From Static to Dynamic: Exploring Self-supervised Image-to-Video Representation Transfer Learning

Supplementary Material

Appendix Contents

A Symbol Definitions 1

B Formal Theorems and Proofs 2
- B.1. Spectral Properties of Optimal Projections . . 2
  - B.1.1. Settings for Linear Projection 2
  - B.1.2. Settings for MLP Projection 2
  - B.1.3. Proof for Theorem 1 3
- B.2. Trade-off Improvement 5

C Supplementary Explanation of Method 7
- C.1. Differences with Previous Methods 7
- C.2. Algorithm of the Framework 7

D Detailed Description of Experiments 7
- D.1. Training Datasets 7
- D.2. Training Settings 8
- D.3. Evaluation Settings 8
  - D.3.1. Evaluation on Dense-level Benchmarks 8
  - D.3.2. Evaluation on Frame-/Video-level Benchmarks 9
  - D.3.3. Distance-based Trade-off Metrics Validation 10
- D.4. Image-pretrained Fundamental Models . . . 11
- D.5. Competitors 12

E Detailed Experiments Results 13
- E.1. Comparison with Task-Specific SOTAs . . . 13
- E.2. Detailed Results of Frame-/Video-Level Tasks 13
- E.3. Training Dynamics 14
- E.4. Shortcut Phenomenon in Training 14
- E.5. Additional Ablation Study 14

F. Additional Visualizations 15
- F.1. Inter-frame Correspondence 15
- F.2. Downstream Task Performance 16

G Detailed Related Work 16
- G.1. Self-supervised Visual Representation 16
- G.2 Image-to-video Transfer Learning 18
- G.3. Temporal Cycle Consistency 18

H Additional Discussions 18
- H.1. Limitation and Future Work 18
- H.2 Broader Impact 18

A. Symbol Definitions

We summarize the key notations used in the Method and Theoretical Analysis sections in Tab. 1 and Tab. 2.

Table 1. A summary of key notations and descriptions used in the Method Section.

Notations	Descriptions
$\mathcal{D}$	Training dataset.
$H/W/C$	The height/width/channel dimension of a frame.
T	The number of frames in a video.
$\boldsymbol{V}$	A video containing T frames.
$\boldsymbol{v}_t$	The frame at the moment t in a video, $\boldsymbol{v}_t \in \mathbb{R}^{H\times W\times C}$.
$\boldsymbol{v}_t(i)$	A frame patch of $\boldsymbol{v}_t$, $\boldsymbol{v}_t \in \mathbb{R}^{H\times W\times C}$.
δ	The temporal offset between two frames, $\delta \in (0, 1)$.
p	The size of a frame patch.
N_H	The patch number on the height dimension, $N_H = H/p$.
N_W	The patch number on the width dimension, $N_W = W/p$.
N	The patch number of a frame, $N = N_H \times N_W$.
d	The embedding dimension of a frame patch.
$f(\cdot)$	The image-pretrained encoder, $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{N\times d}$.
$g(\cdot)$	The projection layer, $g : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$.
α	The amplitude of the positional encoding interpolation.
$\mathbf{E}_{\text{pos}}$	The positional encoding of f.
$\widetilde{\mathbf{E}}_{\text{pos}}$	The augmented version of $\mathbf{E}_{\text{pos}}$.
$\boldsymbol{z}_t$	The original representation of $\boldsymbol{v}_t$, where $\boldsymbol{z}_t = f(\boldsymbol{v}_t)$.
$\boldsymbol{p}_t$	The projected representation of $\boldsymbol{z}_t$, where $\boldsymbol{p}_t = g(\boldsymbol{z}_t)$.
$\boldsymbol{A}_{t_1}^{t_2}$	The transition matrix between representations $\boldsymbol{p}_{t1}$ and $\boldsymbol{p}_{t2}$.
λ	The strength of the constraint term.

Table 2. A summary of key notations and descriptions used in the Theoretical Analysis Section.

Notations	Descriptions
d	The embedding dimension of a frame patch.
λ	The strength of the constraint term.
$f(\cdot)$	The image-pretrained encoder, $f : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{N\times d}$.
$g(\cdot)$	The projection layer, $g : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$.
$\boldsymbol{z}_i$	The latent representation of an input patch.
$\bar{\boldsymbol{z}}_i$	The mean representation of video $\boldsymbol{V}_i$, $\bar{\boldsymbol{z}}_i = \mathbb{E}_{\boldsymbol{z}\in f(\boldsymbol{V}_i)}[\boldsymbol{z}]$.
$\boldsymbol{p}_i$	The projected representation of $\boldsymbol{z}_i$.
$\boldsymbol{W}$	The projection weight of the linear layer.
$\boldsymbol{W}_1/\boldsymbol{W}_2$	The projection weight of the two-layer MLP.
$\phi(\cdot)$	The `tanh` activation function.
$\boldsymbol{J}_g(\cdot)$	The Jacobian matrix of g.
$\boldsymbol{\Sigma}$	The intra-video covariance matrix between two patches.
$\bar{\boldsymbol{\Sigma}}$	The inter-video covariance matrix between two videos.
$\boldsymbol{U}$	The orthogonal basis for spectral decomposition.
$\boldsymbol{\Lambda}_W/\boldsymbol{\Lambda}_1/\boldsymbol{\Lambda}_2$	The eigenvalue matrices of $\boldsymbol{W}/\boldsymbol{W}_1/\boldsymbol{W}_2$.
$\boldsymbol{\Lambda}_\Sigma/\boldsymbol{\Lambda}_{\bar{\Sigma}}$	The eigenvalue matrices of $\boldsymbol{\Sigma}/\bar{\boldsymbol{\Sigma}}$.
$\mu_i/\mu_{1,i}/\mu_{2,i}$	The eigenvalues of $\boldsymbol{W}/\boldsymbol{W}_1/\boldsymbol{W}_2$.
σ_i/τ_i	The eigenvalues of $\boldsymbol{\Sigma}/\bar{\boldsymbol{\Sigma}}$.
D_{intra}	The intra-video distance between two patches.
D_{inter}	The inter-video distance between two videos.
γ	The scale factor between D_{intra} and D_{inter}.
D	The margin of inter-/intra-video distances.
Δ	The improvement of $D(\boldsymbol{z}_1, \boldsymbol{z}_2)$.

B. Formal Theorems and Proofs

In this section, we provide detailed proofs for the theoretical analysis of how the proposed method achieves our target, *i.e.*, improving the **intra-video temporal consistency** without largely affecting the **inter-video semantic separability**. Generally, our analysis leads to two main conclusions: **a)** Within our proposed method, both linear-based and MLP projection rebalance different dimensions of the representation space in a similar mechanism (Theorem 1). **b)** This rebalance yields a better trade-off between the two properties under appropriate conditions (Theorem 2).

Formally, given the original representation of a patch $\boldsymbol{z}_i \in \mathbb{R}^d$, we aim to learn a projection g that maps $\boldsymbol{z}_i$ to $\boldsymbol{p}_i = g(\boldsymbol{z}_i) \in \mathbb{R}^d$. Since directly analyzing the original objectives $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{\text{reg}}$ is challenging, we introduce simplified yet equivalent surrogates to facilitate the analysis.

Objective 1 (Temporal Cycle Consistency). This term encourages alignment between temporally corresponding patches. We quantify it with the metric in Eq. (1). Note that minimizing M_{cyc} is equivalent to minimizing the cycle-consistency loss L_{cyc}, since both decrease as temporal consistency improves and share the same optimality conditions.

$$M_{\text{cyc}} = \frac{1}{2}\,\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|g(\boldsymbol{z}_1) - g(\boldsymbol{z}_2)\|^2\big]. \tag{1}$$

Objective 2 (Semantic Separability Constraint): The KL divergence constraint $\mathcal{L}_{\text{reg}}$ preserves the distance relationships between patches before and after the projection, which is equivalent to constraining the projection to be isometric. This property can be measured by the orthogonality of the Jacobian matrix [14, 40, 47, 50] of g, as formulated in Eq. (2). Therefore, we use it as an approximation of L_{reg}.

$$M_{\text{reg}} = \frac{1}{2}\,\mathbb{E}_{\boldsymbol{z}_i}\big[\|\boldsymbol{J}_g(\boldsymbol{z}_i)\boldsymbol{J}_g(\boldsymbol{z}_i)^\top - \boldsymbol{I}\|_F^2\big]. \tag{2}$$

Combining the two surrogates yields the overall objective:

$$\min_g \; M(g) = M_{\text{cyc}} + \lambda M_{\text{reg}}. \tag{3}$$

We now consider two representative cases for g: **i)** A linear projection: $g(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$; **ii)** A two-layer MLP: $g(\boldsymbol{z}) = \boldsymbol{W}_2\,\phi(\boldsymbol{W}_1\boldsymbol{z})$ with activation function $\phi(\cdot) = tanh(\cdot)$, and this case represents more complex modules. The following theorem analyzes the spectral properties of the optimal solution under both cases, illustrating how the projection affects the quality of the transferred representation.

B.1. Spectral Properties of Optimal Projections

B.1.1. Settings for Linear Projection

For the linear projection $g(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$, the Jacobian matrix can be expressed as $\boldsymbol{J}_g(\boldsymbol{z}_i) = \dfrac{\partial g}{\partial \boldsymbol{z}_i} = \boldsymbol{W}$, thereby the optimization objective can be reformulated as:

$$\begin{aligned}\min_{\boldsymbol{W}} \; M(\boldsymbol{W}) = &\frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2\|^2\big] \\ &+ \frac{\lambda}{2}\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{I}\|_F^2.\end{aligned} \tag{4}$$

To facilitate the analysis, we begin by introducing several definitions and assumptions.

Definition 1 (Intra-video Covariance Matrix)**.** Define the intra-video covariance matrix as the covariance of the patch representations that exhibit corresponding relationships between different frames in a single video: $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\big]$, where $(\boldsymbol{z}_1, \boldsymbol{z}_2)$ denotes a pair of temporally aligned patch representations.

Definition 2 (Inter-video Covariance Matrix)**.** Define the inter-video covariance matrix as the covariance of the video-level representations across the dataset: $\bar{\boldsymbol{\Sigma}} = \mathbb{E}_{\bar{\boldsymbol{z}}_1,\bar{\boldsymbol{z}}_2}\big[(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^\top\big]$, where $\bar{\boldsymbol{z}}_i = \mathbb{E}_{\boldsymbol{z}\in f(\boldsymbol{V}_i)}\,[\boldsymbol{z}]$ denotes the mean representation of video $\boldsymbol{V}_i$.

Assumption 1 (Symmetric Operator)**.** Without loss of generality, $\boldsymbol{W}$, $\boldsymbol{\Sigma}$ and $\bar{\boldsymbol{\Sigma}}$ are constrained to be symmetric ($\boldsymbol{W}^\top = \boldsymbol{W}, \boldsymbol{\Sigma}^\top = \boldsymbol{\Sigma}, \bar{\boldsymbol{\Sigma}}^\top = \bar{\boldsymbol{\Sigma}}$). This is justified because any optimal $\boldsymbol{W}$ can be symmetrized without increasing (3).

Assumption 2 (Positive Semi-definite)**.** The transformation operator $\boldsymbol{W}$, the intra-video covariance matrix $\boldsymbol{\Sigma}$, and the inter-video covariance matrix $\bar{\boldsymbol{\Sigma}}$ are positive semi-definite: $\boldsymbol{W} \succeq 0, \boldsymbol{\Sigma} \succeq 0, \bar{\boldsymbol{\Sigma}} \succeq 0$. This ensures all eigenvalues are non-negative.

Assumption 3 (Commutative Minimizer)**.** we restrict the analysis to real symmetric commuting pairs $(\boldsymbol{W}, \boldsymbol{\Sigma})$ and $(\boldsymbol{W}, \bar{\boldsymbol{\Sigma}})$, *i.e.*, $\boldsymbol{\Sigma}\boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Sigma}$ and $\bar{\boldsymbol{\Sigma}}\boldsymbol{W} = \boldsymbol{W}\bar{\boldsymbol{\Sigma}}$. This allows simultaneous diagonalization with a common orthogonal basis $\boldsymbol{U}$, yielding $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}_W\boldsymbol{U}^\top$, $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}_\Sigma\boldsymbol{U}^\top$, and $\bar{\boldsymbol{\Sigma}} = \boldsymbol{U}\boldsymbol{\Lambda}_{\bar{\Sigma}}\boldsymbol{U}^\top$, where $\boldsymbol{\Lambda}_W, \boldsymbol{\Lambda}_\Sigma, \boldsymbol{\Lambda}_{\bar{\Sigma}}$ denote corresponding eigenvalue matrices.

B.1.2. Settings for MLP Projection

For the MLP projection $g(\boldsymbol{z}) = \boldsymbol{W}_2\phi(\boldsymbol{W}_1\boldsymbol{z})$, the optimization objective can be reformulated as:

$$\begin{aligned}\min_g \; M(g) = &\frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|g(\boldsymbol{z}_1) - g(\boldsymbol{z}_2)\|^2\big] \\ &+ \frac{\lambda}{2}\mathbb{E}_{\boldsymbol{z}_i}\big[\|\boldsymbol{J}_g(\boldsymbol{z}_i)\boldsymbol{J}_g(\boldsymbol{z}_i)^\top - \boldsymbol{I}\|_F^2\big].\end{aligned} \tag{5}$$

To facilitate the analysis, we begin by introducing a set of assumptions analogous to those in Case **i)**. Specifically, we replace the matrix $\boldsymbol{W}$ in Assumptions 1 to 3 with $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$, respectively. In addition to these modifications, we introduce the following additional assumptions:

Assumption 4 (Gaussian Distribution)**.** Without loss of generality, we assume that each patch representation $\boldsymbol{z}_i \in \mathbb{R}^d$

is independently drawn from a multivariate Gaussian distribution: $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. This assumption is justified by the observation that patch-level features extracted from natural videos tend to exhibit approximately Gaussian behavior due to the high-dimensional embedding and the central limit effect. Consequently, $\mathbb{E}_{\boldsymbol{z}_i}[\boldsymbol{z}_i \boldsymbol{z}_i^\top] = \boldsymbol{\Sigma}$.

Assumption 5 (Linear Approximation)**.** Assuming most values fall within the near-linear region of the `tanh` activation, we adopt the approximation $\phi(\boldsymbol{U}\boldsymbol{x}) \approx \boldsymbol{U}\phi(\boldsymbol{x})$, where $\boldsymbol{U}$ is an orthogonal matrix and $\phi(\cdot) = \tanh(\cdot)$ is applied element-wise.

