*electronics*

MDPI

*Article*

# Entropy-Gated Prediction Agreement for Two-View Video Action Recognition

**Young-Jin Park * and Hui-Sup Cho ***

Division of AI, Big Data and Block Chain, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, Republic of Korea

* Correspondence: yjpark@dgist.ac.kr (Y.-J.P.); mozart73@dgist.ac.kr (H.-S.C.)

**Abstract**

Human action recognition (HAR) often struggles to capture important temporal cues distributed across an entire video when relying solely on a single sampled clip. To overcome this limitation, this study proposes a framework that constructs two temporal views from the same video and explicitly learns the prediction consistency between them. Specifically, the prediction-level agreement (AG) loss was introduced to align the class probability distributions of the two views. In addition, conditional gating was applied to adaptively control the contribution of AG loss according to the sample-wise prediction confidence, thereby reducing unstable alignment in temporally ambiguous or information-insufficient segments. The proposed framework was evaluated using both convolutional neural network (CNN)- and Transformer-based backbones on three representative action-recognition benchmark datasets, and it generally improved the performance over the single-view baseline across backbone–dataset combinations. Further empirical analyses, including training behavior, motion magnitude, temporal prediction stability, and qualitative case studies, were conducted to examine the effectiveness and behavior of the proposed two-view framework from multiple perspectives.

**Keywords:** human action recognition; two-view learning; temporal consistency; prediction-level agreement; entropy-based gating

## 1. Introduction

Human action recognition (HAR) has been widely studied as a key research topic in various application domains, including intelligent surveillance, sports analysis, robotic vision, and autonomous systems. Because videos inherently contain spatiotemporal information along the temporal axis, HAR presents unique challenges compared to single-image recognition, as it requires modeling not only the spatial appearance but also the contextual temporal dynamics of actions [1,2].

Early studies employed frame-level features extracted using 2D convolutional neural networks (CNNs) [3]; however, these approaches were limited in effectively capturing temporal information. Subsequently, 3D CNN-based architectures were introduced to jointly model spatial and temporal features [4]. For example, SlowFast proposes a dual-pathway structure with different temporal resolutions to learn semantic information and dynamic motion patterns [5]. More recently, Transformer-based models have been adopted for HAR, and approaches such as TimeSformer have demonstrated the ability to model global spatiotemporal relationships by disentangling spatial and temporal attention mechanisms [6,7].



Academic Editors: Emmanuele Barberi and Emanuele Guardiani

Received: 4 May 2026
Revised: 22 June 2026
Accepted: 25 June 2026
Published: 30 June 2026

Despite these advances, several HAR models still adopt a single-view training strategy in which only one clip is sampled from a video during training. Although this approach is practical in terms of computational efficiency and implementation simplicity, it may fail to sufficiently capture the action cues distributed throughout the entire video. In particular, when key actions are concentrated within specific temporal regions, or when certain segments contain incomplete information, models trained on a single segment may yield uncertain predictions. This ultimately raises the question of how different temporal segments within the same video can be effectively utilized during training.

Existing studies that leverage temporal information in video-based action recognition can be categorized into three types. First, multi-view learning exploits the relationships among different views to improve the representation robustness. Second, multi-segment or multi-branch approaches combine multiple temporal segments or information sources to enhance the video representation and prediction performance. Third, consistency-based learning employs the prediction consistency across different inputs, repeated inferences, or multiple models as a learning signal. However, these studies have mainly focused on representation alignment, information fusion, and prediction stabilization. The problem of directly learning the relationship between prediction distributions obtained from different temporal views of the same video in a fully supervised setting has been relatively underexplored.

To address this limitation, this paper proposes a two-view framework comprising two temporal views extracted from the same video. Different temporal segments can provide complementary information for the same action instance; however, the clarity of action cues and reliability of predictions may vary across segments. Some segments may fail to adequately capture the core action, resulting in uncertain and potentially unstable predictions, depending on their temporal position within the same video. To alleviate this issue, this paper incorporated agreement (AG) regularization into the training objective such that the prediction distributions of the two temporal views maintained consistent relationships. However, applying the same alignment strength to all samples may cause excessive alignment, even for segments with insufficient information.

Therefore, this paper introduced conditional gating, which adaptively controls the strength of the alignment based on prediction uncertainty. This design allows AG regularization to be actively utilized for reliable segments, while suppressing excessive alignment for ambiguous segments. Therefore, this study proposes a unified learning framework that combines a temporal two-view structure, AG regularization, and conditional gating.

The main contributions of this study are summarized as follows:

1. This study proposes a two-view framework for video action recognition that directly aligns the prediction probability distributions of different temporal views extracted from the same video. AG regularization is introduced to encourage prediction-level agreement, while conditional gating adaptively controls its contribution based on sample-wise prediction confidence.
2. Through experiments on three representative action-recognition benchmark datasets using both CNN- and Transformer-based backbones, this paper demonstrates that the proposed method provides consistent performance improvements across various model architectures and datasets.
3. The proposed two-view framework is analyzed through group-wise training metrics, motion magnitude-based performance analysis, prediction-consistency and reliability analysis, and qualitative case analysis. In addition, the single-view baseline, nAG without agreement regularization, and the proposed gAG are compared to distinguish the effects of two-view supervision and gated agreement regularization.

The remainder of this paper is organized as follows. Section 2 reviews the related work and discusses how our approach differs from the existing methods. The proposed method is described in Section 3. Section 4 presents and analyzes the experimental setup and results. Section 5 discusses the key implications and limitations of this study. Finally, Section 6 concludes the paper and suggests directions for future research.

## 2. Related Work

This section reviews previous studies related to the proposed method from three perspectives: multi-view learning, multi-branch or multi-segment action recognition, and consistency-based learning.

First, multi-view learning was developed to improve the representation robustness by utilizing different views generated from the same sample. Refs. [8,9] are contrastive learning methods that learn representations from augmentation-based views, whereas Ref. [10] learns the relationships between different temporal clips in videos through contrastive learning and aligns them in the representation space. In contrast, the present study directly addresses the relationship between the prediction probability distributions at the final classification stage rather than focusing on representation alignment.

Second, in the field of HAR, multi-branch and multi-segment architectures have been actively studied to exploit multiple temporal information sources. Refs. [3,5,11] are representative approaches that utilize different modalities, temporal resolutions, or segment-level information. In particular, SlowFast is based on a dual-pathway architecture that learns semantic and motion-related information separately, and stable performance has been reported on benchmark datasets such as UCF101 [12] and HMDB51 [13] and HAA500 [14]. In contrast, TimeSformer is a Transformer-based backbone that learns video representations using only space-time self-attention without convolution. However, these approaches mainly focus on feature fusion, dual-pathway modeling, or prediction aggregation and do not directly incorporate the relationship between prediction distributions from different temporal views into the training objective.

Table 1 summarizes representative previous results to contextualize dataset difficulty. Since these methods differ in pretraining data, input modality, backbone scale, test-time views, and evaluation protocols, they are not used for direct comparison. The main evaluation is instead conducted under the controlled Baseline–nAG–gAG setting.

**Table 1.** Representative previously reported Top-1 accuracy results on UCF101, HMDB51, and HAA500, used as reference performance levels for each dataset.

| Dataset | Method | Top-1 (%) | Ref. |
|---|---|---|---|
| UCF101 | VideoMAE V2/ViT-g | 99.05 | [15] |
| UCF101 | 3D CNN + BERT temporal modeling | 98.69 | [16] |
| HMDB51 | VideoMAE V2/ViT-g | 86.08 | [15] |
| HMDB51 | 3D CNN + BERT temporal modeling | 85.10 | [16] |
| HAA500 | TSN Two-Stream | 64.40 | [14] |
| HAA500 | I3D Three-Stream | 49.87 | [14] |

Third, consistency-based learning has been widely studied in the context of semi-supervised learning and regularization. Refs. [17–21] are representative methods that use the prediction consistency across different conditions, repeated inferences, and multiple models as learning signals. In video action recognition, attempts have been made to exploit temporal consistency and prediction stability, as described in Ref. [22]. In addition, recent studies have used temporal-view consistency for test-time adaptation in video action recognition [23], temporal consistency as a learning signal for domain adaptation [24], and confidence-weighted divergence to align the prediction distributions between different views [25]. However, these studies focused on test-time adaptation, domain adaptation,

and multicamera settings. Therefore, they differ from the present study in both problem setting and application, as our method directly aligns class-wise prediction probability distributions between two views extracted from the same video in a fully supervised setting and adaptively controls the alignment strength through sample-wise conditional gating.

