using recurrent neural networks or variational autoencoders [16, 17, 32]. To mitigate the computational overhead and the distraction of unpredictable high-frequency visual noise, subsequent architectures transitioned to latent-space dynamics modeling, prominently featured in the Dreamer series [16, 17, 32] and continuous control frameworks (*e.g.* TD-MPC) [18].

Recently, self-supervised video representation learning has advanced significantly through masked modeling techniques (*e.g.* VideoMAE [37, 40], MaskFeat [42]) and diffusion-based predictive models [6, 19]. Among these, the Joint-Embedding Predictive Architecture (JEPA) [2, 3], establishes a rigorous self-supervised paradigm by predicting the latent representations of masked spatiotemporal patches using an exponential moving average (EMA) teacher. While

V-JEPA2 effectively abstracts macroscopic dynamics by minimizing L1 distance in the latent space, it is implicitly formulated for the visible light spectrum. When applied to thermal imaging, V-JEPA2 fails to account for thermal equilibrium. The standard unconstrained self-attention mechanism processes near-zero temperature gradients as homogeneous regions, erroneously smoothing the rigid topological boundaries of distinct physical entities. This intra-frame spatial homogenization degrades the latent structural integrity required for downstream dynamics modeling.

2.2 Cross-Modal Representation Learning

Learning aligned representations across multiple sensor modalities is a cornerstone of robust perception. Conventional cross-modal alignment heavily relies on contrastive learning paradigms (*e.g.* CLIP [35], VideoCLIP [45], ALIGN [22]), which maximize the cosine similarity between paired inter-modal embeddings. Other approaches extend Masked Autoencoders (MAE) to multi-modal inputs via early-fusion or cross-attention mechanisms (*e.g.* MultiMAE [4], OmniMAE [14]), forcing the network to reconstruct masked patches of one modality conditioning on another. Furthermore, knowledge distillation is frequently utilized to transfer rich representations from foundation RGB models to modality-specific encoders [13, 30, 48].

Crucially, these existing paradigms inherently presume a joint distribution, relying on dense feature matching. When bridging the distinct physical mechanisms of RGB and thermal sensors, such direct alignment incurs spurious feature projections [9, 46]. In contrast, we completely abandon dense feature distillation. Instead, we formulate our pre-training under the Learning Using Privileged Information (LUPI) framework [34, 39]. By extracting strictly topological adjacency—rather than high-frequency visual textures—we insulate the thermal encoder from gradient leakage and modal contamination.

2.3 Thermal Video Understanding and Datasets

Thermal infrared computer vision has been extensively studied to address the fragility of RGB sensors under adverse illumination [11, 20]. The majority of literature focuses on discriminative perception tasks, including thermal pedestrian tracking [28, 29, 50], object detection [12, 44], multispectral object tracking [49, 51], and thermal image enhancement or translation [5, 25, 52].

Despite these advancements, existing thermal video research exhibits two critical gaps regarding world modeling. First, current thermal datasets are either predominantly composed of static images, characterized by low frame rates, or lack continuous long-term physical interactions [10, 23, 27], making them insufficient for training predictive spatiotemporal models. Second, the fundamental objective of thermal visual representation remains confined to spatial pattern recognition. The formulation of action-conditioned predictive dynamics and spatiotemporal world modeling in thermal environments remains largely unaddressed.

3 Methodology

As illustrated in Fig. 2, our pipeline consists of three core components: (1) extracting intra-frame spatial topology from visible light (RGB) as privileged information; (2) modulating the thermal teacher's attention mechanism with RGB-derived topological boundaries as a confidence-gated soft prior into the thermal teacher's attention mechanism; and (3) optimizing the student network via masked spatiotemporal latent prediction.

3.1 Preliminaries and Spatial Homogenization

Given a pair of spatiotemporally aligned video sequences: a thermal infrared sequence $\mathbf{X}_{IR} \in \mathbb{R}^{B\times T\times C\times H\times W}$ and a visible sequence $\mathbf{X}_{RGB} \in \mathbb{R}^{B\times T\times C\times H\times W}$, where B is the batch size and T is the number of frames. Following the standard Video Joint-Embedding Predictive Architecture (V-JEPA2) [3], the video is divided into non-overlapping spatiotemporal patches. The sequence is flattened into $L = T \times N$ tokens, where N denotes the number of spatial patches per frame.

The thermal teacher encoder E_ξ, updated via an exponential moving average (EMA) of the student weights, outputs the flattened latent sequence $\mathbf{Z}^{seq}_{EMA} \in \mathbb{R}^{B\times(T\cdot N)\times D}$. However, applying this paradigm directly to thermal data leads to severe Intra-frame Spatial Homogenization. Due to thermal equilibrium, adjacent patches belonging to distinct physical entities often exhibit near-zero temperature gradients. Unconstrained self-attention inevitably homogenizes these patches, blurring rigid boundaries and degrading entities into unstructured continuous feature distributions. This ultimately causes action collapse in downstream dynamics modeling.

3.2 Extraction of Spatial Topology

To recover precise semantic segregation without hallucinating unobservable textures, we treat the aligned RGB data as privileged information (LUPI). Crucially, to preserve the temporal causality essential for a world model, we restrict the topological extraction strictly to the intra-frame spatial domain.