B.1.3. Proof for Theorem 1

Based on the settings above, we establish the following theorem, which characterizes the spectral properties of the optimal solution in both cases and illustrates how the projection affects the quality of the transferred representation.

Theorem 1 (Spectral Properties of Optimal Projections, Formal)**.** *Denote the intra-video covariance matrix as* $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\big]$*. Let* $\{\sigma_i\}_{i=1}^d$ *be the eigenvalues of* $\boldsymbol{\Sigma}$*.*

For case i), assume symmetric matrices $\boldsymbol{W}$ *and* $\boldsymbol{\Sigma}$ *are positive semi-definite and mutually commuting. Let* $\{\mu_i\}_{i=1}^d$ *be the eigenvalues of* $\boldsymbol{W}$*. Then the eigenvalues of the optimal projection* $\boldsymbol{W}^\star$ *obey:*

$$\mu_i^\star = \begin{cases} 0, & \sigma_i > 2\lambda, \\ \sqrt{1 - \dfrac{\sigma_i}{2\lambda}}, & \sigma_i \le 2\lambda. \end{cases} \tag{6}$$

For case ii), assume $\phi(u\boldsymbol{z}_i) \approx u\phi(\boldsymbol{z}_i)$ *holds for* $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$*, and that symmetric matrices* $\boldsymbol{W}_1$*,* $\boldsymbol{W}_2$*,* $\boldsymbol{\Sigma}$ *are positive semi-definite and mutually commuting. Let* $\{\mu_{1,i}\}_{i=1}^d$ *and* $\{\mu_{2,i}\}_{i=1}^d$ *be the eigenvalues of* $\boldsymbol{W}_1$ *and* $\boldsymbol{W}_2$*, respectively. Then the eigenvalues of the optimal projections* ${\boldsymbol{W}_1}^\star$*,* ${\boldsymbol{W}_2}^\star$ *satisfy:*

$$\mu_{1,i}^\star \mu_{2,i}^\star = \begin{cases} 0, & \sigma_i > 2\lambda, \\ \sqrt{1 - \dfrac{\sigma_i}{2\lambda}}, & \sigma_i \le 2\lambda. \end{cases} \tag{7}$$

Proof. We first derive the **case i)** for the optimization objective of linear projection:

$$\begin{aligned} M(\boldsymbol{W}) = &\underbrace{\frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[\|\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2\|^2\big]}_{\text{Term A}} \\ &+ \underbrace{\frac{\lambda}{2}\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{I}\|_F^2}_{\text{Term B}}. \end{aligned} \tag{8}$$

The Term A can be derived as:

$$\begin{aligned} &\text{Term A} \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[\|\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2\|^2\big] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[(\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2)^\top(\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2)\big] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[\mathrm{Tr}\big((\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2)(\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2)^\top\big)\big] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[\mathrm{Tr}\big(\boldsymbol{W}(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\boldsymbol{W}^\top\big)\big] \\ &= \frac{1}{2}\,\mathrm{Tr}\big(\boldsymbol{W}^\top\boldsymbol{W}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\big)\big]\big) \\ &= \frac{1}{2}\,\mathrm{Tr}\big(\boldsymbol{W}^\top\boldsymbol{W}\boldsymbol{\Sigma}\big) \\ &= \frac{1}{2}\,\mathrm{Tr}\big((\boldsymbol{U}\boldsymbol{\Lambda}_W\boldsymbol{U}^\top)^\top(\boldsymbol{U}\boldsymbol{\Lambda}_W\boldsymbol{U}^\top)(\boldsymbol{U}\boldsymbol{\Lambda}_\Sigma\boldsymbol{U}^\top)\big) \\ &= \frac{1}{2}\,\mathrm{Tr}\big(\boldsymbol{\Lambda}_W^2\boldsymbol{\Lambda}_\Sigma\big). \end{aligned} \tag{9}$$

The derivation in Eq. (9) converts the squared ℓ_2 norm into a matrix trace and further reduces it to a product of eigenvalues via orthogonal decomposition.

The Term B can be derived as:

$$\begin{aligned} &\text{Term B} \\ &= \frac{\lambda}{2}\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{I}\|_F^2 \\ &= \frac{\lambda}{2}\,\mathrm{Tr}\big((\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{I})(\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{I})^\top\big) \\ &= \frac{\lambda}{2}\,\mathrm{Tr}\big((\boldsymbol{W}^4 - 2\boldsymbol{W}^2 + \boldsymbol{I})\big) \\ &= \frac{\lambda}{2}\,\mathrm{Tr}\big((\boldsymbol{U}\boldsymbol{\Lambda}_W\boldsymbol{U}^\top)^4 - 2(\boldsymbol{U}\boldsymbol{\Lambda}_W\boldsymbol{U}^\top)^2 + \boldsymbol{I}\big) \\ &= \frac{\lambda}{2}\,\mathrm{Tr}\big(\boldsymbol{\Lambda}_W^4 - 2\boldsymbol{\Lambda}_W^2\big) + \frac{\lambda d}{2}. \end{aligned} \tag{10}$$

Similarly, this derivation transforms the Frobenius norm into a matrix trace and reduces it to a function of eigenvalues via orthogonal decomposition.

Then the original objective can be rewritten as:

$$\begin{aligned} M(\boldsymbol{W}) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\big[\|\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2\|^2\big] \\ &\quad + \frac{\lambda}{2}\|\boldsymbol{W}\boldsymbol{W}^\top - \boldsymbol{I}\|_F^2 \\ &= \frac{1}{2}\,\mathrm{Tr}\big(\boldsymbol{\Lambda}_W^2\boldsymbol{\Lambda}_\Sigma\big) + \frac{\lambda}{2}\,\mathrm{Tr}\big(\boldsymbol{\Lambda}_W^4 - 2\boldsymbol{\Lambda}_W^2\big) + \frac{\lambda d}{2} \\ &= \frac{1}{2}\sum_{i=1}^d ({\mu_i}^2\sigma_i) + \frac{\lambda}{2}\sum_{i=1}^d ({\mu_i}^4 - 2{\mu_i}^2) + \frac{\lambda d}{2}. \end{aligned} \tag{11}$$

Differentiating Eq. (11) w.r.t. μ_i gives:

$$\begin{aligned} \frac{\partial M}{\partial \mu_i} &= \mu_i\sigma_i + 2\lambda\mu_i^3 - 2\lambda\mu_i \\ &= \mu_i(2\lambda\mu_i^2 - (2\lambda - \sigma_i)). \end{aligned} \tag{12}$$

Setting the derivative of the objective function to zero yields two possible solutions for each eigenvalue μ_i:

- $\mu_i = 0$. This is always a solution. It is optimal whenever the cubic term renders the quartic penalization unnecessary, *i.e.*, when $\sigma_i > 2\lambda$.
- $2\lambda\mu_i^2 - (2\lambda - \sigma_i) = 0$. Solving for μ_i yields the non-zero stationary points, which exist precisely when $\sigma_i \leq 2\lambda$.

In summary, the objective Eq. (8) reaches its minimum at $\boldsymbol{W}^\star = \boldsymbol{U}\,\mathrm{diag}\big(\mu_1^\star, \ldots, \mu_d^\star\big)\boldsymbol{U}^\top$, where the eigenvalues of the optimal projection $\boldsymbol{W}^\star$ obey:

$$\mu_i^\star = \begin{cases} 0, & \sigma_i \geq 2\lambda, \\ \sqrt{1 - \dfrac{\sigma_i}{2\lambda}}, & \sigma_i < 2\lambda. \end{cases} \tag{13}$$

Afterward, we derive the **case ii)** for the optimization objective of MLP projection:

$$\begin{aligned} \min_g \; M(g) = & \underbrace{\frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|g(\boldsymbol{z}_1) - g(\boldsymbol{z}_2)\|^2\big]}_{\text{Term C}} \\ & + \underbrace{\frac{\lambda}{2}\mathbb{E}_{\boldsymbol{z}_i}\big[\|\boldsymbol{J}_g(\boldsymbol{z}_i)\boldsymbol{J}_g(\boldsymbol{z}_i)^\top - \boldsymbol{I}\|_F^2\big]}_{\text{Term D}}. \end{aligned} \tag{14}$$

The Term C can be derived as:

$$\begin{aligned} &\text{Term C} \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|g(\boldsymbol{z}_1) - g(\boldsymbol{z}_2)\|^2\big] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|\boldsymbol{W}_2\big(\phi(\boldsymbol{W}_1\boldsymbol{z}_1) - \phi(\boldsymbol{W}_2\boldsymbol{z}_2)\big)\|^2\big] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|\big(\boldsymbol{U}\boldsymbol{\Lambda}_2\boldsymbol{U}^\top\big) \\ &\qquad \big(\phi(\boldsymbol{U}\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_1) - \phi(\boldsymbol{U}\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_2)\big)\|^2\big]. \end{aligned} \tag{15}$$

Following Assumption 5, the MLP projection becomes $g(\boldsymbol{z}) = \boldsymbol{W}_2\phi(\boldsymbol{W}_1\boldsymbol{z}) \approx \boldsymbol{U}\boldsymbol{\Lambda}_2\phi(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z})$, we can continue to derive Term C:

$$\begin{aligned} &\text{Term C} \\ &\approx \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|\big(\boldsymbol{U}\boldsymbol{\Lambda}_2\boldsymbol{U}^\top\big)\boldsymbol{U} \\ &\qquad \big(\phi(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_1) - \phi(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_2)\big)\|^2\big] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|\boldsymbol{\Lambda}_2\big(\phi(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_1) - \phi(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_2)\big)\|^2\big]. \end{aligned} \tag{16}$$

Let $\boldsymbol{Q} = \boldsymbol{U}^\top\boldsymbol{z}$. Since $\boldsymbol{U}$ is orthogonal and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, it follows that $\boldsymbol{Q} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Based on this property, we have:

$$\begin{aligned} &\text{Term C} \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\|\boldsymbol{\Lambda}_2\big(\phi(\boldsymbol{\Lambda}_1\boldsymbol{Q}_1) - \phi(\boldsymbol{\Lambda}_1\boldsymbol{Q}_2)\big)\|^2\big] \\ &= \frac{1}{2}\sum_{i=1}^d \mu_{2,i}^2\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\big(\phi(\mu_{1,i}\boldsymbol{Q}_{1,i}) - \phi(\mu_{1,i}\boldsymbol{Q}_{2,i})\big)^2\big] \\ &\approx \frac{1}{2}\sum_{i=1}^d \mu_{2,i}^2\mu_{1,i}^2\mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\big[\big(\boldsymbol{Q}_{1,i} - \boldsymbol{Q}_{2,i}\big)^2\big] \\ &= \frac{1}{2}\sum_{i=1}^d \mu_{2,i}^2\mu_{1,i}^2\sigma_i. \end{aligned} \tag{17}$$

Next, we consider the derivation of the Term D. Note that $\boldsymbol{J}_g$ is the Jacobian matrix of $g(\boldsymbol{z}_i)$, it can be formulated as:

$$\begin{aligned} \boldsymbol{J}_g(\boldsymbol{z}_i) &\approx \boldsymbol{U}\boldsymbol{\Lambda}_2\phi(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_i) \\ &= \boldsymbol{U}\boldsymbol{\Lambda}_2\phi'(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_i)\boldsymbol{\Lambda}_1\boldsymbol{U}^\top \\ &= \boldsymbol{U}\boldsymbol{\Lambda}_2\big(1 - \phi^2(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_i)\big)\boldsymbol{\Lambda}_1\boldsymbol{U}^\top. \end{aligned} \tag{18}$$

Therefore, the Term D can be derived as:

$$\begin{aligned} &\text{Term D} \\ &= \frac{\lambda}{2}\mathbb{E}_{\boldsymbol{z}_i}\big[\|\boldsymbol{J}_g(\boldsymbol{z}_i)\boldsymbol{J}_g(\boldsymbol{z}_i)^\top - \boldsymbol{I}\|_F^2\big] \\ &= \frac{\lambda}{2}\mathbb{E}_{\boldsymbol{z}_i}\big[\|\big(\boldsymbol{U}\boldsymbol{\Lambda}_2\big(1 - \phi^2(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_i)\big)\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\big) \\ &\qquad \big(\boldsymbol{U}\boldsymbol{\Lambda}_2\big(1 - \phi^2(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_i)\big)\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\big)^\top - \boldsymbol{I}\|_F^2\big] \\ &= \frac{\lambda}{2}\mathbb{E}_{\boldsymbol{z}_i}\big[\|\boldsymbol{\Lambda}_2\big(1 - \phi^2(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_i)\big)\boldsymbol{\Lambda}_1\boldsymbol{\Lambda}_1^\top \\ &\qquad \big(1 - \phi^2(\boldsymbol{\Lambda}_1\boldsymbol{U}^\top\boldsymbol{z}_i)\big)^\top\boldsymbol{\Lambda}_2^\top - \boldsymbol{I}\|_F^2\big] \\ &= \frac{\lambda}{2}\sum_{i=1}^d \big(\mu_{1,i}^2\mu_{2,i}^2\mathbb{E}_{\boldsymbol{z}_i}\big[\big(1 - \phi^2(\mu_{1,i}\boldsymbol{Q}_i)\big)^2\big] - 1\big)^2. \end{aligned} \tag{19}$$

Similarly, based on Assumption 5, we can approximately move the coefficient of $\boldsymbol{z}_i$ outside the activation function $\phi(\cdot)$, leading to the following transformation into the function of eigenvalues:

$$\begin{aligned} &\text{Term D} \\ &\approx \frac{\lambda}{2}\sum_{i=1}^d \big(\mu_{1,i}^2\mu_{2,i}^2\big(1 - \mu_{1,i}^2\mathbb{E}_{\boldsymbol{z}_i}\big[\boldsymbol{Q}_i^\top\boldsymbol{Q}_i\big]\big)^2 - 1\big)^2 \\ &= \frac{\lambda}{2}\sum_{i=1}^d \big(\mu_{1,i}^2\mu_{2,i}^2\big(1 - 2\mu_{1,i}^2\mathbb{E}_{\boldsymbol{z}_i}\big[\boldsymbol{Q}_i^\top\boldsymbol{Q}_i\big] \\ &\qquad + \mathbb{E}_{\boldsymbol{z}_i}\big[(\boldsymbol{Q}_i^\top\boldsymbol{Q}_i)^\top(\boldsymbol{Q}_i^\top\boldsymbol{Q}_i)\big]\big) - 1\big)^2 \\ &= \frac{\lambda}{2}\sum_{i=1}^d \big(\mu_{1,i}^2\mu_{2,i}^2\big(1 - 2\sigma_i\mu_{1,i}^2 + 3\sigma_i^2\mu_{1,i}^4\big) - 1\big)^2. \end{aligned} \tag{20}$$

Then the original objective can be rewritten as:

$$M(g) = \frac{1}{2}\sum_{i=1}^{d} \mu_{2,i}^2 \mu_{1,i}^2 \sigma_i + \frac{\lambda}{2}\sum_{i=1}^{d}\left(\mu_{1,i}^2\mu_{2,i}^2\left(1-2\sigma_i\mu_{1,i}^2+3\sigma_i^2\mu_{1,i}^4\right)-1\right)^2. \tag{21}$$

For simplicity, we assume $\mu_{1,i} \ll 1$, which is a reasonable approximation at initialization when using Kaiming [35] or Xavier [29] schemes. Under this setting, the term $1-2\sigma_i\mu_{1,i}^2+3\sigma_i^2\mu_{1,i}^4$ approaches 1, leading to the following simplification:

$$M(g) = \frac{1}{2}\sum_{i=1}^{d} \mu_{2,i}^2 \mu_{1,i}^2 \sigma_i + \frac{\lambda}{2}\sum_{i=1}^{d}\left(\mu_{1,i}^2\mu_{2,i}^2-1\right)^2. \tag{22}$$

Differentiating Eq. (22) w.r.t. $\mu_{1,i}$ gives:

$$\begin{aligned}\frac{\partial M}{\partial \mu_{1,i}} &= \mu_{1,i}\mu_{2,i}^2\sigma_i + 2\lambda\left(\mu_{1,i}^2\mu_{2,i}^2-1\right)\mu_{1,i}\mu_{2,i}^2 \\ &= \mu_{1,i}\mu_{2,i}^2\left(\sigma_i + 2\lambda\left(\mu_{1,i}^2\mu_{2,i}^2-1\right)\right).\end{aligned} \tag{23}$$

Setting the derivative to zero produces two cases:

- $\mu_{1,i} = 0$ when $\sigma_i > 2\lambda$.
- $\mu_{1,i} = \frac{1}{\mu_{2,i}}\sqrt{1-\frac{\sigma_i}{2\lambda}}$. Solving for $\mu_{1,i}$ yields the non-zero stationary points, which exist precisely when $\sigma_i \le 2\lambda$.

