Towards Generalized Action Recognition on Low-Resolutions with Domain-Invariant Representation

Hao Li[1,2] , **Jinhui Xu**[1,2] and **Dianlong You**[1,2,*]

[1]School of Artificial Intelligence, Yanshan University, China
[2]The Key Laboratory for software engineering of Hebei Province, Yanshan University, China
lihao_0216@sina.com, jinhuixu_ai@163.com, youdl@ysu.edu.cn

Abstract

This paper studies cross-domain/species action recognition from low-resolution videos with sharp appearance variations and domain-specific biases. Its attractive viewpoint is mining domain-invariant representations of cross-domain/species features under rough spatial details to enhance recognition and generalization, overcoming the significant decline of existing methods. To address this, we propose a generalized Action recognition framework for Low-resolution conditions with Domain-invariant Representation learning, named ActLDR, designed to learn domain-invariant representations. First, it decomposes video understanding into spatial and temporal pathways for explicitly separating domain-dependent appearance cues from robust motion dynamics; Second, it constructs a Spatial-Temporal Feature Exchange module to enable cross-branch refinement and suppress domain bias; Third, we inject Gaussian feature interference to simulate feature corruption and enforce prediction-level consistency to encourage stable representations. Empirical results demonstrate that our proposal outperforms previous methods, significantly improving robustness across resolutions, domains, and species, and demonstrating outstanding generalization and transferability.

1 Introduction

Action recognition under low-resolution (LR) conditions [Oguz and Ikizler-Cinbis, 2024] remains a fundamental yet underexplored problem. The resolution mismatch between high-resolution (HR) training data and LR test inputs causes severe performance degradation, as discriminative spatial cues collapse and temporal representations become unstable. Beyond resolution mismatch, low-resolution action recognition suffers from severe cross-domain shifts, including both human-to-human transfer across environments and human-to-non-human transfer across species [Fuchs *et al.*, 2025; Maekawa *et al.*, 2021]. These shifts, driven by differences in anatomy and motion dynamics, are substantially amplified at low resolution. From Fig. 1, low-resolution action recognition is hindered by coupled challenges of spatial detail loss, non-stationary degradation, and severe cross-domain shifts across environments and species. These factors jointly destabilize feature representations and limit the robustness of existing CNN- and Transformer-based models.

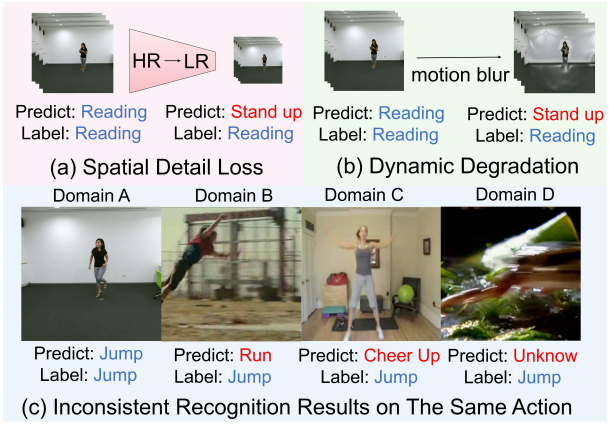

Figure 1: Overview of the key challenges in low-resolution action recognition. (a) Severe loss of discriminative spatial details makes pose and contour ambiguity common. (b) Degradation patterns are dynamic and non-stationary across frames. (c) Significant cross-domain variation arises when the same action appears in different contexts, including human-to-human and human-to-animal scenarios, resulting in substantial semantic and structural disparities.

Existing solutions address these challenges only partially. Super-resolution-based approaches [Xie *et al.*, 2023; Ledig *et al.*, 2017] reconstruct HR frames prior to recognition, but incur high computational cost and often introduce hallucinated artifacts that disrupt temporal consistency. Domain adaptation methods [Ganin *et al.*, 2017; Ghosh *et al.*, 2025] align source and target distributions, yet commonly rely on fixed target domains and struggle under dynamically varying low-resolution conditions. Multi-resolution and ensemble-based strategies [Meng *et al.*, 2020; Li *et al.*, 2025] improve robustness across scales, but remain largely human-centric and do not readily generalize to cross-species scenarios. These limitations motivate a unified framework that directly addresses resolution-induced domain shifts while enabling robust gen-

eralization across both domains and species.

To this end, we propose a generalized **Act**ion recognition framework for **L**ow-resolution conditions with **D**omain-invariant **R**epresentation learning, termed **ActLDR**. It is built upon a three-fold idea: 1) use dual-branch disentanglement and spatial-temporal asymmetry to extract spatial and temporal features from motion dynamics with resolution-sensitive attributes, 2) utilize bidirectional cross-attention to exchange spatial and temporal features, allowing each branch to suppress resolution- and domain-specific noise by attending to complementary cues, 3) exploit invariance induced consistency-driven learning and Gaussian distribution perturbations to enforce stable decision boundaries and distinguish action domain-invariant and bias representations.

The main contributions of this work are as follows: 1) We first construct a generalized action recognition architecture under low-resolutions to address resolution-induced domain shifts and cross-species discrepancies; 2) A unified dual-branch architecture with domain-invariant representations is designed to disentangle invariant features, spatial appearance, and temporal dynamics, yielding high generalization and robustness in low-resolutions; 3) We use feature-level interference and consistency regularization, which generate richer virtual domains instead of target domains, to get stable decision boundaries for mining domain-invariant representations.