Recent video action recognition studies have addressed view bias and view ambiguity in multi-view settings. Ref. [26] proposed a debiasing strategy to reduce excessive dependence on specific temporal thumbnails by strengthening semantic action representations. Ref. [27] addressed the view fuzz problem using a dual-recommendation disentanglement network that separates view-related and action-related features. These studies are meaningful in that they improve robustness against view-dependent representation bias. However, they mainly focus on multi-view feature disentanglement, representation correction, or information fusion across different views.

In contrast, this study focuses on temporal-view consistency within a single RGB video rather than camera-view variation. The proposed framework constructs two temporal views with different temporal coverage and regularizes their prediction-level agreement without modifying the backbone architecture. Entropy-based gating is further introduced to reduce the influence of uncertain view pairs, thereby mitigating excessive regularization when one view contains insufficient action evidence. In this respect, the proposed method provides a lightweight temporal consistency regularization strategy for supervised action recognition while preserving a single-view inference structure.

## 3. Methods

### *3.1. Overview of the Proposed Framework*

The proposed framework constructs two temporal views from the same video and jointly learns supervised classification for the anchor view and AG between the two views. Here, AG refers to a learning signal that encourages the prediction probability distributions of the two views to maintain consistent class-level semantics. Figure 1 illustrates the overall training pipeline.

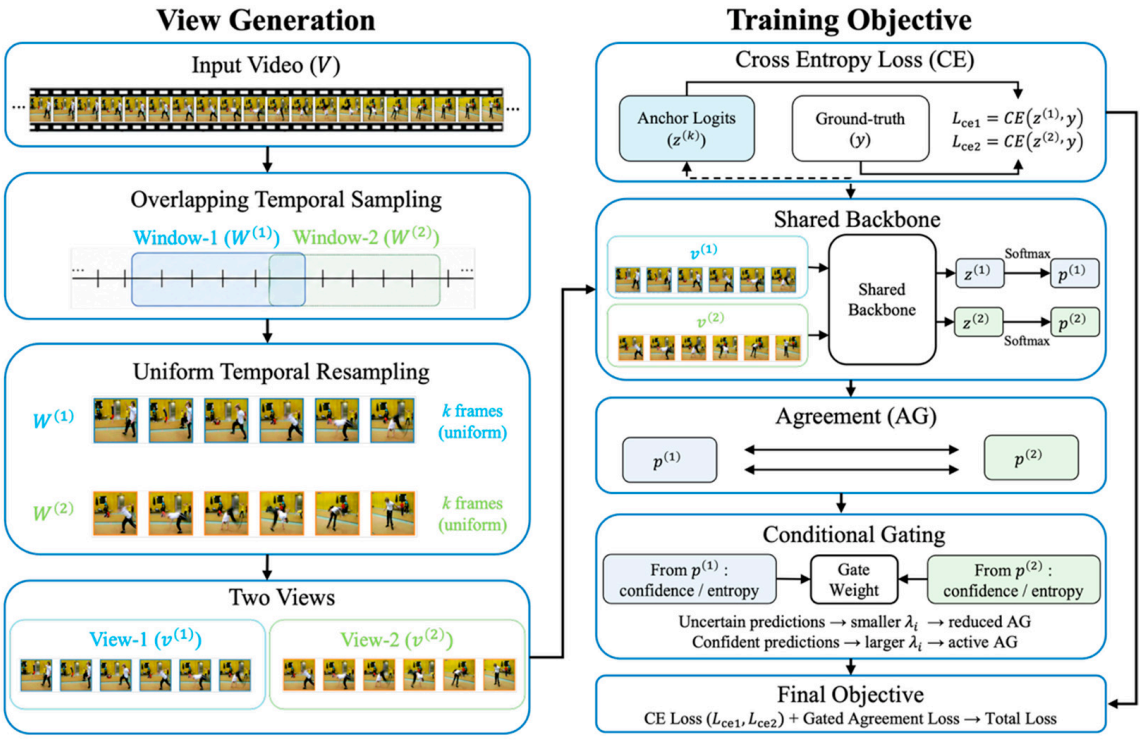


**Figure 1.** Overall framework of the proposed two-view framework learning method.

First, two temporal windows are extracted from the input video using overlapping temporal sampling. Each window is then converted into a fixed-length clip through uniform temporal resampling to form the input for the two-view framework. The generated views are fed into a shared backbone network. Both views are used for supervised classification with the same ground-truth label, and their prediction distributions are additionally used to compute the AG loss. Accordingly, the two views jointly contribute to the classification objective, while the AG term encourages cross-view prediction consistency.

Furthermore, AG loss is applied to prevent the output probability distributions of the two views from becoming excessively different, thereby reducing the prediction variation caused by changes in the temporal position. However, because partial observations may lead to uncertain predictions in certain temporal views, this study introduces conditional gating to adjust the contribution of AG loss according to the sample-wise prediction confidence. Consequently, the proposed framework is designed to jointly learn the classification performance and prediction stability from different temporal observations of the same video.

In summary, this paper proposes a framework that learns AG regularization between two views in a fully supervised setting. In addition, this paper introduces adaptive gating based on prediction uncertainty to mitigate the excessive regularization that may occur when the same alignment constraint is uniformly imposed on all view pairs. This allows the proposed method to explicitly learn the prediction relationships between the two views while flexibly adjusting the alignment strength for each sample.

### *3.2. View Generation*

#### 3.2.1. Overlapping Temporal Sampling

This study trained a two-view framework by generating two temporal windows with different time ranges from a single-input video. The input video $V$ is defined in Equation (1), where $x_t$ denotes the frame at temporal index $t$, and $T$ denotes the total number of frames in the video.

$$V = \{x_0, x_1, \ldots, x_{T-1}\} \tag{1}$$

The center index $m$ of the entire video and the center-based extension length $o$ are defined in Equations (2) and (3), respectively, where $\alpha \in [0, 1]$ is a hyperparameter that controls the degree of overlap. Here, $\alpha$ is not the actual overlap ratio itself, but a hyperparameter that controls how much each window is extended around the center. Therefore, the actual overlap ratio may vary depending on the video and window lengths.

$$m = \left\lfloor \frac{T}{2} \right\rfloor \tag{2}$$

$$o = \text{round}(\alpha m) \tag{3}$$

The two temporal windows, $W^{(1)}$ and $W^{(2)}$, are defined in Equations (4) and (5), respectively. Accordingly, the first view always included the earlier part of the video, whereas the second view always included the later part, with both views partially sharing a temporal context around the center.

$$W^{(1)} = [0,\ \min(T-1, m+o)] \tag{4}$$

$$W^{(2)} = [\max(0, m-o),\ \ T-1] \tag{5}$$

The actual degree of overlap between the two views is measured using the overlap ratio $r$, as defined in Equation (6).

$$r = \frac{|\ W^{(1)} \cap W^{(2)}\ |}{\min(|\ W^{(1)}\ |,\ \ |\ W^{(2)}\ |)} \tag{6}$$

Although it is possible to fix the overlap ratio directly, this strategy requires a joint design of the positions and lengths of the two windows. This makes view construction less intuitive and complicates determining whether performance differences are caused by the overlap itself or by differences in the absolute positions or lengths of the windows. In contrast, the proposed strategy maintains simple and consistent view construction: the first view reflects the earlier part of the video, the second view reflects the later part, and $\alpha$ controls only the shared temporal context around the center.