Differentiating Eq. (22) w.r.t. $\mu_{2,i}$ gives:

$$\begin{aligned}\frac{\partial M}{\partial \mu_{2,i}} &= \mu_{2,i}\mu_{1,i}^2\sigma_i + 2\lambda\left(\mu_{2,i}^2\mu_{1,i}^2-1\right)\mu_{2,i}\mu_{1,i}^2 \\ &= \mu_{2,i}\mu_{1,i}^2\left(\sigma_i + 2\lambda\left(\mu_{2,i}^2\mu_{1,i}^2-1\right)\right).\end{aligned} \tag{24}$$

Setting the derivative to zero produces two cases:

- $\mu_{2,i} = 0$ when $\sigma_i > 2\lambda$.
- $\mu_{2,i} = \frac{1}{\mu_{1,i}}\sqrt{1-\frac{\sigma_i}{2\lambda}}$. Solving for $\mu_{2,i}$ yields the non-zero stationary points, which exist precisely when $\sigma_i \le 2\lambda$.

In summary, the objective Eq. (14) reaches its minimum at $\boldsymbol{W}_1^\star = \boldsymbol{U}\,\mathrm{diag}\left(\mu_{1,1}^\star,\ldots,\mu_{1,d}^\star\right)\boldsymbol{U}^\top$, $\boldsymbol{W}_2^\star = \boldsymbol{U}\,\mathrm{diag}\left(\mu_{2,1}^\star,\ldots,\mu_{2,d}^\star\right)\boldsymbol{U}^\top$, where the eigenvalues of the optimal projection $\boldsymbol{W}_1^\star, \boldsymbol{W}_2^\star$ obey:

$$\mu_{1,i}^\star\mu_{2,i}^\star = \begin{cases} 0, & \sigma_i > 2\lambda, \\ \sqrt{1-\frac{\sigma_i}{2\lambda}}, & \sigma_i \le 2\lambda. \end{cases} \tag{25}$$

This completes the proof. □

B.2. Trade-off Improvement

Since both cases in Sec. B.2 yield similar spectral effects, we conduct the theoretical analysis based on a linear layer. To this end, we first provide the justification of the existence of the trade-off between temporal consistency and semantic separability. Subsequently, we define the distance-based metrics to quantify the two competing objectives and then present a theorem that reveals how the margin evolves after applying the optimal projection.

Lemma 1 (Trade-off between temporal consistency and semantic separability)**.** *For the objective $M(\boldsymbol{W})$ consisting of a temporal consistency term and a semantic separability term, the gradients of these two terms induce opposing directions in a certain parameter space. This misalignment indicates an inherent trade-off between temporal consistency and semantic separability when optimizing $M(\boldsymbol{W})$.*

Proof. According to Eq. (11), the objective $M(\boldsymbol{W})$ of our method can be derived as:

$$\begin{aligned} M(\boldsymbol{W}) &= \underbrace{\frac{1}{2}\mathbb{E}_{z_1,z_2}\left[\|\boldsymbol{W}z_1-\boldsymbol{W}z_2\|^2\right]}_{\text{Temporal Consistency}} + \underbrace{\frac{\lambda}{2}\|\boldsymbol{W}\boldsymbol{W}^\top - I\|_F^2}_{\text{Semantic Separability}} \\ &= \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Lambda}_W^2\boldsymbol{\Lambda}_\Sigma) + \frac{\lambda}{2}\mathrm{Tr}({\boldsymbol{\Lambda}_W}^4 - 2\boldsymbol{\Lambda}_W^2) + \frac{\lambda d}{2} \\ &= \frac{1}{2}\sum_{i=1}^{d}({\mu_i}^2\sigma_i) + \frac{\lambda}{2}\sum_{i=1}^{d}({\mu_i}^4 - 2{\mu_i}^2) + \frac{\lambda d}{2}. \end{aligned} \tag{26}$$

According to the Assumption 2 , $\boldsymbol{W}$ and $\boldsymbol{\Sigma}$ are positive semi-definite, implying that μ_i and σ_i are non-negative. Therefore, by differentiating $M(\boldsymbol{W})$ w.r.t. μ_i gives:

$$\frac{\partial M}{\partial \mu_i} = \underbrace{\mu_i\sigma_i}_{\text{Temporal Consistency}} + \underbrace{2\lambda\mu_i^3 - 2\lambda\mu_i}_{\text{Semantic Separability}}. \tag{27}$$

Based on the formulation in Eq. (27), the gradient of these two derived terms can be inferred as:

- Temporal Consistency: $\mu_i\sigma_i \ge 0$.
- Semantic Separability: $2\lambda\mu_i^3 - 2\lambda\mu_i = 2\lambda\mu_i(\mu_i+1)(\mu_i-1) < 0$ when $\mu_i < 1$.

Therefore, the two terms may change in opposite directions during optimization, since updates that increase temporal consistency tend to decrease semantic separability in the same eigen-direction, and vice versa. This behavior reflects an inherent trade-off between temporal consistency and semantic separability in our objective. A similar argument can be made for the trade-off in the MLP case.

□

Definition 3 (Intra-video Distance)**.** Define the intra-video distance as $D_{intra}(\boldsymbol{z}_1,\boldsymbol{z}_2) = \mathbb{E}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\left[\|\boldsymbol{z}_1-\boldsymbol{z}_2\|^2\right]$, which measures the average distance between temporally corresponding patches within a video.

Definition 4 (Inter-video Distance)**.** Define the inter-video distance as $D_{inter}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[\|\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2\|^2\right]$, calculating the average distance between video-level representations, where $\bar{\boldsymbol{z}}_i = \mathbb{E}_{\boldsymbol{z} \in f(\boldsymbol{V}_i)}[\boldsymbol{z}]$ is the mean representation of the video $\boldsymbol{V}_i$. This reflects the average distance between different video-level representations.

Definition 5 (Distance Margin)**.** Define the margin of these two metrics as $D(\boldsymbol{z}_1, \boldsymbol{z}_2) = D_{inter}(\boldsymbol{z}_1, \boldsymbol{z}_2) - \gamma D_{intra}(\boldsymbol{z}_1, \boldsymbol{z}_2)$, reflecting the degree of separation between the two properties, where a larger value indicates a better trade-off between the two objectives.

Assumption 6 (Mean Eigenvalue Approximation)**.** The eigenvalues of the inter-video covariance matrix approximate the average of those of the intra-video covariance matrix, *i.e.*, $\forall j,\ \tau_j = \frac{1}{d}\sum_{i=1}^{d}\sigma_i$.

Theorem 2 (Trade-off Improvement, Formal)**.** *Let* $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\right]$, $\bar{\boldsymbol{\Sigma}} = \mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^\top\right]$ *be the intra-video and inter-video covariance matrices, with eigenvalues* $\{\sigma_i\}_{i=1}^d$ *and* $\{\tau_i\}_{i=1}^d$*, respectively. Assume symmetric matrices* $\boldsymbol{W}$*,* $\boldsymbol{\Sigma}$*, and* $\bar{\boldsymbol{\Sigma}}$ *are positive semi-definite and mutually commuting, and that* $\forall j,\ \tau_j = \frac{1}{d}\sum_{i=1}^{d}\sigma_i = \tau$*. Let* $\{\mu_i\}_{i=1}^d$ *be the eigenvalues of* $\boldsymbol{W}$*. For the linear projection* $g(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$*, where the optimal eigenvalues are given by* $\mu_i^\star = \sqrt{1 - \frac{\sigma_i}{2\lambda}}$ *for* $\sigma_i \le 2\lambda$*, the improvement in the margin metric is: given by:*

$$\begin{aligned}\Delta &= D(g(\boldsymbol{z}_1), g(\boldsymbol{z}_2)) - D(\boldsymbol{z}_1, \boldsymbol{z}_2) \\ &= \sum_{\sigma_i \le 2\lambda}(\tau - \sigma_i)\left(1 - \frac{\sigma_i}{2\lambda}\right) > 0.\end{aligned} \tag{28}$$

Proof. The margin metric of intra-video distance between the projected representations and the original representations can be derived as:

$$\begin{aligned}\Delta_{intra} &= D_{intra}(\boldsymbol{W}\boldsymbol{z}_1, \boldsymbol{W}\boldsymbol{z}_2) - \gamma D_{intra}(\boldsymbol{z}_1, \boldsymbol{z}_2) \\ &= \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\|\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2\|^2\right] \\ &\quad - \gamma\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|^2\right] \\ &= \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[(\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2)^\top(\boldsymbol{W}\boldsymbol{z}_1 - \boldsymbol{W}\boldsymbol{z}_2)\right] \\ &\quad - \gamma\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top(\boldsymbol{z}_1 - \boldsymbol{z}_2)\right] \\ &= \mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\mathrm{Tr}\big(\boldsymbol{W}(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\boldsymbol{W}^\top\big)\right] \\ &\quad - \gamma\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[\mathrm{Tr}\big((\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\big)\right] \\ &= \mathrm{Tr}\big(\boldsymbol{W}^\top\boldsymbol{W}\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\right]\big) \\ &\quad - \gamma\,\mathrm{Tr}\big(\mathbb{E}_{\boldsymbol{z}_1, \boldsymbol{z}_2}\left[(\boldsymbol{z}_1 - \boldsymbol{z}_2)(\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top\right]\big) \\ &= \mathrm{Tr}\big(\boldsymbol{\Lambda}_W^\top\boldsymbol{\Lambda}_W\boldsymbol{\Lambda}_\Sigma\big) - \gamma\,\mathrm{Tr}(\boldsymbol{\Lambda}_\Sigma) \\ &= \sum_{i=1}^{d}(\mu_i^2 - \gamma)\sigma_i.\end{aligned} \tag{29}$$

The margin metric of inter-video distance between the projected representations and the original representations can be derived as:

$$\begin{aligned}\Delta_{inter} &= D_{inter}(\boldsymbol{W}\boldsymbol{z}_1, \boldsymbol{W}\boldsymbol{z}_2) - D_{inter}(\boldsymbol{z}_1, \boldsymbol{z}_2) \\ &= \mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[\|\boldsymbol{W}\bar{\boldsymbol{z}}_1 - \boldsymbol{W}\bar{\boldsymbol{z}}_2\|^2\right] \\ &\quad - \gamma\mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[\|\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2\|^2\right] \\ &= \mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[(\boldsymbol{W}\bar{\boldsymbol{z}}_1 - \boldsymbol{W}\bar{\boldsymbol{z}}_2)^\top(\boldsymbol{W}\bar{\boldsymbol{z}}_1 - \boldsymbol{W}\bar{\boldsymbol{z}}_2)\right] \\ &\quad - \gamma\mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^\top(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)\right] \\ &= \mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[\mathrm{Tr}\big(\boldsymbol{W}(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^\top\boldsymbol{W}^\top\big)\right] \\ &\quad - \gamma\mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[\mathrm{Tr}\big((\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^\top\big)\right] \\ &= \mathrm{Tr}\big(\boldsymbol{W}^\top\boldsymbol{W}\mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^\top\right]\big) \\ &\quad - \gamma\,\mathrm{Tr}\big(\mathbb{E}_{\bar{\boldsymbol{z}}_1, \bar{\boldsymbol{z}}_2}\left[(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)(\bar{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_2)^\top\right]\big) \\ &= \mathrm{Tr}\big(\boldsymbol{\Lambda}_W^\top\boldsymbol{\Lambda}_W\boldsymbol{\Lambda}_{\bar{\Sigma}}\big) - \gamma\,\mathrm{Tr}(\boldsymbol{\Lambda}_{\bar{\Sigma}}) \\ &= \sum_{i=1}^{d}(\mu_i^2 - \gamma)\tau_i.\end{aligned} \tag{30}$$

Then the improvement of the margin metrics can be formulated as:

$$\begin{aligned}\Delta &= (D_{inter}(\boldsymbol{W}\boldsymbol{z}_1, \boldsymbol{W}\boldsymbol{z}_2) - D_{intra}(\boldsymbol{W}\boldsymbol{z}_1, \boldsymbol{W}\boldsymbol{z}_2)) \\ &\quad - (D_{inter}(\boldsymbol{z}_1, \boldsymbol{z}_2) - D_{intra}(\boldsymbol{z}_1, \boldsymbol{z}_2)) \\ &= (D_{inter}(\boldsymbol{W}\boldsymbol{z}_1, \boldsymbol{W}\boldsymbol{z}_2) - D_{inter}(\boldsymbol{z}_1, \boldsymbol{z}_2)) \\ &\quad - (D_{intra}(\boldsymbol{W}\boldsymbol{z}_1, \boldsymbol{W}\boldsymbol{z}_2) - D_{intra}(\boldsymbol{z}_1, \boldsymbol{z}_2)) \\ &= \sum_{i=1}^{d}(\mu_i^2 - \gamma)\tau_i - \sum_{i=1}^{d}(\mu_i^2 - \gamma)\sigma_i \\ &= \sum_{i=1}^{d}(\mu_i^2 - \gamma)(\tau_i - \sigma_i).\end{aligned} \tag{31}$$

Under Assumptions 1 to 3, the optimal linear projection $\boldsymbol{W}^\star = \boldsymbol{U}\,\mathrm{diag}\big(\mu_1^\star, \ldots, \mu_d^\star\big)\boldsymbol{U}^\top$ has eigenvalues given by:

$$\mu_i^\star = \begin{cases} 0, & \sigma_i \ge 2\lambda, \\ \sqrt{1 - \frac{\sigma_i}{2\lambda}}, & \sigma_i < 2\lambda. \end{cases} \tag{32}$$

Under Assumption 6, due to $\forall j, \tau_j = \frac{1}{d}\sum_{i=1}^{d}\sigma_i = \tau$, we have $\sum_{i=1}^{d}(\tau - \sigma_i) = 0$.