2 Related Work

We review research lines: CNN-based architectures, Transformer-based video models, low-resolution action recognition, and domain generalization.

CNN-Based Action Recognition. Two-stream networks [Simonyan and Zisserman, 2014; Zhu *et al.*, 2019] separate appearance and motion cues using RGB frames and optical flow, while 3D CNNs such as C3D, I3D, R(2+1)D, SlowFast, X3D, and MF-Net learn joint spatio-temporal representations through volumetric convolutions [Tran *et al.*, 2018; Carreira and Zisserman, 2017; Feichtenhofer *et al.*, 2019; Feichtenhofer, 2020]. These architectures remain strong baselines on standard RGB benchmarks such as UCF101, HMDB51, and NTU RGB+D 120 [Soomro *et al.*, 2012; Kuehne *et al.*, 2012; Liu *et al.*, 2019]. CNN-based models strongly depend on local spatial details and fixed receptive fields. Under low-resolution conditions, discriminative appearance cues collapse, resulting in significant performance degradation and a struggle to generalize across domains, which limits the applicability of cross-resolution and cross-species scenarios.

Transformer for Video Understanding. Transformer architectures, such as ViViT, TimeSformer, and VideoMAE [Arnab *et al.*, 2021; Bertasius *et al.*, 2021; Tong *et al.*, 2022] demonstrate strong representation learning through self-attention and masked video modeling. Large-scale frameworks VideoMAE v2, OmniVec/OmniVec2, and BIKE—achieve near-saturated performance on canonical benchmarks through combinations of self-supervised learning, cross-modal pretraining, and image-to-video knowledge transfer [Wang *et al.*, 2023; Srivastava and Sharma, 2024; Wu *et al.*, 2023]. To improve structural efficiency on RGB-only, pose-guided or motion-aware Transformers [Zhang *et al.*, 2024; Long *et al.*, 2022] introduce auxiliary inductive biases to enhance fine-grained temporal modeling, while efficient attention mechanisms and hybrid architectures (e.g., Mamba-style sequence models [Li *et al.*, 2024]) aim to reduce computational overhead. Although some of these works are inspired by additional modalities (e.g., pose or language), the final inference often remains RGB-based, with auxiliary signals used only during training. Meanwhile, their attention mechanisms rely heavily on rich spatial tokens, making them fragile under severe resolution degradation.

Low-Resolution and Domain Generalization. Super-resolution-based methods [Xie *et al.*, 2023; Ledig *et al.*, 2017; Zhang *et al.*, 2019; Isobe *et al.*, 2022] reconstruct high-resolution frames with hallucinated textures and temporal inconsistencies before recognition, resulting harm to recognition accuracy and computational cost. Resolution-invariant or multi-resolution models [Meng *et al.*, 2020] mostly assume fixed degradation patterns and struggle with non-stationary resolution changes commonly observed in practice. Motion-centric or skeleton-inspired approaches [Yan *et al.*, 2018; Zhang *et al.*, 2023] mainly rely on reliable spatial cues during feature extraction and degrade sharply under ultra-low resolutions, resulting in difficulty extending to non-human species without specialized annotations. Domain generalization (DG) [Ganin *et al.*, 2017; Ghosh *et al.*, 2025] aims to improve robustness to unseen environments without access to target data. However, transferring RGB action models from humans to animals presents unique challenges, which exhibit distinct kinematic structures, motion periodicities, and anatomical constraints, even when semantic labels overlap (e.g., running or jumping).

To overcome above, this motivates our proposed dual-branch disentanglement architecture with feature interference and consistency-driven learning, which explicitly models resolution-induced perturbations and constructs domain-invariant representations without relying on additional input modalities for generalized action recognition.

3 Methodology

3.1 Preliminaries

We establish the theoretical foundations of domain-invariant representation learning for generalized low-resolution action recognition. We first introduce the core concepts of domains, domain generalization, and domain-invariant representations.

Definition 1 (Domain / Environment [Ben-David *et al.*, 2009])**.** *A domain (or environment) is defined as a joint probability distribution $P_m(X, Y)$, where $X \in \mathcal{X}$ denotes the input space and $Y \in \mathcal{Y}$ denotes the label space.*

Definition 2 (Domain Generalization [Muandet *et al.*, 2013])**.** *Domain Generalization (DG) aims to learn a predictive model using only a set of source domains $\mathcal{D} = \{\mathcal{D}_m\}_{m=1}^{M}$ that generalizes well to an unseen target domain $\mathcal{D}_t$, where $P_t(X, Y) \neq P_m(X, Y)$ for all m.*

Definition 3 (Domain-Invariant Representation [Muandet *et al.*, 2013])**.** *Given a feature extractor $\phi : \mathcal{X} \to \mathcal{Z}$, the*