Intra-frame Feature Extraction. Since the pre-trained visible-light foundation model E_ψ operates on independent 2D images, we fold the temporal dimension into the batch dimension. After the forward pass, we explicitly reshape the sequence back to decouple the batch size B and the temporal horizon T, yielding the spatial feature tensor $\mathbf{F}^{4D}_{RGB}$:

$$\mathbf{F}^{4D}_{RGB} = \mathrm{Reshape}\Big(E_\psi\big(\mathrm{Flatten}(\mathbf{X}_{RGB}, 0, 1)\big)\Big) \in \mathbb{R}^{B\times T\times N\times d} \tag{1}$$

The output is then reshaped back into a 4D spatial tensor $\mathbf{F}^{4D}_{RGB} \in \mathbb{R}^{B\times T\times N\times d}$.

Intra-frame Topological Adjacency. Within each independent time slice t, we compute the pairwise cosine similarity between spatial patches. For a

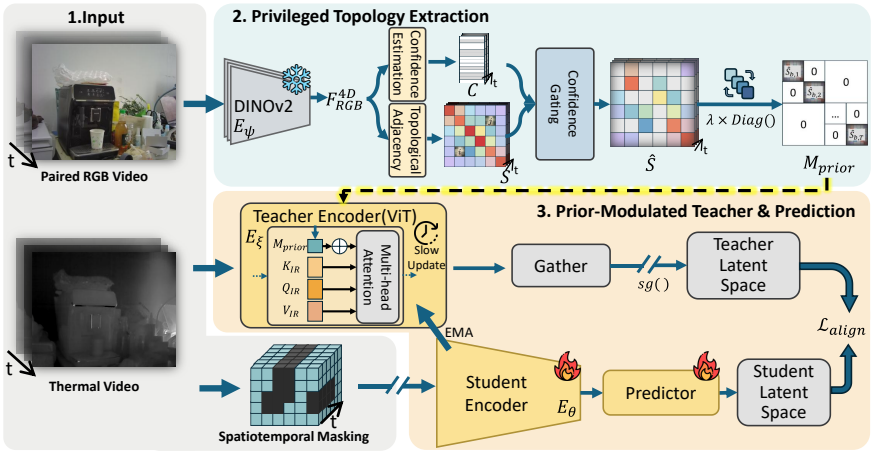

Fig. 2: The Thermo-JEPA framework. We inject the spatiotemporal prior (M_{prior}) into the teacher's attention mechanism to modulate the latent representation learning, ensuring the student predicts semantically aligned dynamics.

given batch index b and frame index t, the topological adjacency matrix $\mathbf{S} \in \mathbb{R}^{B\times T\times N\times N}$ is defined as:

$$S[b,t,i,j] = \frac{\langle \mathbf{F}^{4D}_{RGB}[b,t,i,:], \mathbf{F}^{4D}_{RGB}[b,t,j,:]\rangle}{\|\mathbf{F}^{4D}_{RGB}[b,t,i,:]\|_2 \|\mathbf{F}^{4D}_{RGB}[b,t,j,:]\|_2} \tag{2}$$

RGB Confidence Estimation. To mitigate the risk of visual bias, we explicitly assess the reliability of the visible-light features. We calculate the localized feature variance across the channel dimension of the DINOv2 embeddings to construct a spatial confidence map $\mathbf{C} \in [0,1]^{B\times T\times N}$, which serves as a proxy for structural informativeness:

$$C[b,t,i] = \sigma\left(\mathrm{Var}(\mathbf{F}^{4D}_{RGB}[b,t,i,:])\right) \tag{3}$$

where σ denotes the Sigmoid function. A higher $C[b,t,i]$ indicates salient, well-illuminated visible structures, whereas a lower value signifies degraded RGB regions (*e.g.*, severe darkness), preventing erroneous topology injection.

3.3 Confidence-Gated Structural Attention Modulation

We propose incorporating the RGB-derived topological information as a soft prior into the Multi-Head Self-Attention (MHSA) mechanism of the EMA thermal teacher E_ξ. This effectively decouples semantic boundaries without destroying microscopic temperature gradients.

Adhering to the V-JEPA2 architecture, the teacher's transformer layers process the fully flattened spatiotemporal sequence of length $L = T \times N$. Let

$\mathbf{Q}_{IR}, \mathbf{K}_{IR}, \mathbf{V}_{IR} \in \mathbb{R}^{B\times L\times d}$ denote the query, key, and value matrices. Since standard V-JEPA relies on joint spatiotemporal attention, calculating a full $L \times L$ attention map, we must align our strictly intra-frame RGB prior to this dimension without causing temporal leakage.

For a specific batch b and time step $t \in [1, T]$, let $\mathbf{c}_{b,t} \in \mathbb{R}^N$ denote the spatial confidence vector extracted from $\mathbf{C}[b, t, :]$. We construct the confidence-weighted spatial prior $\hat{\mathbf{S}}_{b,t} \in \mathbb{R}^{N\times N}$ as follows:

$$\hat{\mathbf{S}}_{b,t} = \left(\mathbf{c}_{b,t}\mathbf{c}_{b,t}^T\right) \odot \mathbf{S}[b, t] \tag{4}$$

where $\mathbf{c}_{b,t}\mathbf{c}_{b,t}^T$ computes the pairwise confidence matrix via an outer product, and $\odot$ denotes the Hadamard product.