By substituting $\mu_i^\star = \sqrt{1 - \frac{\sigma_i}{2\lambda}}$ and $\sum_{i=1}^{d}(\tau - \sigma_i) = 0$ under the condition $\sigma_i < 2\lambda$, the change in the margin metric can be expressed as:

$$\Delta = \sum_{i=1}^{d}(\mu_i^2 - \gamma)(\tau_i - \sigma_i) = \sum_{\sigma_i < 2\lambda}(\tau - \sigma_i)\left(1 - \frac{\sigma_i}{2\lambda}\right). \tag{33}$$

When $\lambda < \frac{\tau}{2}$ (*i.e.*, $2\lambda < \tau$), all $\sigma_i \leq 2\lambda$ necessarily obey $\sigma_i < \tau$. In this case, each term in Eq. (33) satisfies $(\tau - \sigma_i)\left(1 - \frac{\sigma_i}{2\lambda}\right) > 0$ because:

- $\tau - \sigma_i > 0$ follows from $\sigma_i < \tau$,
- $1 - \frac{\sigma_i}{2\lambda} \geq 0$ since $\sigma_i \leq 2\lambda$.

Therefore, $\Delta > 0$ holds whenever $\lambda < \frac{1}{2d}\sum_{i=1}^{d} \sigma_i$. □

In summary, this section provides a theoretical analysis of the trade-off between temporal consistency and semantic separability, leading to the following two key insights:

1. **Theorem 1** shows that linear projection exhibits similar behavior to shallow MLPs in adjusting representations, yielding comparable effects in similar feature scaling behavior as the linear layer.
2. **Theorem 2** demonstrates that under optimal conditions, a linear projection is sufficient to improve the trade-off between temporal consistency and semantic separability.

C. Supplementary Explanation of Method

C.1. Differences with Previous Methods

In Figure 1, we provide a comparative overview of several categories of video representation learning works alongside our method.

1) Video-pretrained methods extend the masked image modeling paradigm to the video domain by masking 3D volumes and reconstructing raw pixels for spatiotemporal learning [28, 82, 84]. Subsequent variants incorporate conditional frames to enhance temporal modeling [25, 32, 43, 58]. These approaches typically require large-scale video pretraining from scratch, incurring substantial computational cost due to video redundancy and pixel-level reconstruction overhead.

2) Supervised adaptation methods adapt Vision Transformers pretrained with CLIP [70] by inserting lightweight adapters in serial or parallel configurations [15, 55, 56, 63, 91]. These adapters are usually trained on supervised action recognition datasets [30, 46], making them highly task-dependent and less generalizable without additional task-specific fine-tuning.

3) Video fine-tuning methods follow a two-stage training scheme: models are first pretrained on task-specific datasets to learn static features for instance-level discrimination, then fine-tuned on video datasets with additional temporal branches introduced to handle motion reasoning [22, 38, 52, 53]. Although it can perform well on specific video tasks, its increased model complexity and training cost make it difficult to perform fast cross-domain transfer.

4) Our image-to-video transfer method takes a different approach by leveraging pretrained image representations and adapting them to video tasks via structure-preserving projection. The main advantages are as follows:

Algorithm 1: Consistency-Separability Trade-off Transfer Learning Algorithm

Input: Unlabeled dataset $\mathcal{D}$, number of iterations L, interpolation ratio α, constraint weight λ.
Output: Parameters $\boldsymbol{\theta}_{L+1}$ of projection layer g.

Initialize parameters $\boldsymbol{\theta}_1$ for g.
for $l = 1$ ***to*** L **do**
  Sample a batch of videos $\{\boldsymbol{V}_i\}_{i=1}^{B}$.
  for $i = 1$ ***to*** B **do**
    ▷ Temporal Correspondence Establishment
    Select frames $\boldsymbol{v}_{t_1}^{f}$, $\boldsymbol{v}_{t_2}$, $\boldsymbol{v}_{t_1}^{b}$ from $\boldsymbol{V}_m$ and prepare position encoding $\mathbf{E}_{\text{pos}}$ and $\widetilde{\mathbf{E}}_{\text{pos}}$.
    Extract representations $\boldsymbol{z}_{t_1}^{f}$, $\boldsymbol{z}_{t_2}$, $\widetilde{\boldsymbol{z}}_{t_1}^{b}$ with f and projections $\boldsymbol{p}_{t_1}^{f}$, $\boldsymbol{p}_{t_2}$, $\widetilde{\boldsymbol{p}}_{t_1}^{b}$ via g.
    Calculate correlation matrices $\boldsymbol{A}_{t_1}^{t_2}$ and $\widetilde{\boldsymbol{A}}_{t_2}^{t_1}$.
    ▷ Temporal Consistency and Semantic Separability Trade-off
    Enhance temporal consistency of $\boldsymbol{p}_{t_1}^{f}$, $\boldsymbol{p}_{t_2}$, $\widetilde{\boldsymbol{p}}_{t_1}^{b}$ via $\mathcal{L}_{cyc}$.
    Align the semantic separability of $\{(\boldsymbol{p}_{t_1}^{f}, \boldsymbol{z}_{t_1}^{f}), (\boldsymbol{p}_{t_2}, \boldsymbol{z}_{t_2}), (\widetilde{\boldsymbol{p}}_{t_1}^{b}, \widetilde{\boldsymbol{z}}_{t_1}^{b})\}$ by $\mathcal{L}_{reg}$.
  Update the projection layer g with $\mathcal{L}_{total} = \mathcal{L}_{cyc} + \lambda\mathcal{L}_{reg}$.
return

- **Efficient transfer**: We sample two frames per video and insert a lightweight linear-based projection head after a frozen image encoder, enabling fast transfer with reduced temporal and spatial cost.
- **Joint optimization**: We simultaneously optimize temporal consistency and semantic separability via a temporal cycle-consistency objective and a semantic separability regularization term.
- **Label-free training**: Our method is fully self-supervised, requiring no manual annotations, which enhances scalability and promotes better generalization across diverse video tasks of different granularity.

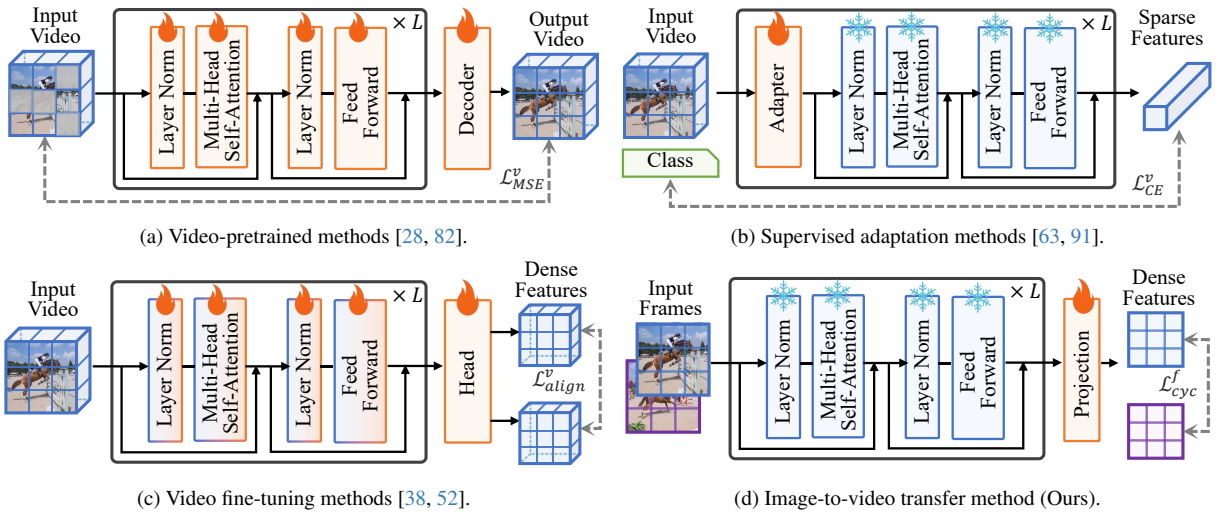
C.2. Algorithm of the Framework

The complete optimization procedure of our framework is summarized in Algorithm 1. The batch-level for-loop can be implemented via matrix operations to reduce computational burden.

D. Detailed Description of Experiments

D.1. Training Datasets

Kinetics-400 [46] is a widely used large-scale video benchmark comprising 400 human action categories collected from YouTube. It provides 239,789 trimmed video clips, each last-



(a) Video-pretrained methods [28, 82].

(b) Supervised adaptation methods [63, 91].

(c) Video fine-tuning methods [38, 52].

(d) Image-to-video transfer method (Ours).

Figure 1. Comparison of several categories of video representation learning methods with ours.

ing around 10 seconds, making it suitable for various video understanding tasks. In our experiments, we sample video frames at 2 FPS for pretraining to reduce redundancy while retaining sufficient temporal cues. In this work, unless otherwise noted, all the models equipped with our method are trained for 5 epochs using the Kinetics-400 training set.

SSV2 (Something-Something V2) [30] is a large-scale video benchmark emphasizing human-object interactions and temporal reasoning. It comprises 220,847 short, crowd-sourced clips across 174 action classes, with each clip lasting a few seconds, making it well-suited for evaluating temporal understanding beyond appearance cues. In our experiments, we only use SSV2 in the ablation study on training datasets, training for 5 epochs on the SSV2 training split.

ImageNet-1k [21] is a well-known image dataset containing over 1.28 million training images across 1,000 real-world object categories. It has played a central role in the development of deep visual representation learning and serves as the pretraining corpus for most high-performance image encoders. In this work, most of the image models are already pretrained on this dataset, providing a strong foundation of semantic separability.

WIT-400M (WebImageText) [70] is a large-scale web-crawled dataset consisting of 400 million image-text pairs, designed to support vision-language pretraining. The dataset was constructed using 500,000 diverse natural language queries to guide image-text pair retrieval, with up to 20,000 pairs per query to encourage approximate class balancing. Its overall scale and linguistic richness make it suitable for training multimodal models such as CLIP [70].

LAION-400M [74] is a large-scale dataset of 400 million image-text pairs designed to support vision-language pretraining. The image-text pairs are extracted from Common Crawl web pages and filtered using CLIP-based similarity to retain pairs with stronger semantic alignment between images and captions. It is used as a pretraining dataset for vision-language models, including BLIP [51].

D.2. Training Settings

During training, we freeze the pretrained image encoder f and update only the projection layer g. The training is performed on the Kinetics-400 dataset for 5 epochs with a total batch size of 512, using the first epoch for learning rate warm-up. We employ the AdamW optimizer [60] with a cosine learning rate decay schedule. The base learning rate is set to $blr = 1 \times 10^{-4}$ and scaled according to the batch size as $lr = blr/256$. For each video clip, two frames are randomly sampled with a temporal interval of $\delta = 0.15$ relative to the total video length. The softmax temperature is set to $\tau = 0.03$. The output dimension of the projection head g is set to $d = 768$ for ViT-Base backbones. Detailed hyperparameter settings for training and method components are summarized in Tab. 3a and Tab. 3b. All experiments are implemented in PyTorch [65] and conducted on a Linux server equipped with an AMD EPYC 9654 96-Core CPU and 4 NVIDIA RTX4090 GPUs.

D.3. Evaluation Settings

D.3.1. Evaluation on Dense-level Benchmarks

We first evaluate the representations on three dense video downstream tasks: video object segmentation on DAVIS-2017 [67], human part segmentation on VIP [96], and human pose propagation on JHMDB [44]. Following previous works [25, 32, 43, 58], all tasks are evaluated under a semi-supervised protocol in which the ground-truth mask of the first frame is given, and the model propagates predictions

Table 3. Summary of hyperparameter settings used during training and evaluation.

(a) Training hyperparameters.

Hyperparameter	Notation	Value
Image size	$H \times W$	224×224
Patch size	p	16
Optimizer	/	AdamW
Scheduler	/	Cosine
Weight decay	/	0.05
Momentum	β_1, β_2	$0.9, 0.95$
Base learning rate	blr	1×10^{-4}
Epochs	/	5
Warm-up Epoch	/	1
Batch size	bs	512

(b) Method hyperparameters.

Hyperparameter	Notation	Value
Temperature of Softmax	τ	0.03
Frame sampling interval	δ	0.15
Feature dim of g	d	768

(c) Evaluation hyperparameters.

Hyperparameter	DAVIS-2017	VIP	JHMDB
Image size	480×880	480×880	320×320
Top-K	7	10	7
Queue Length	20	20	20
Neighborhood Size	20	20	20

to subsequent frames without any task-specific fine-tuning. The hyperparameters used for each evaluation task are listed in Tab. 3c. To ensure fair comparisons, we keep the evaluation hyperparameter settings fixed across all methods and tasks without additional tuning.