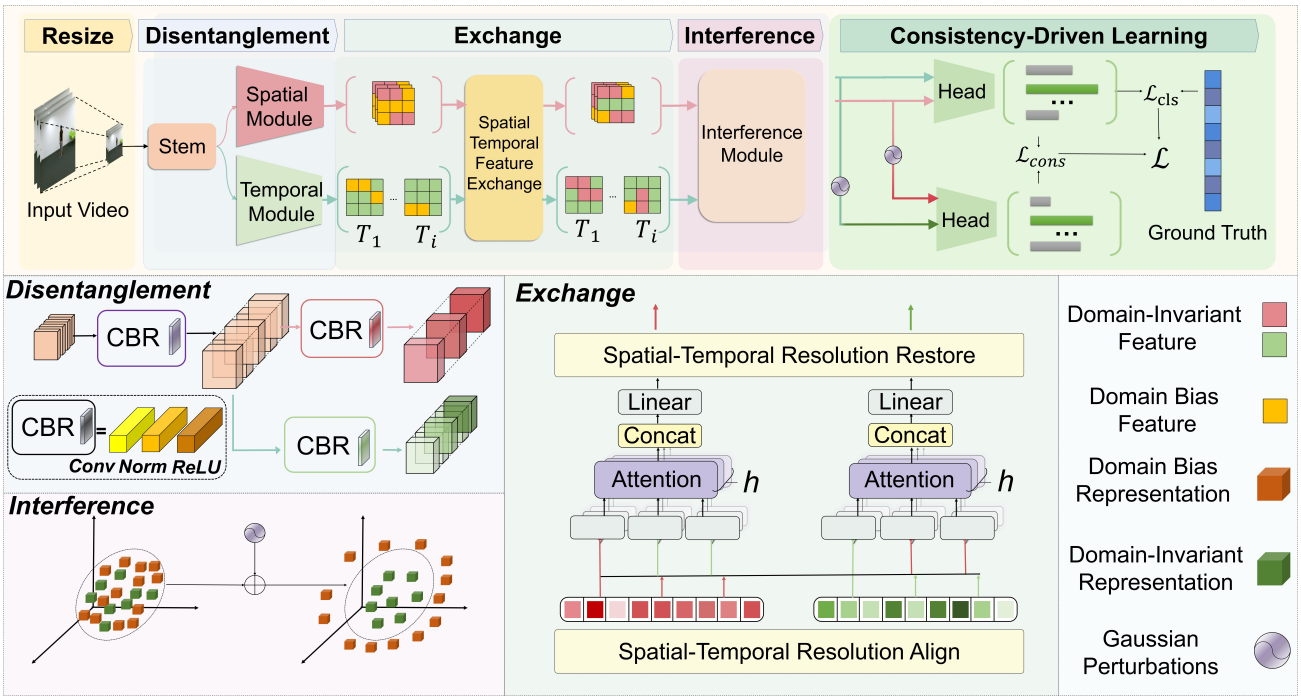

Figure 2: Overview of the ActLDR framework. Spatial and temporal features are disentangled by a dual-branch backbone, regularized through resolution-aligned spatial-temporal feature exchange, and further stabilized by Gaussian feature interference.

induced representation $Z = \phi(X)$ *is said to be domain-invariant if*

$$P(Z \mid D = m) = P(Z \mid D = n), \quad \forall m, n. \tag{1}$$

Definition 4 (Virtual Domain via Feature Intervention [Shankar *et al.*, 2018])**.** *A virtual domain* $\tilde{\mathcal{D}}$ *is induced by feature-level intervention:*

$$\tilde{Z} = Z + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{2}$$

Definition 5 (Prediction Consistency [Xie *et al.*, 2020])**.** *Prediction consistency requires the predictive distribution to remain invariant under feature-level intervention:*

$$\mathcal{L}_{cons} = D_{\mathrm{KL}}(p \,\|\, \tilde{p}), \tag{3}$$

where $p = h(Z)$ *and* $\tilde{p} = h(\tilde{Z})$ *denote predictions before and after intervention.*

These definitions characterize different notions of invariance, from marginal distribution alignment to label-sufficient representations and consistency under perturbations. However, they do not specify which components of the learned representation should be invariant, nor how such invariance can be effectively enforced under severe spatial degradation.

Following subsections, we exploit the distinct domain sensitivities of spatial and temporal information to formulate principled assumptions that guide the design of our domain-invariant representation framework. Additional formal definitions related to feature representations and domain-invariant notions (Definitions 6-8) are provided in the Appendix B.

3.2 Problem Setup

We study action recognition on low-resolutions with a domain-invariant representation. The model is trained on source domain, then generalize to an unseen target domain without accessing any target-domain data during representation learning.

Let $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$ denote a set of M source domains. Each domain $\mathcal{D}_m = \{(x_i^m, y_i^m)\}_{i=1}^{N_m}$ consists of labeled video clips sampled from an unknown joint distribution $P_m(X, Y)$, where $x \in \mathcal{X}$ denotes a low-resolution video clip with T frames and $y \in \mathcal{Y}$ is the corresponding action label.

Domain shifts naturally arise from variations in spatial resolution degradation, acquisition conditions, background statistics, motion blur, and even anatomical or behavioral differences across datasets or species. The goal is to learn a representation that captures task-relevant motion patterns while being robust to such domain-specific variations, so that it generalizes to an unseen target domain $\mathcal{D}_t$ with distribution $P_t(X, Y)$.

Following standard representation-level domain generalization protocols, the predictive model is decomposed as

$$f(x) = h(g(x)), \tag{4}$$

where $g(\cdot)$ denotes a feature extractor and $h(\cdot)$ is a lightweight classification head.

During evaluation on unseen domains, the feature extractor g is completely frozen, and only the classification head h is trained. Under this protocol, the performance on target domains reflects the degree of domain invariance of the learned

representation, as defined in the preceding section.

3.3 Domain-Invariant Representation

Domain-Invariant Feature Extraction. In low-resolution action recognition, domain shifts primarily arise from variations in spatial appearance, while temporal motion patterns remain relatively stable across domains. This motivates learning representations that emphasize domain-invariant temporal cues and suppress domain-dependent spatial variations.