#### 3.2.2. Uniform Temporal Resampling

Although the two temporal windows $W^{(1)}$ and $W^{(2)}$ cover different temporal ranges, the backbone network requires fixed-length input clips. Therefore, this paper applies uniform temporal resampling to convert each window into a clip with the same number of frames. For each window $W^{(k)} = [s_k, e_k]$, $N$ frames are uniformly sampled over the entire interval, where $N$ denotes the number of frames in the input clip generated from each temporal window. Figure 2 shows an example of resampling a video clip.

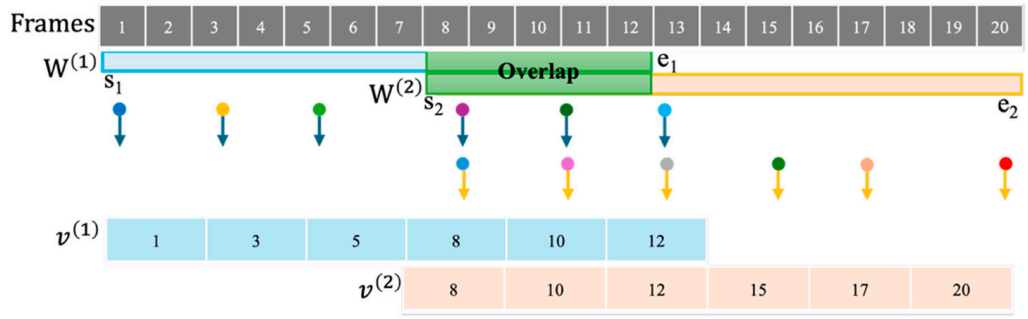


**Figure 2.** Example of uniform temporal resampling for two temporal windows ($T = 20$, $N = 6$). Colors distinguish the views and sampled frames, while arrows indicate the temporal direction.

In this process, the integer index of the $i$-th frame selected from the $k$-th window is denoted by $\hat{t}_i^{(k)}$. This index is obtained by converting the uniformly spaced sampling position within $[s_k, e_k]$ to the nearest integer frame index, as shown in Equation (7). The corresponding frames are then arranged in temporal order to form a fixed-length clip $v^{(k)}$ for each view, as defined in Equation (8).

$$\hat{t}_i^{(k)} = \text{round}\left(s_k + \frac{i}{N-1}(e_k - s_k)\right), i = 0, \ldots, N-1 \tag{7}$$

$$v^{(k)} = \{x_{\hat{t}_0^{(k)}}, x_{\hat{t}_1^{(k)}}, \ldots, x_{\hat{t}_{N-1}^{(k)}}\} \tag{8}$$

This process evenly reflects the temporal range of each window, while converting two views extracted from different temporal segments into fixed-length clips with the same input format for the backbone network. The generated two-view clips are then used in the training objective, which combines classification loss and AG regularization.

### *3.3. Training Objective*

#### 3.3.1. Two-View Classification with Shared Backbone

The two views generated from the same video produce predictions through a shared backbone network, forming the basis for supervised classification and AG regularization. Specifically, the two views $v^{(1)}$ and $v^{(2)}$ are fed into the same backbone $f_\theta$, which outputs prediction logits $z^{(1)}$ and $z^{(2)}$, respectively, as defined in Equation (9). Because the two views share the same backbone parameters, the differences between the outputs are caused by differences in the temporal context rather than differences in the model structure.

$$z^{(k)} = f_\theta(v^{(k)}) \tag{9}$$

In this study, $v^{(1)}$ and $v^{(2)}$ were used as two temporal views, and the cross-entropy (CE) loss with the ground-truth label $y$ was applied to the prediction of each view, as shown in Equations (10) and (11).

$$L_{\text{ce1}} = CE(z^{(1)}, y) \tag{10}$$

$$L_{\text{ce2}} = CE(z^{(2)}, y) \tag{11}$$

Both temporal views contribute to supervised classification. During training, they are processed sequentially: the gradient from $L_{ce2}$ is first accumulated using the second view, after which $L_{ce1}$ and the AG loss are computed using the first view and the stored prediction of the second view. Thus, the AG regularization is applied during the optimization of the first-view prediction. The next subsection describes how consistency between the prediction distributions of the two views is encouraged.

#### 3.3.2. Prediction-Level Agreement

Although these two views contain different temporal contexts, they must maintain consistent predictions regarding class-level semantics. However, conventional CE-based learning focuses on improving the classification accuracy of individual views and does not directly control the prediction consistency between them. Consequently, the model may overly depend on local cues from a specific temporal position or produce unstable predictions when the temporal position changes.

To alleviate this limitation, this paper introduced AG regularization between the two views. Each view produces prediction logits through the shared backbone, which were then converted into probability distributions for AG loss computation. Specifically, the prediction distribution $p^{(k)}$ of each view is defined using softmax with temperature $\tau_{agr}$, as shown in Equation (12).

$$p^{(k)} = \text{softmax}(\frac{z^{(k)}}{\tau_{\text{agr}}}) \tag{12}$$

Both views are supervised using the same ground-truth label and therefore contribute to the classification objective. For AG regularization, the prediction of the second view is stored and used as a detached reference when the first view is processed. Consequently, the AG gradient is propagated through the first-view prediction only, while both prediction distributions are used to measure cross-view inconsistency.

To this end, the discrepancy between the two distributions was computed using symmetric Kullback–Leibler divergence (KL-Div) [28,29], as defined in Equation (13). where $L_{agr,i}$ is the AG loss of the $i$-th sample. This term measures the bidirectional discrepancy between the two prediction distributions, while the detached second-view prediction serves as a fixed reference during optimization. It thereby regularizes the first-view prediction to reduce prediction variation caused by temporal differences within the same video.

$$L_{\text{agr},i} = \frac{1}{2}[D_{\text{KL}}(p_i^{(1)} \parallel p_i^{(2)}) + D_{\text{KL}}(p_i^{(2)} \parallel p_i^{(1)})] \tag{13}$$

By using temperature-scaled probability distributions, the proposed method reduces excessive alignment toward only a few dominant classes caused by overly sharp predictions, and instead captures the overall relationship between the two distributions more smoothly. However, if this regularization is applied with the same strength to all the samples, noisy alignment signals from information-insufficient or uncertain views may also be enforced. To address this issue, the next subsection introduces conditional gating to adaptively control the contribution of AG loss according to the sample-wise prediction reliability.

3.3.3. Entropy-Based Gating

Instead of applying AG regularization equally to all samples, this paper introduced conditional gating to adjust its contribution according to sample-wise prediction reliability. For each sample i, a gate weight $\lambda_i \in [0, 1]$ is defined and multiplied by the sample-wise AG loss to form the gAG loss. A larger $\lambda_i$ increases the contribution of AG loss for that sample, whereas a smaller value reduces its influence. In addition, for samples whose constructed temporal windows cannot share a valid overlapping region, the corresponding gate weight is set to zero, thereby preventing unreliable agreement constraints from contributing to the training objective.

Thus, conditional gating mitigates the performance degradation caused by excessive alignment by adjusting the strength of AG regularization according to the prediction reliability. The gAG loss for batch $B$ is defined by Equation (14). Instead of simply averaging the AG losses of all samples in the batch, different weights were assigned based on their reliability to construct the batch-level agreement regularization.