To map this into the spatiotemporal latent space, we assemble these intra-frame priors into a large block-diagonal matrix $\mathbf{M}_{prior}^{(b)} \in \mathbb{R}^{(T\cdot N)\times(T\cdot N)}$ for each sequence in the batch:

$$\mathbf{M}_{prior}^{(b)} = \lambda \cdot \text{BlockDiag}(\hat{\mathbf{S}}_{b,1}, \hat{\mathbf{S}}_{b,2}, \ldots, \hat{\mathbf{S}}_{b,T}) \tag{5}$$

where λ is a learnable scalar initialized to zero. For any two tokens across different time frames ($t_i \neq t_j$), the prior bias is strictly 0, mathematically guaranteeing zero temporal leakage.

The modified joint spatiotemporal attention within the teacher network is then computed as:

$$\text{Attention}(\mathbf{Q}_{IR}, \mathbf{K}_{IR}, \mathbf{V}_{IR}) = \text{Softmax}\left(\frac{\mathbf{Q}_{IR}\mathbf{K}_{IR}^T}{\sqrt{d}} + \mathbf{M}_{prior}\right)\mathbf{V}_{IR} \tag{6}$$

Optimization Properties. This gated modulation offers two intrinsic advantages. First, it ensures thermal autonomy: the additive prior $\mathbf{M}_{prior}$ acts as a soft constraint rather than a hard replacement. When native thermal gradients are salient, the self-attention mechanism naturally dominates, preserving microscopic thermodynamic semantics. Second, the confidence map $\mathbf{C}$ provides an adaptive fallback. In scenarios with severe visible-light degradation, the gating systematically zeroes out $\mathbf{M}_{prior}$, allowing the network to gracefully degenerate to a pure infrared-driven self-attention without introducing erroneous biases.

3.4 Masked Spatiotemporal Prediction and Alignment

The student encoder E_θ processes only the unmasked tokens from $\mathbf{X}_{IR}$. Combined with learnable mask tokens and 3D-RoPE, the predictor P_ϕ infers the masked patches, yielding $\mathbf{Z}_{pred} \in \mathbb{R}^{B\times M\times D}$ (M is the mask count).

To close the logical loop, we utilize a Gather operation to extract the targets corresponding to the masked indices $\mathcal{M}$ directly from the sequence $\mathbf{Z}_{EMA}^{seq}$ generated by the teacher model. Thanks to the confidence-gated attention modulation, $\mathbf{Z}_{EMA}^{seq}$ intrinsically possesses clear semantic boundaries while preserving thermodynamic details. The objective minimizes the L1 distance:

$$\mathcal{L}_{align} = \frac{1}{B\cdot M}\sum_{b=1}^{B}\sum_{m\in\mathcal{M}} \left\|\mathbf{Z}_{pred}[b, m, :] - \text{sg}\left(\text{Gather}(\mathbf{Z}_{EMA}^{seq}, \mathcal{M})[b, m, :]\right)\right\|_1 \tag{7}$$

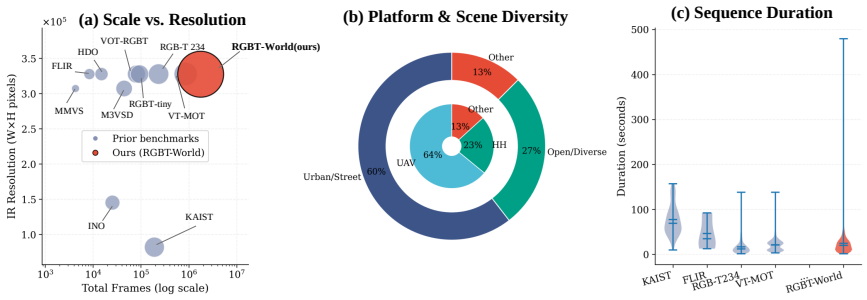

Fig. 3: Statistical superiority of the *RGBT-World* dataset. (a) Compared to prior benchmarks, our dataset achieves unprecedented scale and IR resolution. (b) Multi-level distribution demonstrates extensive platform and environmental diversity. (c) Sequence duration analysis reveals our dataset provides long-term continuous videos essential for spatiotemporal world modeling, overcoming the fragmentation of legacy datasets.

where sg(·) explicitly prevents any visible-light modal information from leaking into the student's gradient flow.

4 Experiments

To validate the theoretical claims of Thermo-JEPA, we design our evaluation protocol around three objectives: (1) benchmarking zero-shot dynamics alignment to address IR data scarcity; (2) quantifying the structural integrity of the latent space against thermal homogenization; and (3) ablating the individual architectural contributions.

4.1 The Aligned RGB-T Video Dataset

Existing thermal datasets are fragmented and tailored for narrow discriminative tasks (*e.g.*, detection, short-term tracking). They critically lack the sensor resolution, platform diversity, and temporal continuity required for self-supervised world modeling, shown in Fig. 3. To address this, we introduce *RGBT-World*, the largest unified, strictly aligned RGB-Thermal video corpus to date.

Construction and Scale. As detailed in Table 1 and Fig.3(a), *RGBT-World* aggregates rigidly filtered sequences from 10 public benchmarks and introduces *Ours-Internal*, a high-resolution subset. The corpus totals 2,622 videos comprising 1.75M synchronized frame pairs (16.78 hours), providing unprecedented scale for generative pre-training.