DAVIS-2017 [67] is a widely used benchmark for video object segmentation. We report three standard metrics to assess overall segmentation quality:

1) $\mathcal{J}_{\mathrm{m}}$ (region similarity) computes the average IoU between the predicted mask P_i and the ground-truth mask G_i across all videos $\boldsymbol{V}_i$:

$$\mathcal{J}_{\mathrm{m}} = \frac{1}{n} \sum_{i=1}^{n} \frac{|P_i \cap G_i|}{|P_i \cup G_i|}. \tag{34}$$

2) $\mathcal{F}_{\mathrm{m}}$ (contour accuracy) evaluates the alignment between the predicted and ground-truth boundaries by calculating the harmonic mean of precision Pre_i and recall Rec_i:

$$\mathcal{F}_{\mathrm{m}} = \frac{1}{n} \sum_{i=1}^{n} \frac{2 \cdot Pre_i \cdot Rec_i}{Pre_i + Rec_i}. \tag{35}$$

3) $\mathcal{J}\&\mathcal{F}_{\mathrm{m}}$ provides an overall performance measure by averaging $\mathcal{J}_{\mathrm{m}}$ and $\mathcal{F}_{\mathrm{m}}$:

$$\mathcal{J}\&\mathcal{F}_{\mathrm{m}} = \frac{\mathcal{J}_{\mathrm{m}} + \mathcal{F}_{\mathrm{m}}}{2}. \tag{36}$$

VIP [96] focuses on fine-grained human part segmentation and is used to evaluate semantic part propagation. The main evaluation metric is the mIoU computed by averaging the IoU across all classes C_j and all videos $\boldsymbol{V}_i$:

$$\mathrm{mIoU} = \frac{1}{|C|} \sum_{j=1}^{|C|} \frac{1}{n} \sum_{i=1}^{n} \frac{|P_{i,j} \cap G_{i,j}|}{|P_{i,j} \cup G_{i,j}|}. \tag{37}$$

JHMDB [44] is commonly used for human pose estimation. We adopt it for the pose propagation task and evaluate performance using the PCK@k metric, which measures the proportion of keypoints predicted within a normalized distance threshold:

$$\mathrm{PCK@}k = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|S_i|} \sum_{j=1}^{|S_i|} \mathbb{1}\left[D(\hat{p}_{i,j}, p_{i,j}) < k \cdot d_i\right], \tag{38}$$

where S_i is the keypoint set in video $\boldsymbol{V}_i$, d_i denotes the scale of the human body, $D(\hat{p}_{i,j}, p_{i,j})$ is the Euclidean distance between the predicted and ground-truth positions, and k is the threshold for the maximum allowable distance error. We report PCK@0.1 and PCK@0.2 in our experiments.

D.3.2. Evaluation on Frame-/Video-level Benchmarks

We further evaluate the transferred models on several frame-level and video-level downstream tasks: temporal action localization on Breakfast [49], video retrieval on UCF101 and HMDB51 [48, 77], and action classification on Something-Something-v2 (SSV2) [30].

Breakfast [49] contains 1,712 untrimmed videos with frame-level annotations of fine-grained actions. We perform temporal action localization on this dataset by extracting frame-wise representations with our transferred image-to-video model and training the FACT [61] backbone on these representations. Following a standard protocol, we train FACT on *split2-4* and evaluate on *split1*. We report three standard metrics as follows:

1) Edit measures sequence-level similarity between the predicted and ground-truth label sequences after collapsing consecutive duplicates. Let $\mathbf{y} = (y_1, \ldots, y_T)$ and $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_T)$ be frame-wise labels, and let $\mathcal{C}(\cdot)$ collapse consecutive identical labels. Denote $\mathrm{Lev}(\cdot, \cdot)$ as the Levenshtein distance and $|\cdot|$ as the sequence length. The normalized edit score is

$$\mathrm{Edit} = 1 - \frac{\mathrm{Lev}\big(\mathcal{C}(\hat{\mathbf{y}}), \mathcal{C}(\mathbf{y})\big)}{\max\{\, |\mathcal{C}(\hat{\mathbf{y}})|, |\mathcal{C}(\mathbf{y})| \,\}}. \tag{39}$$

2) Acc is the frame-wise accuracy, representing the percentage of correctly labeled frames:

$$\mathrm{Acc} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{\hat{y}_t = y_t\}, \tag{40}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

3) F1@k is the segmental F1 at IoU threshold k. Let the ground-truth segment set be $\mathcal{S} = \{(s_j^g, e_j^g, c_j^g)\}$ and the predicted set $\hat{\mathcal{S}} = \{(s_i^p, e_i^p, c_i^p)\}$, where s/e are start/end frames and c is the class. For segments of the same class, define the temporal Intersection-over-Union as:

$$\mathrm{IoU}\big((s_i^p, e_i^p), (s_j^g, e_j^g)\big) = \frac{\big\{\min(e_i^p, e_j^g) - \max(s_i^p, s_j^g)\big\}_+}{\max(e_i^p, e_j^g) - \min(s_i^p, s_j^g)}. \tag{41}$$

A prediction is a true positive (*TP*) if it uniquely matches a ground-truth segment of the same class with $\mathrm{IoU} \geq k$; unmatched predictions are false positives (*FP*), and unmatched ground-truth segments are false negatives (*FN*). With precision $P = \frac{TP}{TP+FP}$ and recall $R = \frac{TP}{TP+FN}$, we compute

$$\mathrm{F1@k} \;=\; \frac{2\,P \cdot R}{P + R}, \tag{42}$$

where $k \in \{0.10, 0.25, 0.50\}$ as standard thresholds.

UCF101 [77] comprises 13,320 videos from 101 human action classes, and **HMDB51** [48] contains 6,766 videos from 51 action classes. For zero-shot video retrieval on the test set, we directly extract video representations using our transferred image-to-video model and perform retrieval following [57]: in each query round, one video is treated as the query and all remaining videos form the reference set. This process is repeated for every video. And we report the following metrics with the average.

1) mAP (Mean Average Precision) is the mean of per-query Average Precision (AP). Let $|\mathcal{Q}|$ be the number of queries, n_j the number of positives for query j, and r_i the rank of the i-th retrieved positive for that query. Then

$$\mathrm{mAP} \;=\; \frac{1}{|\mathcal{Q}|}\sum_{j=1}^{|\mathcal{Q}|}\frac{1}{n_j}\sum_{i=1}^{n_j}\frac{i}{r_i}. \tag{43}$$

2) Recall@K is the fraction of queries for which at least one positive appears in the top-K results. Let $\mathcal{R}_j^{(K)}$ be the set of ranks $\leq K$ among retrieved items for query j, and let $\mathcal{P}_j$ be the set of ranks of its positives. Then

$$\mathrm{Recall@}K \;=\; \frac{1}{|\mathcal{Q}|}\sum_{j=1}^{|\mathcal{Q}|}\mathbb{1}\{\min(\mathcal{P}_j) \leq K\}. \tag{44}$$

Something-Something-v2 (SSV2) [30] is a large-scale action classification benchmark consisting of 220,847 short videos from 174 fine-grained action categories without public labels. It focuses on human-object interactions with subtle motion variations, and is widely used to evaluate a model's capability for temporal reasoning and motion-sensitive action understanding.

For action classification on SSV2, each transferred image-to-video model is fine-tuned on the training set for 25 epochs and then evaluated on the validation set using single-clip sampling. Although this protocol is lighter than commonly used longer-schedule or multi-clip settings, it is applied uniformly to all compared methods to ensure effective and fair comparison. We report the standard top-k accuracy metric: Acc@k measures the percentage of validation videos whose ground-truth label appears among the top-k predicted classes. Let $\mathbf{z}^{(i)} \in \mathbb{R}^C$ be the predicted logits for the i-th video over C classes, and let $y_i \in \{1, \ldots, C\}$ be the ground-truth label. Denote by $\pi_k(\mathbf{z}^{(i)})$ the set of indices corresponding to the top-k largest entries in $\mathbf{z}^{(i)}$. Then

$$\mathrm{Acc@}k \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\Big\{y_i \in \pi_k\Big(\mathbf{z}^{(i)}\Big)\Big\}, \tag{45}$$

where N is the number of validation videos and $\mathbb{1}\{\cdot\}$ is the indicator function. Following common practice to present Acc@1 and Acc@5.

Chiral SSV2 [4] is a temporal order discrimination benchmark constructed from Something-Something-v2 [30]. It groups temporally opposite actions into chiral pairs, such as "sitting down" and "standing up", and evaluates whether a video representation is sensitive to the ordering of visual change over time. Compared with standard action classification, this benchmark places stronger emphasis on time-awareness rather than semantic categorization.

Following [4], we evaluate each model using a linear-probe protocol on frozen representations. Specifically, for each chiral group, we extract frame-level representations from each video, concatenate representations along the temporal dimension to form the video representation, and train a linear classifier for binary classification. This procedure is repeated independently for every chiral group, and the final result is reported as the average classification accuracy across all groups.

Acc measures the percentage of correctly classified videos over all evaluation samples. Let $\mathbf{z}^{(i)} \in \mathbb{R}^2$ be the logits predicted by the linear classifier for the i-th video, and let $y_i \in \{0, 1\}$ denote its ground-truth label within the corresponding chiral pair. Then

$$\mathrm{Acc} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\Big\{\arg\max_c z_c^{(i)} = y_i\Big\}, \tag{46}$$

where N is the total number of evaluation videos and $\mathbb{1}\{\cdot\}$ is the indicator function.

D.3.3. Distance-based Trade-off Metrics Validation

To provide an interpretable assessment, we validate the distance-based metrics proposed in Sec.4. We randomly sample 1,000 videos from the Kinetics-400 validation set and compute each metric for both the original image-pretrained models and our transferred models. All metrics are computed on the same sample set for a fair comparison. We report four metrics as follows.

1) D_{inter} (Inter-video distance). Let the sample set be $\mathcal{V} = \{V^{(n)}\}_{n=1}^{M}$ with $M = 1000$. For each $V^{(n)}$, select the middle frame $v_{t^*}^{(n)}$ and extract N patch representations $\{\boldsymbol{z}_{t^*}^{(n)}(i)\}_{i=1}^{N}$. For each unordered pair (u, v) with $u < v$, define the pair-wise inter-video distance as:

$$d(u,v) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big\| \boldsymbol{z}_{t^*}^{(u)}(i) - \boldsymbol{z}_{t^*}^{(v)}(i) \Big\|_2. \tag{47}$$

The unnormalized inter-video distance is

$$D_{inter}^{ori} \;=\; \frac{2}{M(M-1)} \sum_{1\le u<v\le M} d(u,v). \tag{48}$$

Let the global *center* be the mean patch representation over videos, $\boldsymbol{c}(i) = \frac{1}{M}\sum_{n=1}^{M} \boldsymbol{z}_{t^*}^{(n)}(i)$, and define each video's distance to the center as

$$r(u) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big\| \boldsymbol{z}_{t^*}^{(u)}(i) - \boldsymbol{c}(i) \Big\|_2. \tag{49}$$

The inter-video radius is $R_{inter} = \max_u r(u)$, and the normalized metric is

$$D_{inter} \;=\; \frac{D_{inter}^{ori}}{2R_{inter}}. \tag{50}$$

2) D_{intra} (Intra-video distance). For each $V^{(n)}$, select a set of frame pairs $\mathcal{P}^{(n)} = \{(t_a, t_b)\}$. For a given pair, measure pair-wise intra-video distance as:

$$d^{(n)}(t_a,t_b) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big\| \boldsymbol{z}_{t_a}^{(n)}(i) - \boldsymbol{z}_{t_b}^{(n)}(i) \Big\|_2. \tag{51}$$

The per-video unnormalized intra distance and its normalization radius are

$$\begin{aligned} D_{intra}^{ori,(n)} \;&=\; \frac{1}{|\mathcal{P}^{(n)}|} \sum_{(t_a,t_b)\in\mathcal{P}^{(n)}} d^{(n)}(t_a,t_b), \\ R_{intra}^{(n)} \;&=\; \max_t \frac{1}{N}\sum_{i=1}^{N}\Big\| \boldsymbol{z}_{t}^{(n)}(i) - \bar{\boldsymbol{z}}^{(n)}(i) \Big\|_2, \end{aligned} \tag{52}$$

where $\bar{\boldsymbol{z}}^{(n)}(i)$ is the per-video mean patch representation over the frames used for $\mathcal{P}^{(n)}$. We normalize each video by its own radius and then average:

$$D_{intra} \;=\; \frac{1}{M}\sum_{n=1}^{M} \frac{D_{intra}^{ori,(n)}}{2R_{intra}^{(n)}}. \tag{53}$$

3) D (Distance margin). The trade-off margin balances the two normalized distances with a scale factor γ:

$$\begin{aligned} D \;&=\; D_{inter} - \gamma\, D_{intra}, \\ \gamma \;&=\; \frac{\mathbb{E}_{\mathcal{M}}[D_{intra}^{ori}]}{\mathbb{E}_{\mathcal{M}}[D_{inter}^{ori}]}, \end{aligned} \tag{54}$$

where $\mathcal{M}$ indexes the set of models under comparison. Model-specific values of the scale factor γ are listed in Tab. 4 and concentrate within a narrow range. Therefore, to unify the setting, we use the average $\gamma = 0.3$ in practice.

4) *Cyc. Acc.* (Cycle-consistency accuracy). Given two frames forming a palindrome traversal and N patches per frame, let $\boldsymbol{A}_{t_a}^{t_b}$ and $\boldsymbol{A}_{t_b}^{t_a}$ be the patch-wise correlation transition matrices, and set $\boldsymbol{P} = \boldsymbol{A}_{t_a}^{t_b}\,\boldsymbol{A}_{t_b}^{t_a}$. The cycle-consistency accuracy is the proportion of patches returning to their original indices:

$$\textit{Cyc. Acc.} \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\Big\{ \arg\max_j P_{ij} = i \Big\}. \tag{55}$$

Table 4. The scale factor $\gamma = D_{intra}^{ori}/D_{inter}^{ori}$ and the average scale factor for each model.

Method	γ	Method	γ
MAE	0.1855	MoCov3	0.3321
MAE +***Ours***	0.3289	MoCov3 +***Ours***	0.3053
I-JEPA	0.2283	iBOT	0.2817
I-JEPA +***Ours***	0.2645	iBOT +***Ours***	0.3084
CLIP	0.3365	DINO	0.3378
CLIP +***Ours***	0.3876	DINO +***Ours***	0.3505
BLIP	0.2489	DINOv2	0.2730
BLIP +***Ours***	0.3737	DINOv2 +***Ours***	0.2916
Average γ : 0.3021			

D.4. Image-pretrained Fundamental Models

We evaluate our method using eight representative pretrained image encoders, which can be broadly categorized into three paradigms of self-supervised learning: 1) *Masked modeling*: MAE [37], I-JEPA [2]; 2) *Contrastive learning*: CLIP [70], BLIP [51], MoCo v3 [18]; 3) *Self-distillation*: iBOT [95], DINO [12], DINO v2 [62]. All models are pretrained on ImageNet-1k with self-supervised objectives, except for CLIP and BLIP, which are pretrained with natural language supervision. We adopt ViT-Base [23] architectures with a patch size of 16 as the backbone encoder for each model.