A sample-wise gating mechanism suppresses unreliable agreement signals. Samples with gate weights below the threshold ($\tau$) are excluded from the gAG loss, and their contribution is set to zero. That is, AG regularization is effectively applied only to samples satisfying $\lambda_i \geq \tau$. Consequently, conditional gating flexibly adjusts the contribution of AG regularization according to the sample-wise prediction reliability and alleviates the performance degradation caused by excessive alignment.

$$\widetilde{L}_{\text{agr}} = \frac{1}{B}\sum_{i=1}^{B} \lambda_i L_{\text{agr},i} \tag{14}$$

3.3.4. Final Objective

The final objective function is defined as the sum of the classification loss for the anchor view and the gAG loss, as shown in Equation (15). Here, $L_{ce}$ is the CE loss for the anchor view, and $\beta$ is a hyperparameter that controls the global contribution of the gAG loss. That is, while $\beta$ determines the overall strength of AG regularization, $\lambda_i$ can be interpreted as a local coefficient that determines how much the AG loss should be trusted and reflected for each sample.

$$L = \frac{1}{2}(L_{\text{ce1}} + L_{\text{ce2}}) + \beta\widetilde{L}_{\text{agr}} \tag{15}$$

Consequently, the proposed objective was designed to maintain the supervised classification criterion while more stably learning two-view prediction consistency through reliability-based selective AG regularization.

## 4. Experiments

In this section, this paper evaluates the effectiveness of the proposed two-view framework using two backbone networks and three benchmark datasets: TimeSformer-UCF101 (TSF-UCF), TimeSformer-HMDB51 (TSF-HMDB), TimeSformer-HAA500 (TSF-HAA), SlowFast-UCF101 (SLF-UCF), SlowFast-HMDB51 (SLF-HMDB), SlowFast-HAA500 (SLF-HAA). This paper first describes the datasets, backbones, and experimental settings, and then presents the performance comparison, convergence behavior, and further analyses. All experiments were conducted using the official split and evaluation protocol of each dataset, and Top-1 accuracy (Acc.) was used as the primary performance metric.

### *4.1. Materials and Experimental Settings*

4.1.1. Datasets

Experiments were conducted on three representative video-based human action recognition (HAR) benchmarks: UCF101, HMDB51, and HAA500. UCF101 contains 101 action

classes with relatively clear scene context and stable camera motion, making it a comparatively less challenging benchmark. HMDB51 consists of 51 action classes and provides a more difficult setting due to low video quality, camera motion, occlusion, and large intra-class variations. HAA500 further increases the difficulty by introducing 500 fine-grained atomic action classes, where subtle pose and motion differences are critical for classification. The detailed statistics of each dataset are summarized in Table 2.

**Table 2.** Dataset statistics used in the experiments.

| Dataset | Class | Training | Testing | Total |
|---|---|---|---|---|
| UCF101 | 101 | 9537 | 3783 | 13,320 |
| HMDB51 | 51 | 3570 | 1530 | 5100 |
| HAA500 | 500 | 8000 | 2000 | 10,000 |

#### 4.1.2. Backbones

In this study, TimeSformer and SlowFast pre-trained on Kinetics-400 [30] were used as backbone networks [31–34]. TimeSformer is a Transformer-based backbone that directly learns global dependencies along spatial and temporal dimensions using self-attention. SlowFast is a CNN-based backbone that jointly models long-term semantic and short-term dynamic information using a two-pathway architecture with different temporal resolutions. By evaluating both Transformer- and CNN-based models with different architectural characteristics, this paper aims to examine whether the proposed two-view framework can be applied generally rather than being limited to a specific backbone family. The detailed backbone configurations are listed in Table 3.

**Table 3.** Backbone configurations used in the experiments.

| Backbone | Pretraining | Model Variant | Input Clip |
|---|---|---|---|
| TimeSformer | Kinetics-400 | divST_8×32_224_K400 | 32 frames |
| SlowFast | Kinetics-400 | R101_16×8 | 64 frames (fast pathway) |

#### 4.1.3. Common Experimental Settings

For a fair comparison, the baseline and two-view framework models were trained under the same experimental settings. All experiments were conducted for 30 epochs using a batch size of 2, stochastic gradient descent (SGD) optimizer, initial learning rate of $10^{-3}$, and dropout rate of 0.5. Step decay was used for learning rate scheduling, where the learning rate was multiplied by 0.1 every five epochs. In addition, to reduce the memory usage, the two temporal views were processed sequentially, and gradient accumulation was used before updating the shared backbone parameters. The implementation details are presented in Table 4.

**Table 4.** Implementation environment.

| Item | Setting |
|---|---|
| OS | Ubuntu 20.04 LTS |
| Framework | Anaconda 22.9.0, CUDA 11.1, PyTorch 1.8.0 |
| Python | Python 3.8 (TimeSformer)/Python 3.7 (SlowFast) |
| GPU | Nvidia RTX 3090 Ti (UCF101, HAA500)/Nvidia RTX TITAN (HMDB51) |

### *4.2. Baseline Performance*

To verify the effectiveness of the two-view framework, single-input baseline models were first trained and used as references for all the comparisons. For both backbones, the baseline models were trained using the same dataset and training conditions. Figure 3 shows

the training loss and test-set Top-1 accuracy curves across checkpoints. Table 5 reports the best test Top-1 accuracy of the single-view baseline for each backbone–dataset combination.

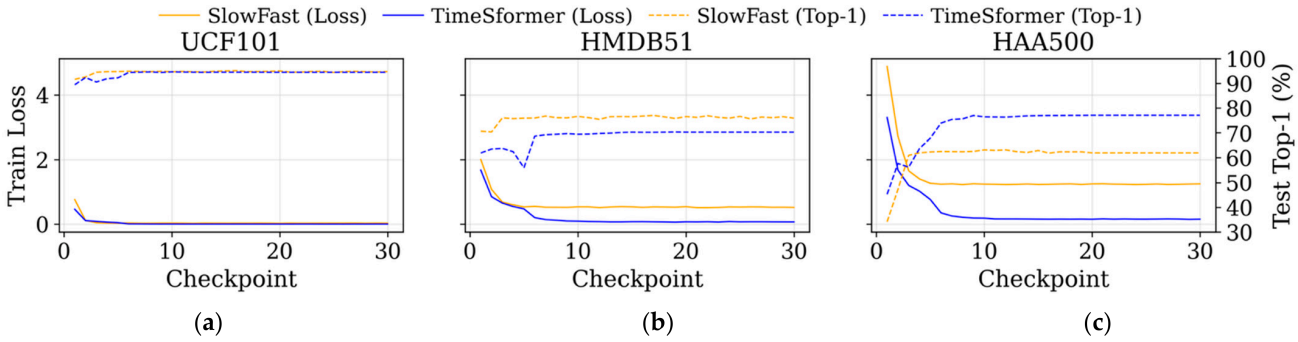


**Figure 3.** Training loss and test Top-1 accuracy curves of the baseline models: (**a**) UCF101; (**b**) HMDB51; (**c**) HAA500.

**Table 5.** Performance comparison of the single-view baseline and the two-view nAG and gAG models. For the baseline, Top-1 denotes the single-view Top-1 accuracy. For nAG and gAG, Top-1 denotes the two-view Top-1 accuracy, Gain denotes the improvement over the baseline, and Time denotes the average training time per epoch (hour/epoch) measured under the same experimental environment.

| Backbone–Dataset | Baseline | nAG | | | gAG | | |
|---|---|---|---|---|---|---|---|
| | Top-1 | Top-1 | Gain | Time | Top-1 | Gain | Time |
| TSF-UCF | 94.66 | 96.46 | +1.80 | 0.88 | 96.59 | +1.93 | 0.86 |
| SLF-UCF | 95.14 | 95.82 | +0.68 | 1.44 | 95.88 | +0.74 | 1.21 |
| TSF-HMDB | 70.33 | 72.81 | +2.48 | 0.42 | 73.07 | +2.74 | 0.42 |
| SLF-HMDB | 77.03 | 78.02 | +0.99 | 0.81 | 79.27 | +2.24 | 0.80 |
| TSF-HAA | 77.10 | 82.50 | +5.40 | 0.95 | 82.10 | +5.00 | 0.93 |
| SLF-HAA | 63.20 | 69.00 | +5.80 | 2.80 | 68.95 | +5.75 | 2.72 |

Based on official split 1, each epoch checkpoint was evaluated on the test set, and the model with the highest Top-1 accuracy was selected as the final baseline. Overall, the training tended to converge stably across all backbone–dataset combinations. These baseline results are compared with the training results of the two-view framework described in the following subsection.