Diversity and Temporal Continuity. To ensure viewpoint robustness, *RGBT-World* unifies four acquisition platforms (UGV, UAV, Handheld, CCTV) across diverse environments (Fig. 3(b)). Furthermore, our *Ours-Internal* subset injects high-resolution data (up to 2700×2160 RGB / 640×512 IR), forcing the encoder to learn scale-invariant topologies rather than overfitting to legacy low-res

Table 1: Comprehensive comparison of RGB-Thermal video datasets. Our *RGBT-World* dataset aggregates 10 public benchmarks and introduces a high-resolution self-collected subset (*Ours-Internal*). Unlike prior datasets limited to specific platforms (UGV/Handheld) or low resolutions, our dataset offers the first unified, high-resolution, multi-platform corpus with 100% spatiotemporal alignment, tailored for world model pre-training. *Platform Key: UGV (Unmanned Ground Vehicle), UAV (Unmanned Aerial Vehicle), HH (Handheld), CCTV (Surveillance).*

Dataset	Platform	Environment	Res. (RGB)	Res. (IR)	Alignment	Videos	Total Frames
Existing Public Benchmarks (Discriminative Tasks)							
KAIST [20]	UGV	Urban Day/Night	640×480	320×256	Aligned	82	190,648
FLIR [10]	UGV	Highway/Street	1600×1200	640×512	Aligned	6	8,390
VT-MOT [50]	UAV	Campus/Street	1920×1080	640×512	Aligned	1344	856,136
RGB-T234 [29]	HH	Open World	Mixed	Mixed	Aligned	468	233,298
VOT-RGBT [24]	Mixed	Diverse	Mixed	Mixed	Aligned	240	80,332
HDO [43]	HH	Office/Urban	1920×1080	640×512	Aligned	54	15,000
MMVS [31]	HH	Indoor/Outdoor	1224×1024	640×480	Aligned	12	4,301
M3VSD [44]	HH	Salient Objects	640×480	640×480	Aligned	60	44,542
INO [21]	SC	Surveillance	448×324	448×324	Aligned	20	25,390
RGBT-tiny [47]	UAV	Tiny Objects	640×512	640×512	Aligned	230	93,402
Ours (Generative World Modeling)							
Ours-Internal	**UAV**	**Urban/Aerial**	$\mathbf{2700 \times 2160}$	$\mathbf{640 \times 512}$	**Strictly Paired**	**106**	**202,810**
RGBT-World	**All**	**Unified**	**Mixed (HQ)**	**Mixed (HQ)**	**100% Aligned**	**2,622**	**1,754,249**

artifacts. Crucially, as shown in Fig. 3(c), our dataset features long-term continuous videos (extending up to 500 seconds), fundamentally solving the temporal fragmentation of prior benchmarks.

4.2 Experimental Setup

Pre-training Configuration. The Thermo-JEPA encoder E_θ is parameterized as a standard ViT-L/16 with 300M parameters, adhering to the architectural design of V-JEPA2 [3]. The video input is partitioned into non-overlapping tubelets of size $16 \times 16 \times 2$. To extract the spatial topology matrix as privileged information, we employ a frozen, pre-trained DINOv2 (ViT-L) [33] as the RGB spatial encoder E_ψ. Pre-training is conducted exclusively on the 15-hour training split of our *RGBT-World* dataset. All model inputs are normalized raw thermal values (single-channel intensity); any colorized visualizations in the figures are solely for human readability.

Implementation Details. For rigorous benchmarking and to avoid hyperparameter tuning biases, we strictly adhere to the default optimization protocols of the official V-JEPA2 [3]. The model is trained using the AdamW optimizer with a cosine learning rate decay (peak learning rate of $1e-4$) and a weight decay of 0.04. The Exponential Moving Average (EMA) momentum for the teacher network is fixed at 0.999. The spatiotemporal masking ratio is set dynamically between 30% and 80% to force deep semantic reconstruction rather than trivial local interpolation. All models are trained across $4 \times$ NVIDIA A6000 GPUs.

Evaluation Protocol. For all quantitative evaluations, we strictly employ a **frozen-backbone protocol**. The parameters of the Thermo-JEPA encoder E_θ are entirely frozen after pre-training. We solely train lightweight task-specific

heads. This protocol ensures that the reported performance reflects the intrinsic topological and kinematic structures embedded within the pre-trained latent space, rather than the capacity of the task-specific heads.

4.3 Zero-Shot Dynamics Alignment

To evaluate the generalization and structural alignment of the learned representations, we propose the Zero-Shot Dynamics Alignment benchmark. The objective is to determine whether the thermal latent space has successfully acquired kinematic and topological structures equivalent to the RGB modality without direct action supervision.

Protocol. We structure this as an Action Anticipation and Classification task on the hold-out evaluation split of our *RGBT-World* dataset. To ensure a strictly fair comparison, all models are evaluated under an identical IR-only frozen-backbone protocol: **Probe Training:** For every model in Table 2, we freeze the pre-trained backbone and extract features from the IR training split of RGBT-World. An attentive action probe is then trained on these frozen IR features following the standard V-JEPA2 [3] configuration. **Evaluation:** The trained probe is frozen. Importantly, every backbone—including image and video foundation models originally pre-trained on RGB data—is given only unannotated IR video sequences from the hold-out test split at inference time, and the extracted latent features are fed into the IR-trained probe. **Fairness Guarantee:** All models (including Thermo-JEPA and all baselines) receive identical IR-only inputs; the only variable is the quality of the pre-trained backbone.