Assumption 1 (Spatial-Temporal Asymmetry)**.** *In low-resolution action recognition, temporal motion patterns are more stable across domains than spatial appearance features. Let the domain discrepancy between representations Z from domains $\mathcal{D}_i$ and $\mathcal{D}_j$ be defined as*

$$\Delta(Z; i, j) disc(P(Z \mid D_i), P(Z \mid D_j)), \tag{5}$$

where D_i and D_j denote different domains (Definition 1).

Then, for spatial and temporal features Z_s and Z_t, we assume

$$\Delta(Z_t; i, j) \;\leq\; \Delta(Z_s; i, j), \quad \forall\, \mathcal{D}_i, \mathcal{D}_j. \tag{6}$$

This assumption reflects an intrinsic asymmetry in low-resolution videos: spatial appearance is highly sensitive to resolution degradation, illumination, and background variations, whereas temporal dynamics are more closely tied to action semantics and therefore exhibit greater cross-domain stability.

Assumption 2 (Label-Conditional Invariance)**.** *We assume that, given the learned representation Z, the relationship between representations and labels is invariant across domains.*

Formally, the conditional label distributions induced by Z are consistent among domains $\mathcal{D}_i$ and $\mathcal{D}_j$:

$$P(Y \mid Z, \mathcal{D}_i) \;=\; P(Y \mid Z, \mathcal{D}_j), \quad \forall\, \mathcal{D}_i, \mathcal{D}_j. \tag{7}$$

This assumption can be viewed as a conditional relaxation of domain-invariant representations (Definition 3). Rather than enforcing full marginal alignment of Z across domains, it only requires invariance of the label conditional distribution $P(Y \mid Z)$. Such conditional invariance is widely adopted in domain generalization and forms the basis for learning transferable predictors.

Under the domain generalization setting defined in Definition 2, we analyze the target-domain risk under the frozen representation evaluation protocol.

Theorem 1 (Target Risk Bound with Frozen Representation)**.** *Let $g : \mathcal{X} \to \mathcal{Z}$ be a fixed feature extractor and $\ell(\cdot, \cdot) \in [0, 1]$ a bounded loss. Under label-conditional invariance across domains, i.e., $P_s(y \mid z) = P_t(y \mid z)$, the target-domain risk of any classifier h satisfies:*

$$\mathcal{R}_t(h \circ g) \;\leq\; \mathcal{R}_s(h \circ g) + disc_{\mathcal{H}}(P_s(Z), P_t(Z)). \tag{8}$$

Proof. See Appendix C.1 □

Theorem 1 shows that, under frozen representation evaluation, target-domain performance is governed by the cross-domain discrepancy of the learned representation. As target data are unavailable during training, we reduce this discrepancy indirectly by exploiting the spatial-temporal asymmetry of domain shifts (Assumption 1), emphasizing temporally stable motion cues while suppressing domain-sensitive spatial appearance variations.

Lemma 1 (Dual-Branch Disentanglement)**.** *Under Assumption 1, the dual-branch architecture with spatial branch ϕ_s and temporal branch ϕ_t satisfies:*

$$\Delta(\phi_t(X); i, j) \leq \Delta(\phi_s(X); i, j) \tag{9}$$

for any domains $\mathcal{D}_i$ and $\mathcal{D}_j$.

Proof. See Appendix C.2 □

Lemma 1 provides a theoretical justification for disentangling spatial and temporal representations in a dual-branch architecture. Motivated by this result, we design a temporal branch to prioritize cross-domain stable motion cues, while the spatial branch captures domain-sensitive appearance information.

Proposition 1 (Temporal Features as Label-Determining Factors)**.** *Under Assumptions 1 and 2, temporal features are more closely aligned with label-determining factors than spatial features.*

Proof. See Appendix C.3 □

Proposition 1 complements Lemma 1 by indicating that temporal features are not only more domain-stable but also more predictive of action labels. Accordingly, we treat the temporal branch as the primary source of domain-invariant semantics, while spatial features are incorporated as complementary cues through controlled interaction and prediction fusion.

Domain-Invariant Representation. A remaining challenge is to ensure that the learned representation is domain-invariant not only at the feature extraction stage but also at the final representation level, which requires that decision boundaries in the representation space depend solely on domain-invariant components.

Assumption 3 (Label Stability under Feature Perturbation)**.** *Following the notion of virtual domains induced by feature-level intervention (Definition 4), we assume that small perturbations in the learned representation do not alter the underlying action semantics. Formally, for a perturbed representation $\tilde{Z} = Z + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and sufficiently small σ, the induced label distribution remains stable:*

$$P(Y \mid Z) \;\approx\; P(Y \mid \tilde{Z}). \tag{10}$$

This assumption characterizes a local stability property of domain-invariant representations in the feature space. It enables enforcing consistency between predictions under feature perturbations, which serves as a practical surrogate for decision-level domain invariance.

Theorem 2 (Invariance Induced by Feature Interference)**.** *Based on the virtual domain construction in Definition 4, under Assumption 3, minimizing the expected loss over feature-perturbed virtual domains:*

$$\min_{g,h} \; E_{(x,y)\sim P_s} E_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)} \big[\mathcal{L}(h(g(x) + \epsilon), y)\big] \tag{11}$$

encourages learning representations that are invariant to feature-level perturbations.

Proof. See Appendix C.4. □

Theorem 2 provides a principled way to promote domain-invariant representations without explicitly measuring cross-domain discrepancy. Motivated by this result, we implement Gaussian feature interference during training to enforce prediction consistency across virtual domains and improve robustness to unseen domain variations.