### *4.3. Two-View Performance*

The performance of the two-view framework was evaluated against the single-view baseline. Along with the proposed gAG model, nAG, which denotes two-view training without the gAG loss, was included to distinguish the effect of two-view supervision from that of agreement regularization.

The two-view framework uses two temporal windows generated from a single video and converted into fixed-length clips. The parameter $\alpha$ controls temporal window generation rather than directly representing the actual overlap ratio. For a fair comparison, the same hyperparameter settings were used across all experiments: $\alpha$ was fixed at 0.5, the softmax temperature for the AG loss was set to 2.0, and the global AG weight $\beta$ was set to 0.05. For all gAG experiments, the gating threshold ($\tau$) was fixed at 0.6.

In Figure 4, for both backbones, the training loss decreases rapidly in the early stage and then remains low, whereas the Top-1 accuracy generally increases and converges. This suggests that the proposed two-view framework forms a stable optimization process across different backbones. The degree of performance improvement and detailed behavior of the learning curves differ depending on the backbone and dataset. TimeSformer tends to show a relatively stable performance improvement as training progresses, whereas SlowFast shows rapid convergence after early fluctuations. Nevertheless, the learning

curves generally exhibit stable convergence across all combinations, indicating that the proposed framework can provide consistent training behavior across different backbones.

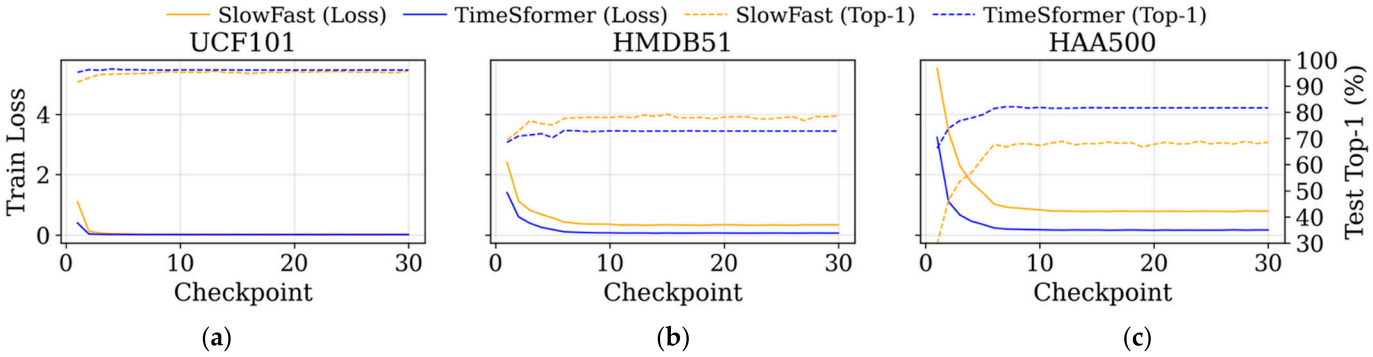


**Figure 4.** Training loss and test Top-1 accuracy curves of the best two-view framework models (gAG): (**a**) UCF101; (**b**) HMDB51; (**c**) HAA500.

Table 5 compares the performance of the single-view baseline with the two-view nAG and gAG models. Overall, both nAG and gAG outperform the baseline across all backbone–dataset combinations, indicating that the two-view learning framework can exploit richer temporal cues than single-view training.

Compared with the baseline, gAG improved Top-1 accuracy by +1.93 and +0.74 percentage points on UCF101, +2.74 and +2.24 percentage points on HMDB51, and +5.00 and +5.75 percentage points on HAA500 for TimeSformer and SlowFast, respectively. The largest gains were observed on HAA500, suggesting that two-view supervision can be particularly useful for complementing temporal evidence in fine-grained atomic action recognition with short video durations. The reported Time value represents the average training time per epoch measured under the same experimental environment and serves as a practical reference for comparing the training costs of nAG and gAG.

Overall, Table 5 shows that two-view learning consistently improves performance over the single-view baseline. The detailed differences between nAG and gAG, including checkpoint-level stability and prediction consistency, are further analyzed in the next subsection.

#### 4.3.1. Prediction-Consistency and Reliability Analysis

In this subsection, we analyze the effect of gated agreement on prediction consistency and reliability under the same two-view setting. nAG and gAG are compared using Top-1 accuracy, symmetric KL divergence, and prediction confidence. KL-Div. measures the discrepancy between the class probability distributions of the two temporal views, where a lower value indicates higher cross-view prediction consistency. Confidence denotes the predicted probability of the selected class, and is used to examine whether the performance change is associated with more reliable predictions.

First, ΔTop-1 shows that the performance effect of gAG is dependent on the backbone–dataset combination. The largest improvement is observed in SLF-HMDB (+1.25), whereas TSF-HAA shows the largest decrease (−0.40). This suggests that gated agreement can improve recognition accuracy when the two temporal views provide reliable complementary evidence, but may be less effective when the shared temporal evidence is insufficient or ambiguous.

Second, ΔStd. indicates that the effect of gAG on checkpoint-level stability is not uniform. Although gAG reduces the standard deviation in several settings, the average ΔStd. is close to zero, suggesting that gated agreement affects stability differently depending on the backbone and dataset.

Third, ΔKL-Div. does not decrease consistently across all settings. SLF-HMDB shows both improved Top-1 accuracy and reduced KL-Div., indicating a favorable case where prediction consistency and recognition performance improve together. In contrast, TSF-HAA shows reduced KL-Div. but lower Top-1 accuracy, suggesting that lower prediction discrepancy alone may not guarantee better performance when the shared temporal evidence is not sufficiently discriminative.

Finally, ΔConf. remains close to zero in most settings. SLF-HMDB achieves the largest Top-1 improvement with almost no confidence change, whereas TSF-HAA shows decreases in both ΔConf. and ΔTop-1. This suggests that the effect of gAG is not simply due to increased prediction confidence, but is more related to the reliability of temporal-view evidence.

The differences between gAG and nAG in terms of Gain, KL-Div., Confidence, and Entropy are summarized in Table 6.

**Table 6.** Auxiliary analysis of gAG relative to nAG. ΔTop-1, ΔStd., ΔKL-Div., and ΔConf. denote the gAG−nAG differences in Top-1 accuracy, checkpoint-level standard deviation, symmetric KL divergence, and prediction confidence, respectively. Negative ΔKL-Div. values indicate reduced two-view prediction discrepancy.

| Backbone–Dataset | ΔTop-1 | ΔStd. | ΔKL-Div. | ΔConf. |
|---|---|---|---|---|
| TSF-UCF | +0.13 | −0.02 | +0.010 | +0.0069 |
| SLF-UCF | +0.06 | +0.17 | +0.003 | +0.0005 |
| TSF-HMDB | +0.26 | −0.14 | +0.004 | −0.0079 |
| SLF-HMDB | +1.25 | +0.38 | −0.002 | +0.0002 |
| TSF-HAA | −0.40 | −0.40 | −0.003 | −0.0054 |
| SLF-HAA | −0.05 | −0.01 | ≈0.000 | −0.0002 |

Overall, the comparison between nAG and gAG shows that gated agreement functions as a selective regularizer rather than a uniform constraint that forces all two-view predictions to be identical. Although gAG improved Top-1 accuracy in most backbone–dataset combinations, the changes in KL divergence and confidence were condition-dependent, indicating that its effect varies with the backbone architecture and dataset characteristics. The contrasting results of SLF-HMDB and TSF-HAA further support this interpretation: gAG improved both accuracy and two-view consistency in SLF-HMDB, whereas it reduced checkpoint-level variation and KL divergence in TSF-HAA without improving Top-1 accuracy. These findings suggest that gated agreement is most effective when reliable and discriminative temporal evidence is available, while its benefit can be limited for ambiguous or phase-dependent views.