Baselines. We compare Thermo-JEPA against an extensive suite of state-of-the-art foundation models: *Image-based Foundation Models*: DINOv2 [33] and SigLIP [38], applied frame-by-frame. *General Video Foundation Models (Feature Extraction)*: InternVideo2-6B [41], VideoPrism [53], and the recent physics-aware Cosmos-World [1] (utilizing open-weight inference). *Thermal-Specific & Cross-Modal Variants*: A Vanilla V-JEPA2 trained exclusively on our IR data to demonstrate the homogenization failure, and a contrastive RGB-T baseline that directly aligns IR [CLS] tokens with RGB [CLS] tokens via InfoNCE loss.

Results and Analysis. As reported in Table 2, while generic massive VFMs (Cosmos-World, InternVideo2) suffer from a severe domain gap, the most stringent comparison lies in models pre-trained on the identical RGBT-World dataset. Vanilla V-JEPA2 trained exclusively on our IR data yields the lowest performance (18.6% Acc), corroborating the spatial homogenization collapse. Contrastive-IR (early fusion) achieves 24.2% Acc, verifying that dense cross-modal imitation incurs modality pollution rather than dynamic alignment. By explicitly isolating gradients, Thermo-JEPA (45.8% Acc) demonstrates massive, consistent gains over all strictly comparable baselines.

4.4 Diagnostic Probing

Beyond macroscopic dynamics, we conduct diagnostic linear probing to decipher the topological integrity within the frozen latent space. We formulate localiza-

Table 2: Zero-shot Action Anticipation on RGBT-World (Hold-out Set). All models utilize a frozen backbone. The attentive probe is trained and tested exclusively on IR data. All models receive identical IR-only inputs; the only variable is the pre-trained backbone. Thermo-JEPA significantly outperforms massive video foundation models, demonstrating that explicit thermodynamic decoupling is more critical than raw data scale in thermal environments.

Category	Method	Year	Pre-train Params	Pre-train Data	Top-1 Acc (%)	Recall@5
Image-based Modality Transfer						
	DINOv2 [33]	2023	1.1B	LVD-142M	21.4	34.2
	SigLIP 2 [38]	2025	1.2B	WebLI	23.8	36.1
General Video Foundation Models (Data-Driven)						
	InternVideo2 [41]	2024	6B	Internal	29.6	41.3
	VideoPrism [53]	2024	1B	VideoPrism-Corpus	31.2	43.5
	Cosmos-World [1]	2025	7B	Cosmos-Data	33.5	46.8
Thermal/Cross-modal Pre-training on RGBT-World						
	Vanilla V-JEPA2 (IR-Only)	2024	300M	RGBT-World	18.6	28.4
	Contrastive-IR (Early Fusion)	2025	300M	RGBT-World	24.2	35.5
	Thermo-JEPA (Ours)	**2026**	**300M**	**RGBT-World**	**45.8**	**62.1**

Table 3: Diagnostic Probing of Spatial and Kinematic Integrity. Lower ADE/FDE and higher mIoU indicate that the frozen latent space successfully isolates distinct physical entities without thermodynamic blurring.

Method	Spatial mIoU (↑)	Trajectory ADE (↓)	Trajectory FDE (↓)
DINOv2 (Frozen) [33]	48.2	15.6	21.3
VideoMAE v2 (IR-Supervised) [40]	55.4	12.3	17.8
Vanilla V-JEPA2 (IR-Only) [3]	32.1	28.5	39.4
Thermo-JEPA (Ours)	**68.7**	**8.4**	**11.2**

tion and trajectory forecasting strictly as analytical probes (rather than seeking tracking SOTA), evaluating along two axes: 1. **Spatial Topology (Current Location):** We attach a randomly initialized 1×1 convolutional layer to the frozen encoder outputs to predict frame-level bounding boxes $[x, y, w, h]$. Performance is quantified by Mean Intersection over Union (**mIoU**). 2. **Kinematic Consistency (Future Trajectory):** Given $T = 8$ context frames, a lightweight single-layer Transformer predicts the object center coordinates for the subsequent $K = 8$ frames. Performance is measured by Average Displacement Error (**ADE**) and Final Displacement Error (**FDE**).

Results and Analysis. Table 3 provides quantitative evidence of the homogenization phenomenon. Vanilla V-JEPA2 exhibits a drastically low spatial mIoU (32.1) and high trajectory ADE (28.5), indicating that objects blend into the background and temporal gradients dissipate. Thermo-JEPA fundamentally reverses this degradation. By explicitly modulating the teacher's attention manifold with RGB-derived semantic boundaries, the frozen features achieve a high mIoU of 68.7 and an ADE of 8.4. This proves that the soft prior successfully guides the model to extract sharp, predictable macro-physical entities from blurry thermal inputs without destroying intra-object variance.