Corollary 1. *Under the conditions of Theorem 2, minimizing the expected loss over virtual domains leads to representations whose predictions are locally stable with respect to feature perturbations.*

Proof. See Appendix C.5. □

Corollary 1 indicates that feature interference encourages locally stable representations under domain-induced perturbations. This supports our use of Gaussian feature interference in Module 3 to improve robustness without explicitly modeling domain shifts.

Based on the prediction consistency objective defined in Definition 5, we establish the following result.

Theorem 3 (Consistency-Induced Invariant Decision Boundaries)**.** *Based on the prediction consistency objective in Definition 5, under Assumptions 2 and 3, minimizing prediction inconsistency encourages decision boundaries to rely primarily on domain-invariant components of the learned representation.*

Proof. See Appendix C.6. □

Theorem 3 provides a theoretical justification for enforcing prediction consistency as a mechanism to promote decision-level domain invariance. Motivated by this result, we design a consistency loss $\mathcal{L}_{\text{cons}}$ that penalizes prediction divergence between original and perturbed features, encouraging robust decision boundaries under domain shifts.

Theorem 4 (Generalization Bound with Frozen Representation [Ben-David *et al.*, 2009; Muandet *et al.*, 2013])**.** *If there exists a representation $\phi : \mathcal{X} \to \mathcal{Z}$ such that the conditional distribution $P(Z \mid Y, D)$ is invariant across domains and a classifier h achieves low empirical risk on source domains, then the target-domain risk is bounded by:*

$$R_t(h \circ \phi) \leq \max_{m=1,\ldots,M} R_m(h \circ \phi) + \epsilon, \tag{12}$$

where ϵ depends on domain divergence and sample complexity.

Theorem 4 provides a general theoretical context for domain generalization under frozen representations, highlighting the importance of conditional invariance for transferable prediction. Our framework instantiates this principle through a set of complementary mechanisms that progressively promote representation- and decision-level invariance.

Based on above, Fig. 2 manifests ActLDR architecture with domain-invariant feature extraction and representation. Specially, utilize dual-branch feature disentanglement and resolution-aligned spatial-temporal feature exchange to extract domain-invariant features. Then, use Gaussian feature interference and consistency-driven learning to generate domain-invariant representations for action recognition.

3.4 Dual-Branch Feature Disentanglement

Motivated by the spatial-temporal asymmetry(Assumption 1) in low-resolution videos, domain-invariant feature extraction requires explicitly separating domain-sensitive appearance cues from more stable motion patterns. To this end, we design a dual-branch architecture that disentangles spatial and temporal information streams, enabling specialized modeling of complementary features.

Given an input video $x \in R^{T \times H \times W \times C}$, ActLDR adopts a dual-branch backbone consisting of a Spatial Branch and a Temporal Branch, producing features $\phi_s(x) \in R^{d_s}$ and $\phi_t(x) \in R^{d_t}$, respectively.

- **Spatial Branch.** The spatial branch focuses on appearance cues and local context, employing convolutional kernels with larger spatial extent $(k_s \times k_s)$ and limited temporal support $(k_t = 1)$. While sensitive to domain shifts, these features provide complementary discriminative information.
- **Temporal Branch.** The temporal branch emphasizes motion dynamics and long-range temporal dependencies using larger temporal receptive fields $(k_t \geq 3)$ with compact spatial footprints $(k_s = 3)$. Such temporal patterns are more stable across domains and closely aligned with action semantics.

Through this disentanglement, the temporal branch serves as the primary carrier of domain-invariant and label-relevant information, as theoretically justified in Proposition 1, while the spatial branch complements it with appearance cues that are subsequently regularized through cross-branch interaction.

3.5 Resolution-Aligned Spatial-Temporal Feature Exchange

While dual-branch disentanglement separates complementary information streams, Theorem 1 indicates that improving target-domain performance requires reducing cross-domain representation discrepancy. We therefore introduce controlled cross-branch interaction, allowing stable temporal cues to regularize domain-sensitive spatial features.

At selected depths $l \in \{l_1, l_2, \ldots, l_L\}$, we introduce a Spatial-Temporal Feature Exchange (STFE) module. Let $\phi_s^{(l)} \in R^{d_s^{(l)}}$ and $\phi_t^{(l)} \in R^{d_t^{(l)}}$ denote the intermediate features at depth l from spatial and temporal branches, respectively.

The STFE module performs cross-branch feature exchange:

$$\tilde{\phi}_s^{(l)} = \mathcal{F}_{st}(\phi_s^{(l)}, \phi_t^{(l)}), \quad \tilde{\phi}_t^{(l)} = \mathcal{F}_{ts}(\phi_t^{(l)}, \phi_s^{(l)}) \tag{13}$$

where $\mathcal{F}_{st}$ and $\mathcal{F}_{ts}$ are learnable transformation functions implemented as:

$$\begin{aligned} \mathcal{F}_{st}(\phi_s, \phi_t) &= \phi_s + \alpha_{st} \cdot \text{Attention}(\phi_s, \phi_t) \odot \phi_t \\ \mathcal{F}_{ts}(\phi_t, \phi_s) &= \phi_t + \alpha_{ts} \cdot \text{Attention}(\phi_t, \phi_s) \odot \phi_s \end{aligned} \tag{14}$$

Here, $\text{Attention}(\cdot, \cdot)$ denotes cross-attention mechanism, $\odot$ is element-wise multiplication, and $\alpha_{st}, \alpha_{ts} \in [0, 1]$

are learnable gating coefficients. Resolution-aligned projections ensure correspondence between branches, while cross-attention highlights mutually consistent patterns. This exchange steers both branches toward a shared invariant subspace, reducing representation discrepancy across domains.