#### 4.3.2. Motion Magnitude and Two-View Performance Analysis

This subsection analyzes how the performance difference between the baseline and two-view framework models changes according to the magnitude of video motion. The optical flow [35] magnitude between consecutive frames was computed for each video, and the average value was defined as the motion score. Optical flow was only used as an analysis metric to quantify the degree of dynamic change in the videos. All test samples were divided into low-, medium-, and high-motion groups based on the motion score, and the Top-1 accuracy of the baseline and best-performing two-view framework models for each backbone–dataset combination was compared in each group.

The line plots in Figure 5 show the Top-1 accuracy of the baseline and two-view framework models in each motion group, whereas the bar plots represent the absolute performance gain of the two-view framework model over the baseline (gain, %p). Thus, the line plots indicate the overall recognition performance level in each motion group, whereas

the bar plots show the additional improvement provided by the two-view framework over the baseline at that performance level.

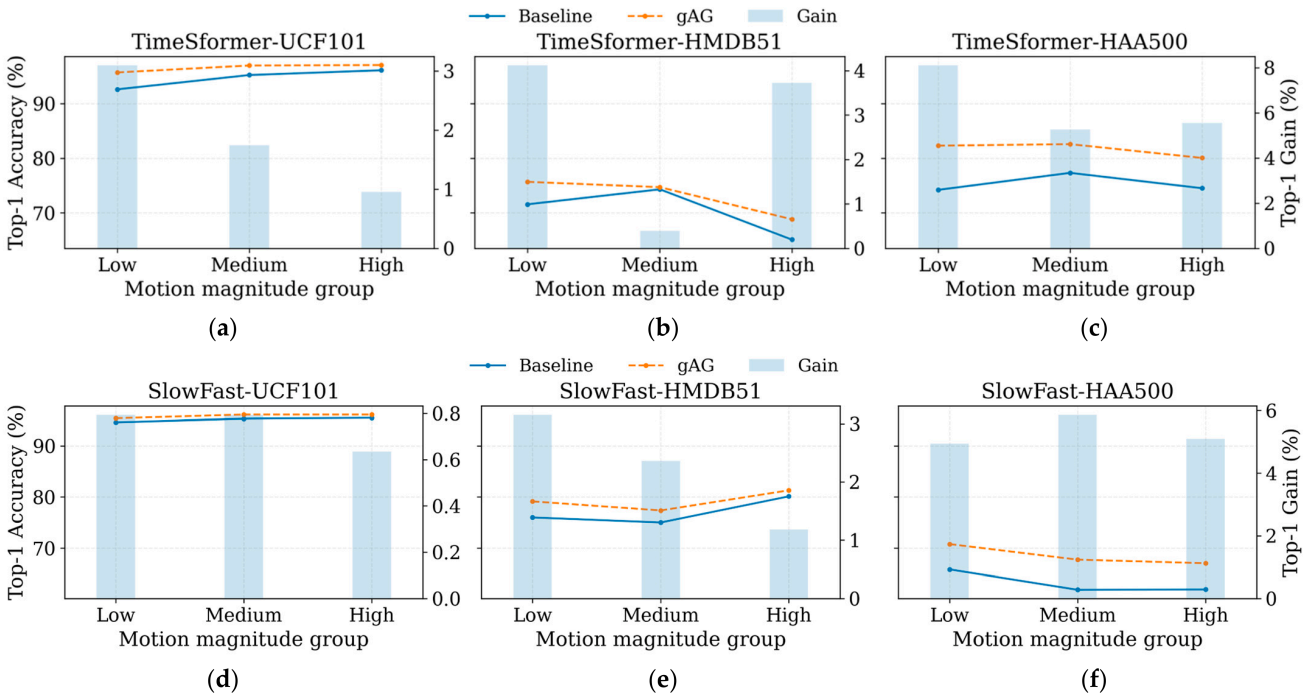


**Figure 5.** Performance comparison across motion magnitude groups: (**a**) TSF–UCF; (**b**) TSF–HMDB; (**c**) TSF-HAA; (**d**) SLF–UCF; (**e**) SLF–HMDB; (**f**) SLF-HAA. Lines indicate Top-1 accuracy, and bars indicate absolute gain over the baseline.

This experiment was conducted to examine whether the effect of two-view learning varies with the motion magnitude of videos. Overall, gAG achieved higher Top-1 accuracy than the baseline in most motion groups, indicating that two-view learning contributes to performance improvement across different motion levels. However, the magnitude of the gain showed dataset-dependent patterns. In TimeSformer–UCF101, both models maintained high accuracy across all motion groups, while the gain was largest in the low-motion group and gradually decreased toward the high-motion group. In contrast, TimeSformer–HMDB51 and TimeSformer–HAA500 showed more pronounced gains in the low- and high-motion groups than in the medium-motion group. These results suggest that two-view learning is generally effective across motion levels, but its benefit depends on the dataset-specific motion characteristics and the distribution of temporal cues.

#### 4.3.3. Sample-Level Analysis

This section analyzes whether the prediction consistency between the two temporal views is related to sample-level classification correctness. For each test sample, the symmetric KL divergence between the prediction distributions of the two views was computed. The samples were then divided into “Correct” and “Wrong” groups according to the Top-1 prediction result, and the distributions were compared across datasets, backbones, and training strategies.

Figure 6 shows the KL divergence distributions for UCF101, HMDB51, and HAA500. In all datasets, correctly classified samples are concentrated in relatively low-divergence regions, whereas wrongly classified samples show higher divergence and larger variation. This trend indicates that when the two temporal views produce similar prediction distributions, the final prediction is more likely to be correct. In contrast, large view-wise prediction

discrepancies are frequently associated with ambiguous or insufficient temporal evidence, leading to unstable predictions.

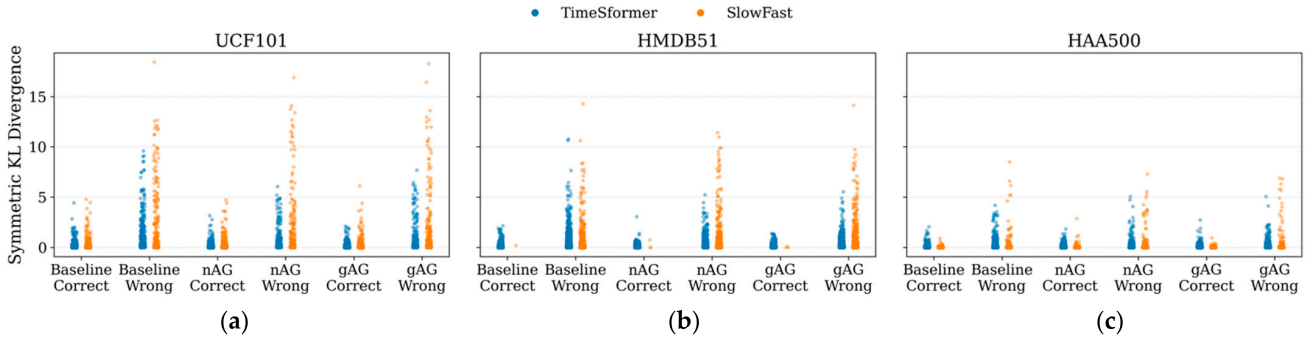


**Figure 6.** Distribution of the KL divergence according to prediction correctness. Comparison baseline and gAG: (**a**) UCF101; (**b**) HMDB51; (**c**) HAA500.