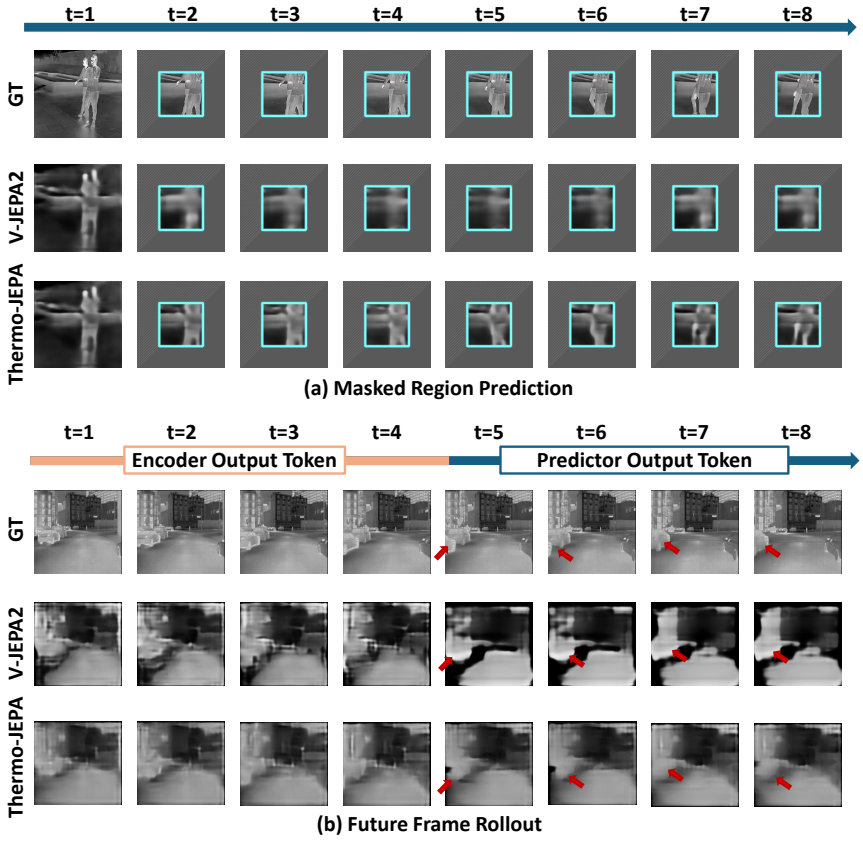

Fig. 4: Dynamic Visual Forecasting of Latent Representations. (a) Masked Region Prediction: Cyan boxes highlight the masked spatiotemporal tubelets targeted for reconstruction. Vanilla V-JEPA2 suffers from severe thermal blending, whereas Thermo-JEPA, guided by the soft topological prior, distinctly resolves rigid physical boundaries (*e.g.*, human limbs). **(b) Future Frame Rollout:** Given context frames ($t = 1 \rightarrow 4$), models predict future representations ($t = 5 \rightarrow 8$). Red arrows track a moving vehicle. Vanilla V-JEPA2's representations suffer from catastrophic structural smearing, dissipating into amorphous noise. In contrast, Thermo-JEPA sharply preserves the dynamic topological structure of moving entities over a longer horizon.

4.5 Qualitative Analysis

To visually diagnose how the geometric constraints affect spatiotemporal understanding, we train a lightweight CNN decoder using an MSE reconstruction loss to map the frozen latent features back to the thermal pixel space. To prevent temporal leakage during visualization, the decoder operates strictly frame-by-frame with shared weights, guaranteeing that all observed dynamics are intrinsic to the latent space itself.

Masked Spatiotemporal Reconstruction. In Fig. 4(a), Vanilla V-JEPA2 reconstructions of the masked regions exhibit severe feature adhesion; adjacent entities are rendered as a single unresolved thermal mass due to unconstrained self-attention in near-zero gradient regions. Conversely, reconstructions from Thermo-JEPA clearly maintain the separation between distinct entities. The boundaries are significantly sharper, providing visual proof that the confidence-gated topological prior successfully prevents inter-object semantic bleeding during pre-training.

Future Frame Rollout. In Fig. 4(b), as the forecasting horizon extends ($t = 5 \rightarrow 8$), Vanilla V-JEPA2 experiences a catastrophic geometric collapse. Instead of maintaining entity boundaries, moving targets (*e.g.*, the vehicle) smear into massive amorphous artifacts and ultimately become entirely indiscernible, severely corrupting the spatiotemporal latent space. In stark contrast, Thermo-JEPA sustains relative dynamic clarity for a significantly longer duration. Crucially, it preserves temporal consistency, enabling reliable semantic identification of the targets over time. The semantic occupancies tightly maintain their shape without spatial diffusion, demonstrating that the geometrically anchored latent space is inherently more stable for long-term transition modeling.

Latent Dimensionality and Feature Collapse. A Principal Component Analysis (PCA) of the deepest encoder layers (Fig. 5) reveals that Vanilla V-JEPA2 undergoes dimensional collapse: its deep features align almost exclusively with PC1 (representing global thermal averages). In contrast, the deepest features of Thermo-JEPA maintain significant correlations with both PC1 and PC2. This geometric distribution objectively indicates that our attention modulation mechanism forces the network to retain a richer, higher-dimensional information capacity.

4.6 Ablation Studies

We ablate the core architectural components of Thermo-JEPA to validate their individual contributions. All ablation variants are pre-trained on *RGBT-World* for 50 epochs. We evaluate using Spatial mIoU (topological integrity) and Zero-shot Action Anticipation Top-1 Accuracy (kinematic transferability).

Table 4: Ablation Study. We validate injection strategies, prior types, and gating mechanisms. Random or edge-based priors degrade performance, while DINOv2 semantic topology with confidence gating and stop-gradient achieves the optimal balance.