From a theoretical standpoint, this exchange explicitly reduces cross-domain representation discrepancy, as formalized in Proposition 2 (Appendix C.7).

3.6 Gaussian Feature Interference

To explicitly reduce representation sensitivity to domain shifts, we instantiate Definition 4 and Theorem 2 by constructing virtual domains via feature-level perturbation.

For an intermediate representation $h^{(l)} \in R^{d^{(l)}}$ at depth l, Gaussian Feature Interference (GFI) generates perturbed features:

$$\hat{h}^{(l)} = h^{(l)} + \epsilon^{(l)}, \quad \epsilon^{(l)} \sim \mathcal{N}(0, \sigma_l^2 I_{d^{(l)}}) \tag{15}$$

where σ_l^2 is a depth-dependent variance parameter. By minimizing the expected loss over these virtual domains, the model suppresses residual domain-sensitive components, as implied by Theorem 2.

From a theoretical perspective, this perturbation-based training enforces local smoothness in the representation space and penalizes features that are highly sensitive to perturbations. This property is formalized in Proposition 3, with a complete proof provided in Appendix C.8.

3.7 Consistency-Driven Learning Objective

Following Theorem 3, enforcing prediction consistency under feature perturbation encourages invariant decision boundaries.

We fuse spatial and temporal predictions:

$$p_f(y \mid x) = \alpha p_s(y \mid x) + (1 - \alpha) p_t(y \mid x) \tag{16}$$

where $p_s(y \mid x) = h_s(\phi_s(x))$ and $p_t(y \mid x) = h_t(\phi_t(x))$ are predictions from spatial and temporal branches, respectively, and $\alpha \in [0, 1]$ is a fusion weight.

We enforce consistency between clean and perturbed predictions:

$$\mathcal{L}_{\text{cons}} = D_{\text{KL}}(p_f(y \mid x) \| \tilde{p}_f(y \mid x)) \tag{17}$$

where $\tilde{p}_f(y \mid x)$ is the prediction from perturbed features $\tilde{\phi}_s(x)$ and $\tilde{\phi}_t(x)$.

The final training objective combines classification loss and consistency loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(p_f, y) + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}} \tag{18}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss on the fused prediction, and $\lambda_{\text{cons}} > 0$ balances empirical risk minimization and invariance promotion.

Theoretical analysis shows that this consistency constraint enforces decision boundaries that depend only on domain-invariant components (Propositions 4 and 5 provided in Appendix C.9.).

4 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of the proposed ActLDR framework under three complementary protocols: (1) In-Domain, (2) Cross-Domain Generalization, and (3) Cross-Species Transfer Learning. These protocols are designed to systematically assess resolution robustness, domain invariance, and generalization under increasingly severe distribution shifts.

An overview of the experimental results is summarized in Fig. 3, where Fig. 3(a) reports performance under the normal low-resolution setting, and Fig. 3(b) presents cross-domain and cross-species generalization results. Detailed numerical results and experimental settings are provided in appendix D.

We evaluate our method on four standard human action recognition benchmarks and one animal action dataset: **NTU RGB+D 120** [Liu *et al.*, 2019], used as the primary pretraining dataset; **UCF101** [Soomro *et al.*, 2012], **HMDB51** [Kuehne *et al.*, 2012], and **NW-UCLA** [Wang and Nie, 2014], used for within-dataset and cross-dataset evaluation; and **LoTE-Animal** [Liu *et al.*, 2023], used to assess cross-species generalization. All videos are uniformly resized to 56×56 to simulate low-resolution inputs.

4.1 In-Domain

We evaluate our method under the *In-Domain*, where models are trained and tested on the same dataset at low resolution (56×56). Implementation details, dataset-specific protocols are provided in Appendix D.1. As summarized in Fig. 3(a) and Table 1, ActLDR consistently outperforms recent CNN- and Transformer-based baselines across all evaluated RGB benchmarks under low-resolution conditions. Notably, ActLDR achieves strong and stable gains on both large-scale (NTU120) and small-scale (HMDB51, NW-UCLA) datasets, while using significantly fewer parameters than most Transformer-based models.

4.2 Cross-Domain/Species Generalization

We further evaluate the generalization ability of ActLDR under out-of-domain settings, including both cross-dataset and cross-species transfer. Implementation details, dataset-specific protocols are provided in Appendix D.2. As illustrated in Fig. 3(b) and table 2, ActLDR consistently outperforms CNN- and Transformer-based baselines when transferring from a human source dataset to unseen human datasets as well as to animal action recognition. Notably, competitive performance is maintained even under human-to-animal transfer without any target-domain supervision, indicating that the learned representations emphasize motion-centric and domain-invariant spatiotemporal cues rather than appearance-specific patterns.