The tendency is particularly clear in the "Wrong" groups, where large outliers appear more frequently than in the "Correct" groups. This suggests that two-view prediction consistency can be interpreted as an indicator of temporal prediction stability. Overall, these results support the motivation of the proposed framework, suggesting that improving agreement between reliable temporal views can enhance the stability and accuracy of action recognition.

#### 4.3.4. Group-Wise Analysis of Training Metrics Across Performance Levels

In this subsection, the epochs of each backbone–dataset combination are sorted by the test Top-1 accuracy and divided into bottom-5, middle-5, and top-5 groups. The average cross-entropy loss (CE), mean gate weight ($\lambda$), and gAG loss were then compared across the groups.

Figure 7 shows the changes in these training metrics according to the performance level. Overall, CE tended to decrease as the performance improved across the datasets, indicating that higher-performing epochs were generally associated with more stable classification optimization. This trend was particularly pronounced in HAA500, where the SlowFast CE sharply decreased from the bottom-5 to the middle-5 and top-5 groups.

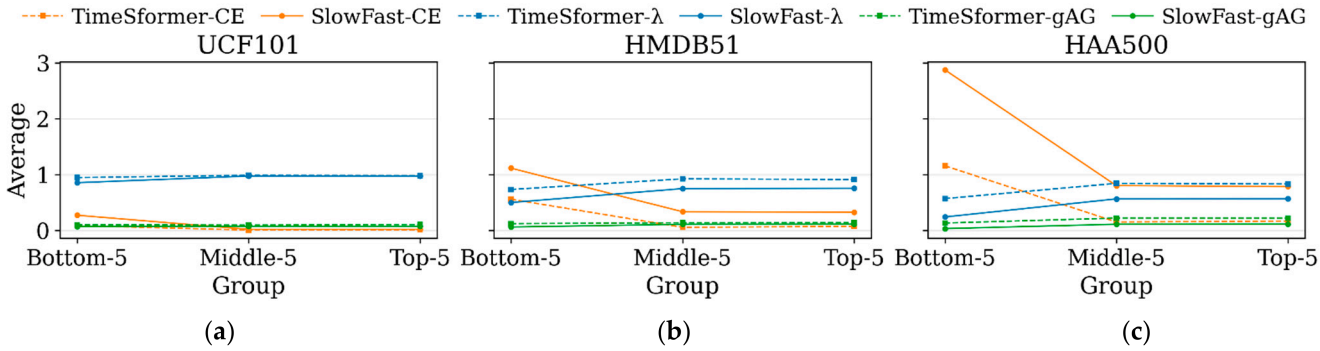


**Figure 7.** Group-wise comparison of cross-entropy loss, mean gate weight, and gated agreement loss across performance levels: (**a**) UCF101; (**b**) HMDB51; (**c**) HAA500.

CE reflects the supervised discriminative learning of each temporal view, whereas the gate weight $\lambda$ determines how strongly the agreement term is activated for each sample. The gAG loss is therefore not a pure error term like CE, but a gated consistency signal whose magnitude depends on both prediction discrepancy and sample-wise reliability. In contrast to CE, the mean gate weight and gAG loss did not show a strictly monotonic pattern in

all cases. This suggests that gAG loss should not be interpreted as better simply when lower. Rather, the gated agreement term acts as an auxiliary consistency constraint whose effect depends on the reliability and temporal compatibility of the two views. Therefore, Figure 7 indicates that higher-performing groups were mainly characterized by lower classification loss, while agreement regularization was selectively reflected during training.

#### 4.3.5. Qualitative Success and Failure Case Analysis

A qualitative analysis was conducted on the TSF-UCF results to examine when two-view learning improves or degrades recognition performance. Samples with different predictions between the baseline and gAG were selected and categorized into success and failure cases. For each case, the input frames from the two temporal views, predicted classes, confidence (Conf.), and entropy (Ent.) are presented.

Figure 8a,b show success cases in which the proposed method corrects the misclassifications of the baseline. In the BlowDryHair and TableTennisShot cases, the baseline incorrectly predicts Haircut and JumpRope, respectively, whereas gAG correctly predicts the ground-truth classes by exploiting action cues commonly shared by the two views. Both cases show high Conf. and low Ent., suggesting that agreement learning can serve as a reliable regularization signal when the two temporal views share consistent class-discriminative evidence.


(**a**) (**b**) (**c**) (**d**)

**Figure 8.** Qualitative success and failure cases of the proposed two-view learning method on TSF-UCF. Confidence (Conf.) and entropy (Ent.) values are reported for each prediction. (**a**) Success: GT—BlowDryHair, Baseline—Haircut, Proposed—BlowDryHair; (**b**) Success: GT—TableTennisShot, Baseline—JumpRope, Proposed—TableTennisShot; (**c**) Failure: GT—Nunchucks, Baseline—Nunchucks, Proposed—PizzaTossing; and (**d**) Failure: GT—CricketBowling, Baseline—CricketBowling, Proposed—Shotput.

In contrast, Figure 8c,d show failure cases. In the Nunchucks and CricketBowling cases, the action cues are not stably shared between the two views due to dark backgrounds, ambiguous object motion, or the emphasis on only a specific motion phase. As a result, gAG misclassifies them as PizzaTossing and Shotput, respectively, with low Conf. and high Ent. This suggests that when shared cues between views are insufficient or temporal phases are inconsistent, agreement learning may reinforce uncertain predictions.

Therefore, gAG can improve recognition performance by enhancing prediction consistency for reliable two-view pairs, but its effect may be limited for ambiguous view pairs. These results support the need for confidence- and entropy-aware gating to mitigate unreliable agreement.

#### 4.3.6. Summary of Experimental Findings

The overall experimental results show that the proposed two-view framework generally improves Top-1 accuracy over the single-view baseline. In particular, the largest gains

are observed on HAA500, suggesting that two-view supervision can help complement temporal evidence in fine-grained atomic action recognition.

However, the comparison between nAG and gAG shows that the effect of agreement regularization is condition-dependent. gAG improves the average Top-1 accuracy over nAG, but the improvement is not uniform across all backbone–dataset combinations. The KL-divergence analysis further shows that reduced prediction discrepancy does not always directly lead to higher accuracy, suggesting that agreement is most useful when the two temporal views share reliable and class-discriminative evidence.

The qualitative success and failure cases support this interpretation. Reliable view pairs with consistent action cues tend to benefit from agreement, whereas ambiguous view pairs or temporally mismatched phases can limit its effect. These findings suggest that the proposed gated agreement should be viewed as a selective regularization mechanism rather than a uniform constraint applied equally to all two-view pairs.

## 5. Discussion

The results indicate that two-view learning can improve action recognition performance by exploiting temporal evidence from different segments of the same video. Nevertheless, the effect of agreement regularization was not uniform across all experimental conditions. This suggests that prediction-level agreement should be applied with consideration of the reliability of each view pair, rather than being treated as a universally beneficial constraint.

The current framework also has several limitations. First, the experiments were conducted in a supervised trimmed-video setting. Therefore, further validation is required to assess its applicability to untrimmed videos, temporal localization, or more complex real-world scenarios. Second, the framework uses only two temporal views because simultaneous multi-view processing substantially increases GPU memory usage. Although sequential processing with gradient accumulation enabled CE supervision for both views, extending the framework to more than two views remains an important future direction. Third, entropy was used as a simple proxy for prediction uncertainty. More advanced confidence estimation methods or learnable gating functions may further improve the robustness of agreement regularization.

These limitations provide meaningful directions for future research. Extending the proposed framework to multi-view temporal coverage, untrimmed video understanding, and adaptive gating mechanisms may further improve the generality and reliability of two-view prediction agreement learning.

## 6. Conclusions

This study proposes a two-view framework that directly learns prediction-level temporal alignment to address prediction discrepancies between different temporal views of the same video in video-based action recognition. The proposed method combines supervised classification of the anchor view with symmetric agreement regularization between the two views, and mitigates the performance degradation caused by excessive alignment by adjusting the alignment strength for each sample through entropy-based conditional gating.