- **MAE** [37] follows an encoder-decoder architecture, where random image patches are masked and the model is trained to reconstruct the missing content at the pixel level.
- **I-JEPA** [2] learns representations by predicting latent representations of masked regions. It discards the decoder and instead relies on semantic-level prediction to better capture high-level image structures.
- **CLIP** [70] is a vision-language model trained with natural language supervision. It learns to align image and text embeddings in a shared feature space using a contrastive objective on WIT dataset.
- **BLIP** [51] is a vision-language model that extends CLIP-style contrastive pretraining with additional image-text matching and language modeling objectives. By jointly

Table 5. Evaluation results on frame-level and video-level tasks based on representative image models. The best results are marked in **bold**.

Model	Action Localization					Video Retrieval				Action Classification		Temporal Order Discrimination
	Breakfast					UCF101		HMDB51		SSV2		Chiral SSV2
	Edit	Acc	F1@0.10	F1@0.25	F1@0.50	mAP	R@1	mAP	R@1	Acc@1	Acc@5	Acc
CLIP	53.8	40.1	**52.9**	46.0	33.8	45.9	90.8	25.5	70.1	34.6	67.8	79.1
CLIP ***+Ours***	**54.9**	**40.9**	52.5	**46.4**	**34.8**	**49.0**	**96.0**	**27.1**	**71.3**	**35.5**	**68.8**	**80.5**
BLIP	57.3	47.3	55.6	49.4	36.7	54.4	96.4	29.4	73.3	39.9	72.4	81.8
BLIP ***+Ours***	**58.6**	**49.1**	**57.5**	**51.0**	**38.1**	**55.2**	**97.0**	**29.9**	**73.5**	**41.4**	**72.6**	**82.8**
iBOT	55.5	40.7	53.2	**47.6**	35.3	33.4	92.0	18.1	59.9	38.1	68.9	80.1
iBOT ***+Ours***	**56.3**	**42.9**	**53.5**	47.5	**37.3**	**34.6**	**94.9**	**18.8**	**63.9**	**40.6**	**71.3**	**82.3**
DINO v2	58.5	44.1	57.0	51.3	38.2	37.1	93.3	18.7	62.1	39.2	71.0	84.8
DINO v2 ***+Ours***	**61.2**	**50.5**	**60.0**	**54.0**	**41.2**	**39.3**	**95.2**	**20.0**	**67.0**	**40.4**	**72.1**	**85.2**

optimizing these objectives on large-scale web data, BLIP learns richer cross-modal representations and achieves stronger performance on image captioning and visual question answering.

- **MoCo v3** [18] is a contrastive learning method that adapts Momentum Contrast to Vision Transformers, employing a siamese architecture with an online encoder and a momentum-updated target encoder, and discards the negative sample queue used in earlier versions.
- **DINO** [12] adopts a self-distillation structure with Vision Transformers as the encoder. It encourages consistent representations across different views of the same image.
- **iBOT** [95] builds upon DINO by introducing additional alignment on dense patch tokens. It aligns both the global `[CLS]` token and local patch-level features between two views, thus encouraging fine-grained spatial consistency in the learned representations.
- **DINO v2** [62] extends iBOT by incorporating various design improvements, including better centering techniques [72], regularization strategies like KoLeo loss [73], and resolution-adaptive training [83]. In our experiments, we exclude computationally intensive techniques to ensure a consistent and fair comparison across models.

D.5. Competitors

We compare our approach against two categories of strong baselines: recent state-of-the-art (SOTA) video representation learning methods and image-to-video adaptation frameworks. For all baselines, we use the officially released pretrained weights without any additional training or fine-tuning.

1) Video representation learning methods: These methods are specifically designed to learn spatiotemporal representations from raw video inputs, often relying on temporal masking or reconstruction-based objectives.

- **VideoMAE** [82] extends the masked modeling paradigm to videos by randomly masking spatiotemporal tubes and reconstructing the missing pixels. It adopts a high masking ratio to encourage the encoder to capture both appearance and motion features.
- **MAE-ST** [28] adapts MAE to spatiotemporal data by explicitly incorporating temporal modeling modules into the encoder to better capture dynamic patterns.
- **DropMAE** [87] applies spatial-attention dropout in masked modeling, encouraging the model to attend to motion cues for temporal discriminability.
- **SiamMAE** [32] adopts a Siamese structure where the past frame and a masked version of the current frame are jointly encoded. A conditional decoder is employed to reconstruct the missing patches, thereby promoting temporal consistency across frames.
- **CropMAE** [25] generalizes SiamMAE by using different crops or augmentations of the same frame as input, encouraging invariance under intra-frame transformations.
- **RSP** [43] formulates temporal modeling as a stochastic frame prediction task. It learns to reconstruct future frames from current ones by modeling both prior and posterior distributions over latent motion variables.

2) Image-to-video adaptation methods: These methods aim to adapt pretrained image models to instance-level video understanding tasks by integrating lightweight modules that enable temporal reasoning, while keeping most of the backbone parameters frozen and only updating a small subset.

- **AIM** [91] introduces a lightweight adapter into a frozen ViT backbone, enabling spatiotemporal adaptation along spatial and temporal dimensions, facilitating efficient transfer from static to dynamic inputs.
- **ST-Adapter** [63] proposes a 3D bottleneck adapter into the CLIP-pretrained ViT model, which enables the model to reason about dynamic video content at a small task-specific parameter cost.
- **ZeroI2V** [55] introduces spatial-temporal dual-headed attention mechanism combined with a linear adaptation layer, thus enabling the transfer of frozen image models to video tasks and supporting zero additional inference cost via structural reparameterization.

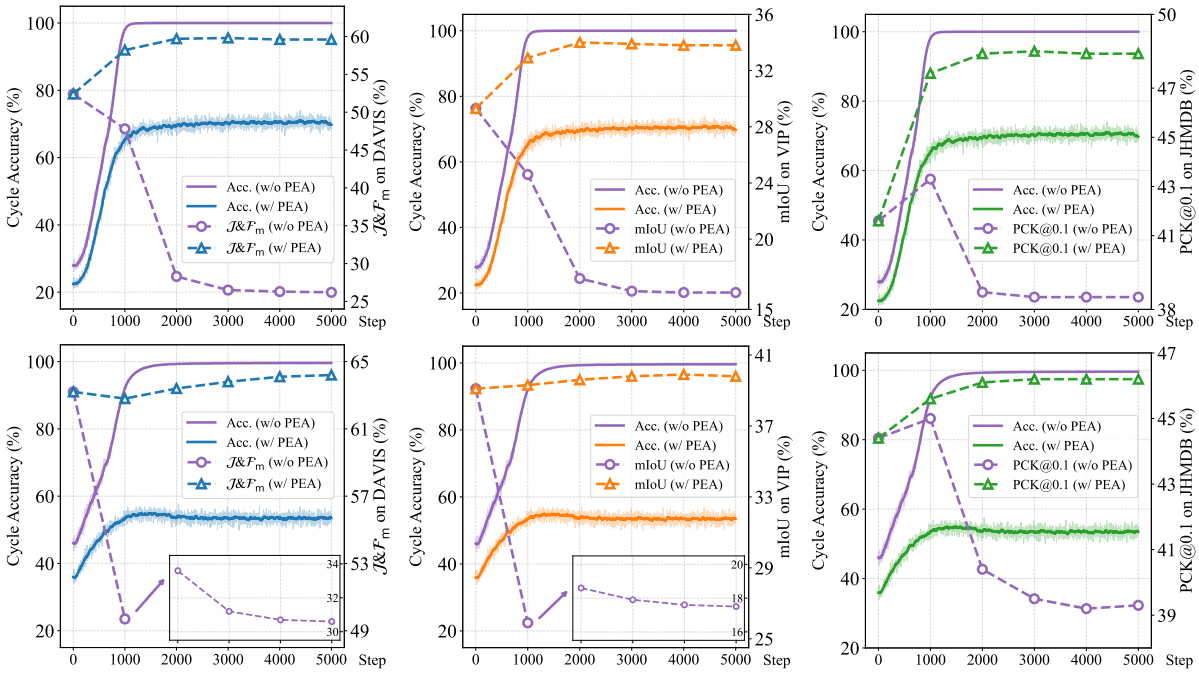


Figure 2. Cycle-consistent accuracy and downstream performance during training with or without the PEA strategies across three tasks based on MAE (*line 1*) and DINO (*line 2*).

Table 6. Extended comparison with representative methods on the DAVIS-2017 validation set. Methods are grouped by their core settings for a broader reference.

Type	Method	Backbone	DAVIS-2017		
			$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Dedicated VOS Systems	STCN [19]	ResNet50	85.4	82.2	88.6
	SwinB-AOT-L [92]	Swin-B	85.4	82.4	88.4
	SimVOS-B [88]	ViT-B/16	88.0	85.0	91.0
	Cutie-base [20]	ResNet50	87.9	84.6	91.1
Segmentation Foundation Models	SAM 2 [71]	Hiera-B+	90.2	87.0	93.4
	SAM 2 [71]	Hiera-L	90.7	87.5	94.0
Self-Supervised Video Pre-training	VideoMAE [82]	ViT-L/16	45.0	43.6	46.5
	MAE-ST [28]	ViT-L/16	54.6	55.5	53.6
	SiamMAE [32]	ViT-B/16	60.9	59.4	62.4
	CropMAE [25]	ViT-B/16	57.8	56.9	58.7
	RSP [43]	ViT-B/16	60.5	57.8	63.2
Self-Supervised Image Pre-training ***+Ours***	DINO [12]	ViT-B/16	63.2	60.9	65.5
	DINO + ***Ours***	ViT-B/16	64.2	62.3	66.0
	DINOv2 [62]	ViT-B/16	63.1	61.6	64.5
	DINOv2 + ***Ours***	ViT-B/16	63.7	61.9	65.4
	iBOT [95]	ViT-B/16	64.6	63.0	66.1
	iBOT + ***Ours***	ViT-B/16	65.1	63.3	66.9

E. Detailed Experiments Results

E.1. Comparison with Task-Specific SOTAs

To provide a broader view, we compare our method with several representative VOS systems and recent segmentation foundation models on DAVIS-2017 validation in Tab. 6. Our work focuses on general representation pre-training for direct **transfer** across multiple tasks rather than task-specific designs, and thereby applies lightweight transfer for evaluation per standard self-supervised learning protocols. Thus, our datasets, computing resources, and architectures are not aligned with specialized SoTA methods for individual tasks such as video object segmentation (VOS).

E.2. Detailed Results of Frame-/Video-Level Tasks

We further evaluate the transferred models on several frame- and video-level downstream tasks: temporal action localization on Breakfast [49] using the FACT [61] backbone, zero-shot video retrieval on UCF101 and HMDB51 [48, 77], fine-tuned action classification on Something-Something-v2 (SSV2) [30], and temporal order discrimination via linear probing on Chiral SSV2 [4].

The quantitative results of transferred representations from four representative image models on both frame-level and video-level downstream tasks are depicted in Tab. 5. Our method delivers steady performance improvements across these tasks. For instance, on frame-level tasks, it achieves an average improvement of 2.80% *Acc* on Breakfast, indicating enhanced temporal awareness in image models. On video-level tasks, it brings a 2.58% $R@1$ improvement on HMDB51 and a 1.53% $Acc@1$ gain on SSV2, which validates the preserved semantic discrimination ability. These results indicate that our method generalizes well across different task granularities, highlighting its potential as a versatile solution for image-to-video transfer.

E.3. Training Dynamics

We visualize the training dynamics of MAE and DINO across three downstream tasks in Figure 2. The plots show the cycle-consistency accuracy (*i.e.*, the percentage of patches that return to their original positions after a cycle traversal) together with the downstream performance over training steps. Without the PEA strategy, the downstream performance drops sharply within the first two epochs, even when the cycle-consistency accuracy is close to 100%. This indicates that the model exploits the absolute positional encoding as a shortcut instead of learning temporal correspondences that remain reliable when the temporal distance between frames grows.

In contrast, when we apply the proposed PEA strategy, the cycle-consistency accuracy increases gradually, and the final value converges to a small stable range that depends on the model architecture and hyperparameter settings. This behavior is reasonable, since in real-world videos, correspondence quality naturally degrades as time passes: the first and last frames in a propagation chain can differ greatly due to camera motion and non-rigid object deformation, which leads to unavoidable information loss. On the Kinetics-400 dataset, the empirical cycle-consistency accuracy stabilizes around $50\% \sim 70\%$ when the temporal interval is set to $\delta = 0.15$. By promoting effective dense correspondences between frames and reducing reliance on positional cues, PEA leads to more stable improvements in downstream performance and highlights its role in learning robust temporal representations.

E.4. Shortcut Phenomenon in Training

Tab. 7 compares the performance of our method trained with and without the proposed Positional Encoding Augmentation (PEA) strategy. As shown, removing PEA consistently leads to substantial performance degradation, with $4.4\% \sim 37.3\%$ drop in $\mathcal{J}\&\mathcal{F}_{\mathrm{m}}$ on DAVIS and $4.9\% \sim 22.8\%$ drop in mIoU on VIP. This is primarily due to the model exploiting absolute positional encodings as shortcuts, resulting in dimensional collapse and degraded representations. The issue is particularly severe in self-distillation architectures, which rely heavily on positional alignment between teacher and student branches. This highlights the brittleness of image-pretrained representations when transferred to video and underscores that image-to-video transfer is a non-trivial challenge. In contrast, applying PEA consistently improves performance across all three downstream tasks, indicating the effectiveness of resisting shortcuts induced by the positional encoding mechanism of ViT.

E.5. Additional Ablation Study

In Figure 3, we study the effects of the interpolation ratio α and the regularization weight λ. A moderate value of α gives the best performance since a small α cannot effectively suppress shortcut learning, while a large one disrupts relative positional cues and harms correspondence learning. Similarly, λ controls the strength of the semantic separability constraint: too small values may cause dimensional collapse in the projection space, whereas overly strong regularization reduces the flexibility needed to adapt the representations. Overall, both hyperparameters influence performance in a relatively mild range, and good results can be obtained with moderate choices.

Table 7. Impact of Positional Encoding Augmentation (PEA) strategy on representation quality across three downstream tasks.