4.3 Ablation Study

We conduct ablation studies on NW-UCLA and HMDB51 under the 56×56 low-resolution setting to systematically analyze the contribution of each component in our framework. All variants are trained using the same protocol. In particular, we examine the impact of dual-branch disentanglement, spatial-temporal feature exchange depth and gaussian feature

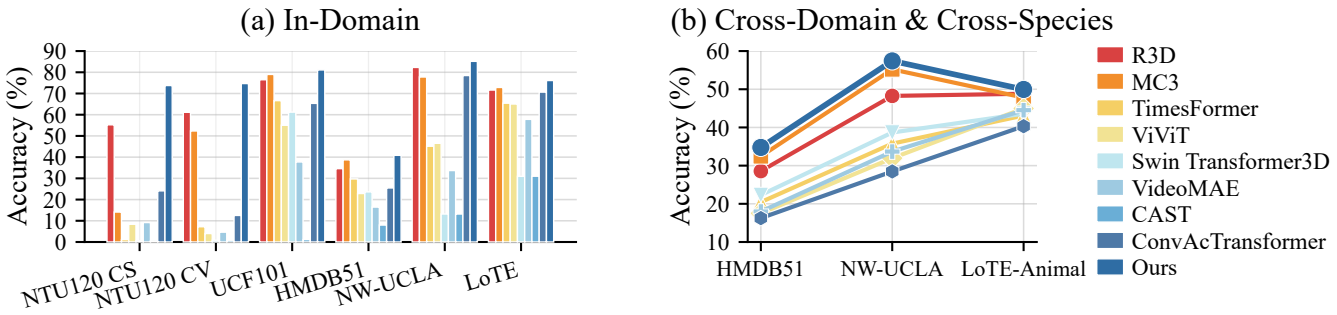

Figure 3: Performance comparison under normal and out-of-domain evaluation protocols. **(a) In-domain.** Models are trained and evaluated on the same dataset under the low-resolution setting (56×56). **(b) Cross-domain and cross-species generalization.** ActLDR achieves superior performance across both in-domain, cross-domain, and cross-species.

Methods	**Size(M)**	**NTU120**		**UCF101**	**HMDB51**	**NW_UCLA**	**LoTE**
		CS	**CV**				
R3D [Hara *et al.*, 2017]	33.2	55.25	61.14	76.48	34.58	82.29	71.63
MC3 [Tran *et al.*, 2018]	11.7	14.12	52.34	78.99	38.72	77.78	72.85
Timesformer [Bertasius *et al.*, 2021]	121.2	1.13	7.20	66.64	29.73	45.14	65.43
Vivit [Arnab *et al.*, 2021]	86.3	8.36	4.00	55.03	22.86	46.53	65.04
Swin Transformer 3D [Yang *et al.*, 2023]	87.8	0.55	0.85	61.22	23.63	13.19	30.96
VideoMAE [Tong *et al.*, 2022]	86.3	9.17	4.64	37.73	16.46	33.68	57.81
CAST [Lee *et al.*, 2023]	215.2	0.54	0.83	1.13	7.99	13.19	31.01
ConvAcTransformer [Shi and Liu, 2024]	12.9	24.11	12.53	65.35	25.48	78.47	70.65
Ours	**27.5**	**73.76**	**74.69**	**81.11**	**40.86**	**85.16**	**76.12**

Table 1: Performance comparison on NTU120,UCF101,HMDB51,NW_UCLA, LoTE Animal dataset under resolution 56×56 (Only RGB).

Methods	**NW_UCLA(%)**	**HMDB51(%)**	**LoTE(%)**
R3D	48.26	28.57	48.82
MC3	55.21	32.44	47.82
Timesformer	35.76	20.45	43.12
Vivit	31.94	17.51	44.97
Swin Transformer 3D	38.67	22.30	43.60
VideoMAE	33.68	17.98	44.52
ConvAcTransformer	28.47	16.25	40.32
Ours	**57.42**	**34.77**	**49.95**

Table 2: Cross-dataset transfer performance. Models are pretrained on UCF101 (56×56 LR) and evaluated by freezing the backbone and training only the classification head.

Variant	**NW-UCLA**	**HMDB51**
Spatial-only	61.32	16.92
Temporal-only	82.81	38.34
Dual-Branch(Ours)	**86.33**	**40.86**
w/o Feature Exchange	78.12	39.15
1 Exchange Block	83.59	35.64
2 Exchange Blocks(Ours)	**86.33**	**40.86**
3 Exchange Blocks	75.35	35.81
w/o Feature Interference	80.21	37.23
w/ Feature Interference(Ours)	**86.33**	**40.86**

Table 3: Ablation study on NW-UCLA and HMDB51 under 56×56 resolution.

interference by selectively removing each module. Quantitative results are summarized in Table 3, while a more fine-grained analysis is provided in Appendix E.

5 Conclusion

In this work, we studied generalized action recognition under low-resolution and cross-species conditions, where severe visual degradation and domain heterogeneity challenge conventional human-centric models. We proposed **ActLDR**, a unified framework that exploits spatial-temporal asymmetry by disentangling resolution-sensitive appearance cues from stable motion dynamics and enabling their interaction through feature exchange. Building on domain-invariant representation, ActLDR enforces decision-level invariance via feature perturbation and consistency-driven learning, yielding stable predictions. Extensive in-domain, cross-domain, and cross-species experiments show that ActLDR achieves strong performance, robustness, and generalization, with human-to-animal transfer highlighting the effectiveness of temporally grounded and domain-invariant representations.

Future work will strengthen temporal modeling under extreme low resolutions and improve scalability to broader actions and species. We also plan to explore large-scale pretraining and weak or multimodal supervision for better real-world generalization.

Acknowledgments

This work has been partially supported by grants from the Natural Science Foundation of Hebei Province, No.F2025203033; the National Natural Science Foundation of China, No.62276226.

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. *Computing Research Repository*, 2021.

[Ben-David *et al.*, 2009] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2009.