Experiments on two representative benchmark datasets and structurally different backbones showed that the proposed method consistently improved performance over the baseline across various settings. These results indicate that the improvement is related to prediction-level consistency learning between temporal views, rather than simply multi-view input expansion.

Consequently, this study demonstrates that combining prediction-level alignment with adaptive gating is an effective training strategy for handling temporal variations in video action recognition. It also suggests future directions for multi-view extensions, untrimmed video settings, and more sophisticated confidence modeling.

**Author Contributions:** Conceptualization, Y.-J.P.; methodology, Y.-J.P.; software, Y.-J.P.; validation, Y.-J.P.; formal analysis, Y.-J.P.; investigation, Y.-J.P.; resources, Y.-J.P. and H.-S.C.; data curation, Y.-J.P.; writing—original draft preparation, Y.-J.P.; writing—review and editing, Y.-J.P. and H.-S.C.; visualization, Y.-J.P.; supervision, Y.-J.P.; project administration, Y.-J.P. and H.-S.C.; funding acquisition, H.-S.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was supported by the DGIST R&D Program of the Ministry of Science and ICT of Korea (26-IT-03).

**Institutional Review Board Statement:** Not applicable. This study used publicly available benchmark datasets and did not involve new data collection from human participants.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets analyzed in this study are publicly available benchmarks: UCF101 [12], HMDB51 [13] and HAA500 [14]. No new raw data were collected. The data and source code supporting the findings of this study are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| HAR | Human Action Recognition |
| CNN | Convolutional Neural Network |
| AG | Prediction-Level Agreement |
| gAG | Gated Agreement |
| nAG | Without Gated Agreement |
| CE | Cross-Entropy |
| TSF-UCF | TimeSformer-UCF101 |
| TSF-HMDB | TimeSformer-HMDB51 |
| TSF-HAA | TimeSformer-HAA500 |
| SLF-UCF | SlowFast-UCF101 |
| SLF-HMDB | SlowFast-HMDB51 |
| SLF-HAA | SlowFast-HAA500 |
| KL-Div | Kullback–Leibler Divergence |
| SGD | Stochastic Gradient Descent |
| FPS | Frames Per Second |
| STD | Standard Deviation |

## References

1. Kaseris, M.; Kostavelis, I.; Malassiotis, S. A comprehensive survey on deep learning methods in human activity recognition. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 842–876. [CrossRef]
2. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 6299–6308. [CrossRef]
3. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'14), Montreal, QC, Canada, 8–13 December 2014*; MIT Press: Cambridge, MA, USA, 2014; Volume 1, pp. 568–576.

4. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), Santiago, Chile, 7–13 December 2015*; IEEE: New York, NY, USA, 2015; pp. 4489–4497. [CrossRef]
5. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019*; IEEE: New York, NY, USA, 2019; pp. 6202–6211.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
7. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Virtual, 18–24 July 2021*; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2021; Volume 139, pp. 813–824.
8. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020), Virtual, 13–18 July 2020*; Proceedings of Machine Learning Research: Cambridge, MA, USA, 2020; Volume 119, pp. 1597–1607.
9. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Virtual, 14–19 June 2020; pp. 9729–9738.
10. Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S.; Cui, Y. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Virtual, 19–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 6964–6974.
11. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; Volume 9912, pp. 20–36. [CrossRef]
12. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Action Classes from Videos in the Wild. CRCV-TR. 2012. Available online: https://www.crcv.ucf.edu/data/UCF101.php (accessed on 21 April 2026).
13. Kuehne, H.; Jhuang, H.; Garrote, E.; Serre, T.; Poggio, T. HMDB: A large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011*; IEEE: New York, NY, USA, 2011; pp. 2556–2563. Available online: https://serre.lab.brown.edu/hmdb51.html (accessed on 21 April 2026).
14. Chung, J.; Wuu, C.; Yang, H.; Tai, Y.; Tang, C. HAA500: Human-Centric Atomic Action Dataset with Curated Videos. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021*; IEEE: New York, NY, USA, 2021; pp. 13465–13474.
15. Qian, Y.; Sun, Y.; Kargarandehkordi, A.; Azizian, P.; Mutlu, O.C.; Surabhi, S.; Chen, P.; Jabbar, Z.; Wall, D.P.; Washington, P. Advancing Human Action Recognition with Foundation Models Trained on Unlabeled Public Videos. *arXiv* **2024**, arXiv:2402.08875.
16. Kalfaoglu, M.E.; Kalkan, S.; Alatan, A.A. Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition. In *Computer Vision—ECCV 2020 Workshops*; Springer: Cham, Switzerland, 2020; pp. 731–747. [CrossRef]
17. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. In *Proceedings of the 5th International Conference on Learning Representations*; ICLR: Toulon, France, 2017.
18. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NIPS 2017); Long Beach, CA, USA, 4–9 December 2017*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
19. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E.D.; Kurakin, A.; Li, C.-L. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NIPS 2020), Virtual, 4–9 December 2020*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33.
20. Liang, X.; Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T.-Y. R-drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems (NIPS 2021)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34.
21. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; CVPR: Salt Lake City, UT, USA, 2018; pp. 4320–4328. [CrossRef]
22. Li, T.-K.; Chan, K.-L.; Tjahjadi, T. Variable temporal length training for action recognition CNNs. *Sensors* **2024**, *24*, 3403. [CrossRef] [PubMed]
23. Lin, W.; Mirza, M.J.; Kozinski, M.; Possegger, H.; Kuehne, H.; Bischof, H. Video test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; CVPR: Vancouver, BC, Canada, 2023; pp. 22952–22961. [CrossRef]

24. Xu, Y.; Yang, J.; Cao, H.; Wu, K.; Wu, M.; Chen, Z. Source-Free Video Domain Adaptation by Learning Temporal Consistency for Action Recognition. In *Computer Vision—ECCV 2022. ECCV 2022*; Lecture Notes in Computer Science; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; Volume 13694. [CrossRef]
25. Nguyen, T.T.; Kawanishi, Y.; John, V.; Komamizu, T.; Ide, I. View-aware cross-modal distillation for multi-view action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 6–10 March 2026.
26. Feng, W.; Zhu, Z.; Liu, W.; Wang, X.; Liu, B.; Yu, X.; Zhong, X. From temporal thumbnail to semantics: Debiasing multi-view action recognition. *Pattern Recogn.* **2026**, *175*, 113055. [CrossRef]
27. Liu, W.; Zhong, X.; Zhou, Z.; Jiang, K.; Wang, Z.; Lin, C.-W. Dual-recommendation disentanglement network for view fuzz in action recognition. *IEEE Trans. Image Process.* **2023**, *32*, 2719–2733. [CrossRef] [PubMed]
28. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
29. PyTorch Documentation. KLDivLoss. Available online: https://pytorch.org/docs/stable/generated/torch.nn.KLDivLoss.html (accessed on 21 April 2026).
30. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
31. Fan, H.; Murrell, T.; Wang, H.; Alwala, K.V.; Li, Y.; Li, Y.; Xiong, B.; Ravi, N.; Li, M.; Yang, H.; et al. PyTorchVideo: A deep learning library for video understanding. In *Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021*; ACM: New York, NY, USA, 2021; pp. 3783–3786. [CrossRef]
32. Facebook Research. TimeSformer. Available online: https://github.com/facebookresearch/TimeSformer (accessed on 21 April 2026).
33. Facebook Research. SlowFast. Available online: https://github.com/facebookresearch/SlowFast (accessed on 21 April 2026).
34. SlowFast, GitHub Repository. Available online: https://github.com/leftthomas/SlowFast (accessed on 21 April 2026).
35. Horn, B.K.P.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203. [CrossRef]