Image Model	Method	VIP mIoU	DAVIS17 $\mathcal{J}\&\mathcal{F}_{\mathrm{m}}$	JHMDB PCK@0.1
MAE	Vanilla	29.3	52.4	41.6
	w/o PEA	$16.2_{-13.1}$	$26.2_{-26.2}$	$38.5_{-3.1}$
	w/ PEA	$33.8_{+4.5}$	$59.6_{+7.2}$	$48.4_{+6.8}$
I-JEPA	Vanilla	31.5	53.9	42.6
	w/o PEA	$26.6_{-4.9}$	$49.5_{-4.4}$	$44.1_{+1.5}$
	w/ PEA	$35.3_{+3.8}$	$58.7_{+4.8}$	$44.4_{+1.8}$
MoCo v3	Vanilla	38.8	62.6	43.6
	w/o PEA	$23.8_{-15.0}$	$42.8_{-19.8}$	$42.2_{-1.4}$
	w/ PEA	$39.8_{+1.0}$	$62.9_{+0.3}$	$45.3_{+1.7}$
iBOT	Vanilla	39.6	64.6	45.7
	w/o PEA	$16.8_{-22.8}$	$27.3_{-37.3}$	$38.2_{-7.5}$
	w/ PEA	$40.8_{+1.2}$	$65.1_{+0.5}$	$46.1_{+0.4}$
DINO	Vanilla	39.1	63.2	44.4
	w/o PEA	$17.5_{-21.6}$	$30.6_{-32.6}$	$39.3_{-5.1}$
	w/ PEA	$39.8_{+0.7}$	$64.2_{+1.0}$	$46.2_{+1.8}$
DINO v2	Vanilla	38.4	63.1	46.6
	w/o PEA	$17.7_{-20.7}$	$30.0_{-33.1}$	$39.1_{-7.5}$
	w/ PEA	$39.9_{+1.5}$	$63.7_{+0.6}$	$47.3_{+0.7}$

Tab. 8 analyzes the sensitivity of the temporal sampling interval δ and the softmax temperature τ. A suitable δ balances visible motion and visual continuity, which is important for learning meaningful frame-level correspondences, while a moderate τ maintains an appropriate level of similarity sharpening. The model shows limited sensitivity to variations in these two hyperparameters, and the performance remains stable across a reasonable range. To ensure consistency and fair comparison across all experiments, we fix $\tau = 0.03$ and $\delta = 0.15$.

We further conduct a robustness and generalization analysis by varying the PEA crop strategy (Tab. 9a) and the model patch size alongside positional encoding variants (Tab. 9b). PEA remains stable across a wide range of crop choices and generalizes well across various patch sizes (e.g., 8, 14, and 16). Moreover, PEA is compatible with modern designs such as RoPE [78]. By interpolating and cropping on the RoPE coordinate grid, PEA effectively mitigates shortcut behaviors, further demonstrating the robustness of our method.

As shown in Tab. 10, we investigate the impact of different regularization objectives. The KL-based regularization matches the distribution of transferred video representations to that of frozen image features. This helps preserve the

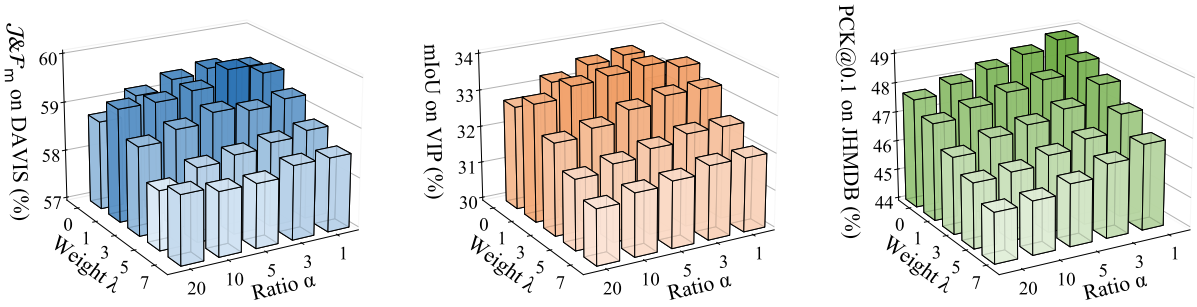


Figure 3. 3D bar charts for ablation results on interpolation ratio α and regularization weight λ across three tasks using MAE.

Table 8. Sensitivity analysis on temporal interval δ and softmax temporature τ. Default settings are highlighted with blue.

(a) Ablation on temporal interval δ.

Base Model	δ	VIP mIoU	DAVIS17 $\mathcal{J}\&\mathcal{F}_{\text{m}}$	JHMDB PCK@0.1
MAE	0.05	33.9	59.6	48.6
	0.10	33.9	59.7	48.4
	0.15	33.8	59.6	48.4
	0.20	33.6	59.7	48.5
	0.25	33.6	59.6	48.4
DINO	0.05	39.7	63.9	46.2
	0.10	39.9	64.0	46.1
	0.15	39.8	64.2	46.2
	0.20	39.9	64.2	46.1
	0.25	39.9	64.2	46.1

(b) Ablation on softmax temporature τ.

Base Model	τ	VIP mIoU	DAVIS17 $\mathcal{J}\&\mathcal{F}_{\text{m}}$	JHMDB PCK@0.1
MAE	0.01	32.6	60.0	48.2
	0.02	33.4	60.0	48.5
	0.03	33.8	59.6	48.4
	0.04	33.7	58.9	48.5
	0.05	33.6	58.6	48.4
DINO	0.01	39.2	63.7	46.0
	0.02	40.0	63.9	46.1
	0.03	39.8	64.2	46.2
	0.04	39.9	64.0	46.0
	0.05	39.9	63.8	45.9

Table 9. Robustness and generalization analysis of PEA strategy across DINO series features.

(a) Robustness across PEA crop manners.

Base Model	PEA Crop	VIP mIoU	DAVIS17 $\mathcal{J}\&\mathcal{F}_{\text{m}}$	JHMDB PCK@0.1
DINO	center	64.1	39.3	**46.9**
	random	64.2	39.8	46.2
	edge	64.1	**39.9**	46.2
	multiple	**64.3**	39.8	46.2

(b) Generalization across PE variants and patch sizes.

PEA	L_{reg}	DINO (Abs. PE) ViT-S/8	DINO v2 (Abs. PE) ViT-S/14	DINO v3 (RoPE) ViT-S/16
Vanilla		71.7	64.7	67.3
✗	✓	71.1	63.9	65.8
✓	✓	**72.3**	**65.1**	**67.9**

inherited semantic geometry and prevents feature collapse by aligning distance relationships (as discussed in Sec. 4). Compared to a strict element-wise MSE loss, KL divergence provides a softer, distribution-level constraint. This allows for sufficient temporal adaptation while effectively maintaining semantic separability. Consequently, KL regularization tends to increase the normalized inter-video distance

Table 10. Ablation study on the choice of regularization loss (L_{reg}) using the DINO backbone.

PEA	L_{reg}	DAVIS	VIP	JHMDB	D_{inter}	D_{intra}	D (↑)
DINO		63.2	39.1	44.4	0.5756	0.2144	0.5112
✗	KL	61.8	38.0	46.1	0.6241	0.2404	0.5520
✓	MSE	62.2	38.1	46.1	0.6142	0.2370	0.5431
✓	KL	**64.2**	**39.8**	**46.2**	0.6246	0.2316	**0.5551**

(D_{inter}), which aligns perfectly with the observed improvements in downstream task performance.

F. Additional Visualizations

F.1. Inter-frame Correspondence

We visualize the inter-frame correspondence learned by the projection layer g in Figure 4. The results indicate that most patches establish consistent matches across frames and successfully return to their original locations through the forward-backward cycle. Notably, due to factors such as camera motion and non-rigid object deformation, patch correspondences between $\boldsymbol{v}_{t_1}^f$ and $\boldsymbol{v}_{t_2}$ are not strictly bijective. A single patch in $\boldsymbol{v}_{t_1}^f$ often correlates to multiple adjacent regions in $\boldsymbol{v}_{t_2}$, resulting in a correlation matrix product

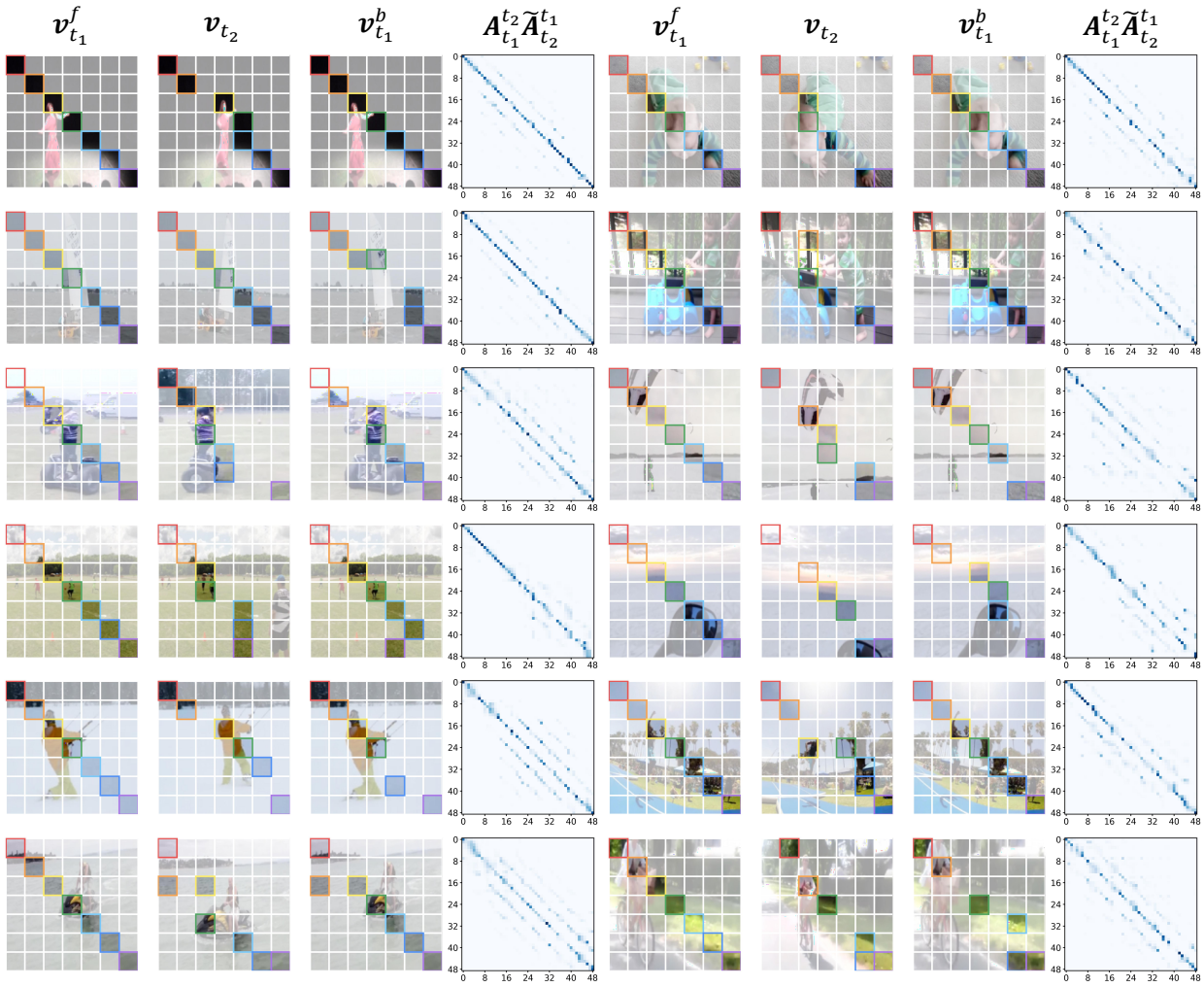


Figure 4. Cross-frame correspondence learned with our method. Patches with the same color box represent correspondence.

$A^{t_2}_{t_1}\widetilde{A}^{t_1}_{t_2}$ that exhibits a diagonally dominant structure rather than an exactly equal to the identity matrix I. This observation reveals the dilemma of the original contrastive random walk strategy: it needs to constrain the matrix to the identity matrix to ensure good cyclic consistency, but we cannot make it a perfect identity matrix because it would allow the model to take advantage of shortcuts in displaying positional encoding. This further justifies the necessity of our proposed PEA strategy, which effectively suppresses shortcut matching to stabilize correspondence learning.

F.2. Downstream Task Performance

In Figures 5 and 6, we compare the performance of original image-pretrained models and our transferred models across three downstream tasks. Our method shows visible improvements in several challenging scenarios, such as rapid movements, complex object boundaries, and motion-induced artifacts, where the original models often underperform. These results suggest that incorporating temporal correspondence and strengthening semantic structure improves image-to-video representation transfer, validating the effectiveness of our method.

G. Detailed Related Work

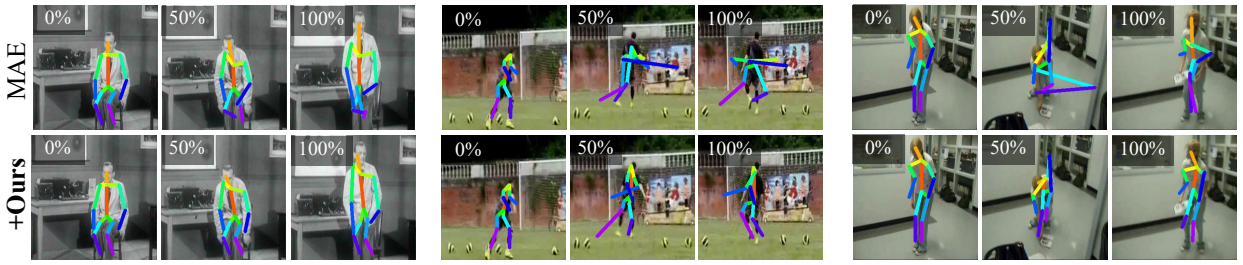
G.1. Self-supervised Visual Representation

The rapid progress of self-supervised learning has enabled models to acquire generalizable representations for diverse downstream tasks in both the image and video domains. Depending on the nature of the pretraining objective, existing approaches can be broadly categorized into three paradigms.