Variant	Prior Source	Confidence Gating (C)	Gradient Flow	Spatial mIoU	Zero-Shot Acc (%)
(a) Vanilla V-JEPA2	None	No	N/A	32.1	18.6
(b) Random Topology	Uniform Noise	Yes	Stopped	15.2	12.4
(c) Edge-based Prior	Sobel Filter	Yes	Stopped	41.5	28.7
(d) Un-gated Soft Attention	DINOv2	No	Stopped	61.2	38.5
(e) End-to-End Leakage	DINOv2	Yes	Allowed	54.8	31.2
(f) Thermo-JEPA (Ours)	**DINOv2 (Variance)**	**Yes**	**Stopped**	**68.7**	**45.8**

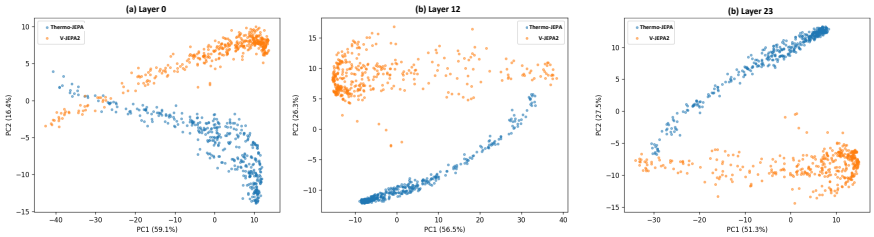

Fig. 5: PCA projection of encoder layers. V-JEPA2 exhibits dimensional collapse, primarily retaining variance only along PC1. Thermo-JEPA preserves a higher-dimensional representation, maintaining significant correlations with both PC1 and PC2.

1. The Necessity of RGB Confidence Gating. We first examine the Ungated Soft Attention variant (d), which omits the confidence gating term $\mathbf{c}_{b,t}\mathbf{c}_{b,t}^T$ in Eq. 4 and directly adds the DINOv2-derived spatial prior into the attention logits. As reported in Table 4, comparing Variant (d) and our full model (Variant f, 45.8% Acc) highlights the critical role of the confidence gating map $\mathbf{C}$. Without gating, the model blindly trusts RGB topological priors even in degraded conditions, injecting erroneous attention biases that disrupt the thermal latent space. The confidence gating ensures adaptive fallback, significantly boosting the robustness of the representations.

2. Preventing Modal Pollution via Gradient Isolation. Variant (e) removes the stop-gradient ($\text{sg}(\cdot)$) operator during the alignment phase, allowing gradients to backpropagate through the target representations. This results in a performance drop to 31.2% Acc. Without gradient isolation, the student network attempts to implicitly decode visible-light textures to minimize the loss leading to feature collapse during zero-shot inference. This confirms our hypothesis: the EMA teacher's representations must remain strictly gradient-free targets to prevent the thermal latent space from implicitly decoding unobservable RGB textures.

3. Semantic Topology vs. Low-level Edges. As shown in Table 4, injecting a random prior (Var b) collapses the representation. Substituting DINOv2 semantics with low-level RGB edge density (Var c) yields sub-optimal accuracy (28.7%), proving that high-level semantic grouping—not just raw pixel gradients—is essential. Furthermore, variance-based confidence naturally outperforms edge density, as edges fail under motion blur.

5 Conclusion

In this paper, we identified intra-frame spatial homogenization as the primary bottleneck for extending masked world modeling to the thermal infrared domain. To tackle this, we introduced Thermo-JEPA, a novel pre-training framework that leverages visible-light topologies as a confidence-gated structural prior. By

explicitly modulating the teacher's attention mechanism with explicitly segregating semantic boundaries, Thermo-JEPA prevents representation collapse and modal pollution, ensuring strict zero-shot temporal robustness on pure thermal inputs. Concurrently, we presented RGBT-World, the most comprehensive spatiotemporally aligned video dataset to date, to fuel future generative thermal representation learning. Extensive evaluations demonstrate that our approach significantly outperforms generic video foundation models, paving the way for reliable, all-weather physical perception and autonomous spatiotemporal forecasting.

Acknowledgements

Acknowledgments: This work was funded by the National Natural Science Foundation of China under grants 62475016, the Beijing Natural Science Foundation under Grant L252142, and the National Key Laboratory of Proximity Detection and Control Funding under grant 6142601012402.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15619–15629 (2023)
3. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
4. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: European conference on computer vision. pp. 348–367. Springer (2022)
5. Berg, A., Ahlberg, J., Felsberg, M.: Generating visible spectrum images from thermal infrared. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1143–1152 (2018)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
7. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
8. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)