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

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition*, 2017.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *Computing Research Repository*, 2019.

[Feichtenhofer, 2020] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Computer Vision and Pattern Recognition*, 2020.

[Fuchs *et al.*, 2025] Michael Fuchs, Emilie Genty, Adrian Bangerter, Klaus Zuberbühler, Jean-Marc Odobez, and Paul Cotofrei. From forest to zoo: great ape behavior recognition with chimpbehave. *International Journal of Computer Vision*, pages 1–21, 2025.

[Ganin *et al.*, 2017] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 2017.

[Ghosh *et al.*, 2025] Partho Ghosh, Raisa Bentay Hossain, Mohammad Zunaed, and Taufiq Hasan. Domain generalization for improved human activity recognition in office space videos using adaptive pre-processing. *Computing Research Repository*, abs/2503.12678, 2025.

[Hara *et al.*, 2017] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *IEEE International Conference on Computer Vision*, 2017.

[Isobe *et al.*, 2022] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. *Computing Research Repository*, 2022.

[Kuehne *et al.*, 2012] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Thomas Serre. Hmdb: A large video database for human motion recognition. In *Computer Vision*, pages 571–582, 2012.

[Ledig *et al.*, 2017] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Computer Vision and Pattern Recognition*, 2017.

[Lee *et al.*, 2023] Dongho Lee, Jongseo Lee, and Jinwoo Choi. Cast: Cross-attention in space and time for video action recognition. In *Conference on Neural Information Processing Systems*, 2023.

[Li *et al.*, 2024] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *Computing Research Repository*, 2024.

[Li *et al.*, 2025] Xiaoyang Li, Wenzhu Yang, Kanglin Wang, Tiebiao Wang, and Qingsong Fei. Context-aware network based on multi-scale spatio-temporal attention for action recognition in videos. 2025.

[Liu *et al.*, 2019] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2019.

[Liu *et al.*, 2023] Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang. Lote-animal: A long time-span dataset for endangered animal behavior understanding. In *IEEE International Conference on Computer Vision*, pages 20064–20075, 2023.

[Long *et al.*, 2022] Fuchen Long, Zhaofan Qiu, Yingwei Pan, Ting Yao, Chong-Wah Ngo, and Tao Mei. Dynamic temporal filtering in video models. *Lecture Notes in Computer Science*, 13695, 2022.

[Maekawa *et al.*, 2021] Takuya Maekawa, Daiki Higashide, Takahiro Hara, Kentarou Matsumura, Kaoru Ide, Takahisa Miyatake, Koutarou D. Kimura, and Susumu Takahashi. Cross-species behavior analysis with attention-based domain-adversarial deep neural networks. *Nature communications*, 12(1), 2021.

[Meng *et al.*, 2020] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. *Lecture Notes in Computer Science*, pages 86–104, 2020.

[Muandet *et al.*, 2013] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

[Oguz and Ikizler-Cinbis, 2024] Oguzhan Oguz and Nazli Ikizler-Cinbis. Leveraging cross-resolution attention for

effective extreme low-resolution video action recognition. *Signal, Image and Video Processing*, 18(1):399–406, 2024.

[Shankar *et al.*, 2018] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *Computing Research Repository*, 2018.

[Shi and Liu, 2024] Chengcheng Shi and Shuxin Liu. Human action recognition with transformer based on convolutional features. *Intelligent Decision Technologies*, 18(2):881–896, 2024.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *Computing Research Repository*, abs/1212.0402, 2012.

[Srivastava and Sharma, 2024] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1236–1248, 2024.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Computing Research Repository*, 2022.

[Tran *et al.*, 2018] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Computer Vision and Pattern Recognition*, 2018.

[Wang and Nie, 2014] Jiang Wang and Xiaohan Nie. Northwestern-ucla multiview action 3d dataset. https://wangjiangb.github.io/my_data.html, 2014.

[Wang *et al.*, 2023] 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. *Computing Research Repository*, pages 14549–14560, 2023.

[Wu *et al.*, 2023] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6620–6630, 2023.

[Xie *et al.*, 2020] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *Computing Research Repository*, 2020.

[Xie *et al.*, 2023] Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. Desra: Detect and delete the artifacts of gan-based real-world super-resolution models. *Computing Research Repository*, 2023.

[Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI Conference on Artificial Intelligence*, 2018.

[Yang *et al.*, 2023] Yu-Qi Yang, Yu-Xiao Guo, J Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding (2023). *arXiv preprint arXiv:2304.06906*, 2023.

[Zhang *et al.*, 2019] Haochen Zhang, Dong Liu, and Zhiwei Xiong. Two-stream action recognition-oriented video super-resolution. *Computing Research Repository*, 2019.

[Zhang *et al.*, 2023] Haiping Zhang, Xinhao Zhang, Dongjin Yu, Liming Guan, Dongjing Wang, Fuxing Zhou, and Wanjun Zhang. Multi-modality adaptive feature fusion graph convolutional network for skeleton-based action recognition. *Sensors*, 23(12):5414–5414, 2023.

[Zhang *et al.*, 2024] Haosong Zhang, Mei Chee Leong, Liyuan Li, and Weisi Lin. Pgvt: Pose-guided video transformer for fine-grained action recognition. *IEEE Winter Conference on Applications of Computer Vision. IEEE Winter Conference on Applications of Computer Vision*, 2024.

[Zhu *et al.*, 2019] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. Hidden two-stream convolutional networks for action recognition. *Lecture Notes in Computer Science*, pages 363–378, 2019.