Contrastive learning learns invariant representations by maximizing agreement between relevant instances while pushing apart representations of different instances. Early methods in the image domain construct positive and negative pairs [16, 18, 36] or apply diverse augmentations [11, 17, 31] to generate contrasting views. These approaches demonstrate strong generalization capabilities [39] and have been

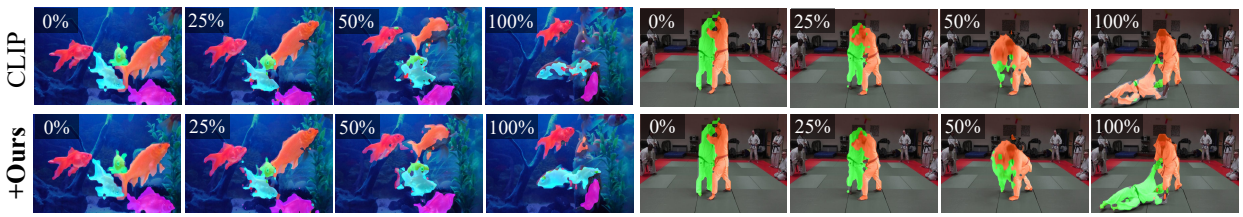
(a) Video Object Segmentation on DAVIS-2017

(b) Semantic Part Propagation on VIP



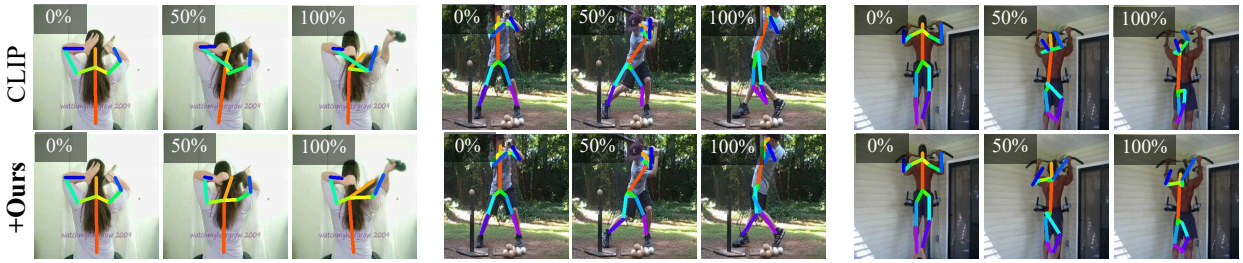
(c) Human Pose Propagation on JHMDB

Figure 5. Visualization comparison across three downstream tasks based on MAE.



(a) Video Object Segmentation on DAVIS-2017

(b) Semantic Part Propagation on VIP



(c) Human Pose Propagation on JHMDB

Figure 6. Visualization comparison across three downstream tasks based on CLIP.

successfully extended to the video domain. By leveraging 3D convolutions [27], temporal self-attention [1, 7, 10], or inter-frame contrastive objectives [34, 41, 80, 93], such methods benefit from spatiotemporal cues and have shown promising results on discriminative tasks such as action recognition and video retrieval [48, 77].

Masked modeling aims to reconstruct the original RGB values of masked image patches in the pixel space [2, 5, 6, 37, 79, 89]. A representative method is MAE [37], which employs an encoder-decoder architecture based on Vision Transformers [23] to restore the masked regions, thereby capturing structural dependencies within the images. By incorporating the additional temporal dimension, MAE can be naturally extended for video representation learning [28, 66, 82, 84, 87]. To alleviate the computational cost of dense modeling, recent methods focus on more efficient designs. SiamMAE [32] leverages sparsely sampled frames, asymmetric masking, and a conditional Siamese architecture, motivating subsequent works that improve frame selection and predictive mechanisms [25, 43, 90].

Self-distillation methods supervise a student network using outputs from a teacher network without relying on explicit labels, often focusing on restoring latent representations rather than raw pixels. This encourages the learning of high-level semantic information, aligning with principles of information compression [45, 81]. DINO [12] adopts a self-distillation framework with Vision Transformers to align patch-level representations across views. Subsequently, iBOT [95] and DINO v2 [62] extend this paradigm by enforcing consistency in both global `[CLS]` tokens and dense patch representations.

G.2. Image-to-video Transfer Learning

Temporal structure enhancement methods typically design training objectives in a two-stage training manner based on self-supervised image contrastive learning frameworks [31, 36]. In the first stage, models are pretrained on image datasets to learn static representations for instance-level discrimination [38, 52], or on synthetic videos to capture object motion patterns [22, 53]. In the second stage, the models are fine-tuned on real video datasets to refine temporal correspondences, enabling them to perform specific video tasks. However, the high spatiotemporal complexity hinders swift cross-domain representation transfer, motivating the exploration of parameter-efficient fine-tuning alternatives in subsequent works.

Parameter-efficient fine-tuning methods aim to adapt pretrained models to video tasks by updating only a small fraction of parameters. Specifically, several methods insert adapters into Vision Transformer [23] pretrained by CLIP [70] in a series or parallel way, enabling spatial-temporal joint adaptation through expanded convolution or attention modules [15, 55, 56, 63, 91]. Other approaches decouple spatial and temporal modeling using dual-branch architectures [64, 69], enabling separate reasoning across spatial and temporal dimensions. These adaptation methods are often trained on supervised action recognition datasets [30, 46], which require further fine-tuning when applied to different benchmarks. More recent work explores object-centric adaptation via slot attention [59], demonstrating the potential of using image-pretrained encoders for dense prediction tasks [68]. In a related direction, ProLIP [26] shows that fine-tuning only the visual projector is effective for few-shot CLIP adaptation, showing the strong transfer capacity of lightweight projection-based adaptation.

G.3. Temporal Cycle Consistency

The inherent visual correspondence between temporally adjacent observations provides a powerful supervisory signal to capture spatiotemporal coherence in videos [3, 76]. Leveraging this property, numerous studies attempt to learn semantically consistent representations with a cycle structure, showing effectiveness in dense-level video tasks, including object segmentation [67, 96], motion estimation [44], and point tracking [9, 13]. Early methods mainly focus on tracking patches or objects across frames in a bidirectional manner [54, 85, 94], while others align the feature distributions among videos from the same category to enforce semantic consistency [24, 33, 86]. Another line of work introduces random walk strategies [8, 42, 75], where representation learning is guided by maximizing the probability of each patch returning to itself via a forward-backward cycle.

H. Additional Discussions

H.1. Limitation and Future Work

This work explores a more efficient and effective approach to transferring image representations to the video domain. In this work, we mainly focus on ViT-based backbones under the evaluated settings. For future work, we plan to extend the method to other visual backbones, including lightweight architectures (*e.g.*, CNNs, ResNets) and emerging large-scale vision models, to assess whether the observed trade-off is a general property of visual representations for video understanding.

H.2. Broader Impact

We examine the trade-off between intra-video temporal consistency and inter-video semantic separability in visual representations and, based on this view, propose a method for image-to-video representation transfer learning. The proposed method achieves competitive or superior performance compared with models pretrained on video from scratch, providing a lightweight alternative for video representation learning. It may also provide a useful perspective for future research on image-to-video transfer in broader scenarios.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *International Conference on Computer Vision*, pages 6836–6846, 2021. 18

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 11, 18

[3] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954. 18

[4] Piyush Bagad and Andrew Zisserman. Chirality in action: Time-aware video representation learning by latent straightening. *arXiv preprint arXiv:2509.08502*, 2025. 10, 13

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. 2022. 18

[6] Amir Bar, Florian Bordes, Assaf Shocher, Mido Assran, Pascal Vincent, Nicolas Ballas, Trevor Darrell, Amir Globerson, and Yann LeCun. Stochastic positional embeddings improve masked image modeling. In *International Conference on Machine Learning*, 2024. 18

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021. 18

[8] Zhangxing Bian, Allan Jabri, Alexei A Efros, and Andrew Owens. Learning pixel trajectories with multiscale contrastive random walks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6508–6519, 2022. 18

[9] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Asian Conference on Computer Vision*, pages 3–19. Springer, 2019. 18

[10] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *Advances in Neural Information Processing Systems*, 34:19594–19607, 2021. 18

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 16

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision*, pages 9650–9660, 2021. 11, 12, 13, 18

[13] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 18

[14] Nutan Chen, Alexej Klushyn, Francesco Ferroni, Justin Bayer, and Patrick Van Der Smagt. Learning flat latent manifolds with vaes. In *International Conference on Machine Learning*, pages 1587–1596, 2020. 2

[15] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 7, 18

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 16

[17] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 16

[18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *International Conference on Computer Vision*, pages 9640–9649, 2021. 11, 12, 16

[19] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 13

[20] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024. 13

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 8

[22] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, et al. Bootstap: Bootstrapped training for tracking-any-point. In *Asian Conference on Computer Vision*, pages 3257–3274, 2024. 7, 18

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 11, 18

[24] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 18

[25] Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient image pre-training with siamese cropped masked autoencoders. In *European Conference on Computer Vision*, 2024. 7, 8, 12, 13, 18

[26] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Clip's visual embedding projector is a few-shot cornucopia. pages 3254–3264, 2026. 18

[27] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 3299–3309, 2021. 18

[28] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in Neural Information Processing Systems*, 35:35946–35958, 2022. 7, 8, 12, 13, 18

[29] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 5

[30] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *International Conference on Computer Vision*, pages 5842–5850, 2017. 7, 8, 9, 10, 13, 18

[31] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 16, 18

[32] Agrim Gupta, Jiajun Wu, Jia Deng, and Fei-Fei Li. Siamese masked autoencoders. *Advances in Neural Information Processing Systems*, 36:40676–40693, 2023. 7, 8, 12, 13, 18

[33] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11068–11077, 2021. 18

[34] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 18

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *International Conference on Computer Vision*, pages 1026–1034, 2015. 5

[36] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 16, 18

[37] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 11, 18

[38] Yingdong Hu, Renhao Wang, Kaifeng Zhang, and Yang Gao. Semantic-aware fine-grained correspondence. In *European Conference on Computer Vision*, pages 97–115. Springer, 2022. 7, 8, 18

[39] Weiran Huang, Mingyang Yi, Xuyang Zhao, and Zihao Jiang. Towards the generalization of contrastive self-supervised learning. In *International Conference on Learning Representations*, 2023. 16

[40] In Huh, Jae Myung Choe, YOUNGGU KIM, Daesin Kim, et al. Isometric quotient variational auto-encoders for structure-preserving representation learning. *Advances in Neural Information Processing Systems*, 36:39075–39087, 2023. 2

[41] Yuqi Huo, Mingyu Ding, Haoyu Lu, Nanyi Fei, Zhiwu Lu, Ji-Rong Wen, and Ping Luo. Compressed video contrastive learning. *Advances in Neural Information Processing Systems*, 34:14176–14187, 2021. 18

[42] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in Neural Information Processing Systems*, 33:19545–19560, 2020. 18

[43] Huiwon Jang, Dongyoung Kim, Junsu Kim, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Visual representation learning with stochastic frame prediction. In *International Conference on Machine Learning*, pages 21289–21305, 2024. 7, 8, 12, 13, 18

[44] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013. 8, 9, 18

[45] Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *International Conference on Machine Learning*, pages 16049–16096. PMLR, 2023. 18

[46] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7, 18

[47] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 2

[48] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 9, 10, 13, 18

[49] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–787, 2014. 9, 13

[50] Yonghyeon Lee, Sangwoong Yoon, MinJun Son, and Frank C Park. Regularized autoencoders for isometric representation learning. In *International Conference on Learning Representations*, 2022. 2

[51] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 8, 11

[52] Rui Li and Dong Liu. Spatial-then-temporal self-supervised learning for video correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2279–2288, 2023. 7, 8, 18

[53] Rui Li, Shenglong Zhou, and Dong Liu. Learning fine-grained features for pixel-wise video correspondences. In *Interna-*

tional Conference on Computer Vision, pages 9632–9641, 2023. 7, 18

[54] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. *Advances in Neural Information Processing Systems*, 32, 2019. 18

[55] Xinhao Li, Yuhan Zhu, and Limin Wang. Zeroi2v: Zero-cost adaptation of pre-trained transformers from image to video. In *European Conference on Computer Vision*, pages 425–443. Springer, 2024. 7, 12, 18

[56] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 7, 18

[57] Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, and Qingming Huang. Not all pairs are equal: Hierarchical learning for average-precision-oriented video retrieval. In *ACM International Conference on Multimedia*, pages 3828–3837, 2024. 10

[58] Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, and Qingming Huang. When the future becomes the past: Taming temporal correspondence for self-supervised video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24033–24044, 2025. 7, 8

[59] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020. 18

[60] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 8

[61] Zijia Lu and Ehsan Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18175–18185, 2024. 9, 13

[62] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 11, 12, 13, 18

[63] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 7, 8, 12, 18

[64] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2203–2213, 2023. 18

[65] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[66] Gensheng Pei, Tao Chen, Xiruo Jiang, Huafeng Liu, Zeren Sun, and Yazhou Yao. Videomac: Video masked autoencoders meet convnets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22733–22743, 2024. 18

[67] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 8, 9, 18

[68] Rui Qian, Shuangrui Ding, and Dahua Lin. Rethinking image-to-video adaptation: An object-centric perspective. In *European Conference on Computer Vision*, pages 329–348. Springer, 2024. 18

[69] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Disentangling spatial and temporal learning for efficient image-to-video transfer learning. In *International Conference on Computer Vision*, pages 13934–13944, 2023. 18

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 7, 8, 11, 18

[71] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 13

[72] Yangjun Ruan, Saurabh Singh, Warren Richard Morningstar, Alexander A Alemi, Sergey Ioffe, Ian Fischer, and Joshua V Dillon. Weighted ensemble self-supervised learning. In *International Conference on Learning Representations*, 2023. 12

[73] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *International Conference on Learning Representations*, 2019. 12

[74] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 8

[75] Ayush Shrivastava and Andrew Owens. Self-supervised any-point tracking by contrastive random walks. In *European Conference on Computer Vision*, pages 267–284. Springer, 2024. 18

[76] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001. 18

[77] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 9, 10, 13, 18

[78] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 14

[79] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2132–2141, 2023. 18

[80] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. Tubelet-contrastive self-supervision for video-efficient generalization. In *International Conference on Computer Vision*, pages 13812–13823, 2023. 18

[81] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015. 18

[82] 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. 7, 8, 12, 13, 18

[83] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in Neural Information Processing Systems*, 32, 2019. 12

[84] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 7, 18

[85] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 18

[86] Haiping Wu and Xiaolong Wang. Contrastive learning of image representations with cross-video cycle-consistency. In *International Conference on Computer Vision*, pages 10149–10159, 2021. 18

[87] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14561–14571, 2023. 12, 18

[88] Qiangqiang Wu, Tianyu Yang, Wei Wu, and Antoni B Chan. Scalable video object segmentation with simplified framework. In *International Conference on Computer Vision*, pages 13879–13889, 2023. 13

[89] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 18

[90] Jiange Yang, Bei Liu, Jianlong Fu, Bocheng Pan, Gangshan Wu, and Limin Wang. Spatiotemporal predictive pre-training for robotic motor control. *arXiv preprint arXiv:2403.05304*, 2024. 18

[91] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *International Conference on Learning Representations*, 2023. 7, 8, 12, 18

[92] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. 13

[93] Youngjae Yu, Sangho Lee, Gunhee Kim, and Yale Song. Self-supervised learning of compressed video representations. In *International Conference on Learning Representations*, 2020. 18

[94] Yurong Zhang, Liulei Li, Wenguan Wang, Rong Xie, Li Song, and Wenjun Zhang. Boosting video object segmentation via space-time correspondence learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2246–2256, 2023. 18

[95] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations*, 2022. 11, 12, 13, 18

[96] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1527–1535, 2018. 8, 9, 18