9. Chen, C., Ye, M., Qi, M., Wu, J., Jiang, J., Lin, C.W.: Structure-aware positional transformer for visible-infrared person re-identification. IEEE Transactions on Image Processing **31**, 2352–2364 (2022)
10. FLIR Systems: Flir thermal dataset for algorithm training. `https://www.flir.com/oem/adas/adas-dataset-form/` (2018)
11. Gade, R., Moeslund, T.B.: Thermal cameras and applications: a survey. Machine vision and applications **25**(1), 245–262 (2014)
12. Ghose, D., Desai, S.M., Bhattacharya, S., Chakraborty, D., Fiterau, M., Rahman, T.: Pedestrian detection in thermal images using saliency maps. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
13. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
14. Girdhar, R., El-Nouby, A., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Omnimae: Single model masked pretraining on images and videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10406–10417 (2023)
15. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3), 440 (2018)
16. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)
17. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023)
18. Hansen, N.A., Su, H., Wang, X.: Temporal difference learning for model predictive control. In: International Conference on Machine Learning. pp. 8387–8406. PMLR (2022)
19. 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)
20. Hwang, S., Park, J., Kim, N., Choi, Y., So Kweon, I.: Multispectral pedestrian detection: Benchmark dataset and baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
21. INO: Video analytics dataset. `https://www.ino.ca/en/technologies/video-analytics-dataset/` (2026), accessed: 2024-05-21
22. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
23. Jia, X., Zhu, C., Li, M., Tang, W., Zhou, W.: Llvip: A visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3496–3504 (2021)
24. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J.K., Zajc, C.L., Drbohlav, O., Lukezic, A., Berg, A., Eldesokey, A., Kapyla, J., Fernandez, G., Gonzalez-Garcia, A., Memarmoghadam, A., Lu, A., He, A., Varfolomieiev, A., Chan, A., Tripathi, S.A., Smeulders, A., Pedasingu, S.B., Chen, X.B., Zhang, B., Wu, B., Li, B., He, B., Yan, B., Bai, B., Li, B., Li, B., Kim, H.B., Ki, H.B.: The seventh visual object tracking vot2019 challenge results. international conference on computer vision (2019)

25. Kuang, X., Sui, X., Liu, Y., Chen, Q., Gu, G.: Single infrared image enhancement using a deep convolutional neural network. Neurocomputing **332**, 119–128 (2019)
26. LeCun, Y., et al.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
27. Lee, A.J., Cho, Y., Shin, Y.s., Kim, A., Myung, H.: Vivid++: Vision for visibility dataset. IEEE Robotics and Automation Letters **7**(3), 6282–6289 (2022)
28. Li, C., Cheng, H., Hu, S., Liu, X., Tang, J., Lin, L.: Learning collaborative sparse representation for grayscale-thermal tracking. IEEE Transactions on Image Processing **25**(12), 5743–5756 (2016)
29. Li, C., Liang, X., Lu, Y., Zhao, N., Tang, J.: Rgb-t object tracking: Benchmark and baseline. Pattern Recognition **96**, 106977 (2019)
30. Maheshwari, P., Karhade, J., Chawla, Y., Adu, I., Heisen, F., Porco, A., Jong, A., Liu, Y., Pitla, S., Scherer, S., et al.: Anythermal: Towards learning universal representations for thermal perception. arXiv preprint arXiv:2602.06203 (2026)
31. Meng Sang, H.X., Yang, Y.: Vrff: Video registration and fusion framework. In: Proceedings of 2024 International Joint Conference on Neural Networks (IJCNN). IEEE (Institute of Electrical and Electronics Engineers) (2024)
32. Okada, M., Taniguchi, T.: Dreamingv2: Reinforcement learning with discrete world models without reconstruction. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 985–991. IEEE (2022)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
34. Pechyony, D., Vapnik, V.: On the theory of learnining with privileged information. Advances in neural information processing systems **23** (2010)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
36. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
37. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
38. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
39. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural Networks **22**(5), 544–557 (2009). `https://doi.org/10.1016/j.neunet.2009.06.042`, advances in Neural Networks Research: IJCNN2009
40. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14549–14560 (2023)
41. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European conference on computer vision. pp. 396–416. Springer (2024)

42. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14668–14678 (2022)
43. Xie, H., Sang, M., Zhang, Y., Yang, Y., Zhao, S., Zhong, J.: Rcvs: A unified registration and fusion framework for video streams. IEEE Transactions on Multimedia **26**, 11031–11043 (2024)
44. Xu, D., Ouyang, W., Ricci, E., Wang, X., Sebe, N.: Learning cross-modal deep representations for robust pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5363–5371 (2017)
45. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 6787–6800 (2021)
46. Yao, H., Yang, B., Huang, W., Du, B., Ye, M.: Unsupervised visible-infrared person re-identification under unpaired settings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11916–11926 (2025)
47. Ying, X., Xiao, C., An, W., Li, R., He, X., Li, B., Cao, X., Li, Z., Wang, Y., Hu, M., et al.: Visible-thermal tiny object detection: A benchmark dataset and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
48. Zha, Y., Sun, J., Zhang, P., Zhang, L., Gonzalez-Garcia, A., Huang, W.: Self-supervised cross-modal distillation for thermal infrared tracking. IEEE MultiMedia **29**(4), 80–96 (2022)
49. Zhang, H., Zhang, L., Zhuo, L., Zhang, J.: Object tracking in rgb-t videos using modal-aware attention network and competitive learning. Sensors **20**(2), 393 (2020)
50. Zhang, L., Danelljan, M., Gonzalez-Garcia, A., Van De Weijer, J., Shahbaz Khan, F.: Multi-modal fusion for end-to-end rgb-t tracking. In: Proceedings of the IEEE/CVF International conference on computer vision workshops. pp. 0–0 (2019)
51. Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., Liu, Z.: Weakly aligned cross-modal learning for multispectral pedestrian detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5127–5137 (2019)
52. Zhang, T., Wiliem, A., Yang, S., Lovell, B.: Tv-gan: Generative adversarial network based thermal to visible face recognition. In: 2018 international conference on biometrics (ICB). pp. 174–181. IEEE (2018)
53. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217 (2024)