A. Proof of Theorem 3.1

Lemma A.1 (Stability of Dominant Subspace; see Theorem 1 in [52]). *Let $X, X' \in \mathbb{R}^{T\times D}$ with $\Delta := X' - X$, and $\delta_k := \sigma_k(X) - \sigma_{k+1}(X) > 0$. If $\|\Delta\|_2 < \delta_k$, then:*

$$\|\sin\Theta(\mathcal{U}_k, \mathcal{U}'_k)\| \leq \frac{\|\Delta\|_2}{\delta_k - \|\Delta\|_2} \tag{6}$$

where $\mathcal{U}_k, \mathcal{U}'_k$ are top-k left singular subspaces of X and X'.

Proof. **Step 1: Perturbation bounds.** In this step, we derive an upper bound for the spectral norm of the perturbation matrix Δ. We aim to confirm that the perturbation is small enough to satisfy the conditions of the subspace stability (Lemma A.1). Let $\Delta_c = (1-\alpha_c)(y_c - x_c)$ for $c \in \mathcal{C}_{\text{low}}$, else 0. By the triangle inequality,

$$\|\Delta_c\|_2 \leq |1-\alpha_c|(\|y_c\|_2 + \|x_c\|_2) \tag{7}$$

From the condition $|1-\alpha_c| \leq 1$, and *Assumption 1* of Theorem 3.1 ($\|y_c\|_2 \leq \beta$):

$$\|\Delta_c\|_2 \leq |1-\alpha_c|(\|y_c\|_2 + \|x_c\|_2) \leq \beta + \sqrt{I_c} \leq \beta + \sqrt{\eta} \tag{8}$$

Since the Frobenius norm of a matrix Δ is the square root of the sum of the squared L2 norms of its column vectors (Δ_c), and the perturbation Δ is only non-zero for columns in the set $\mathcal{C}_{\text{low}}$:

$$\|\Delta\|_F^2 = \sum_{c\in\mathcal{C}_{\text{low}}} \|\Delta_c\|_2^2 \tag{9}$$

By Equation (8) and Equation (9):

$$\|\Delta\|_F^2 \leq \sum_{c\in\mathcal{C}_{\text{low}}} (\beta + \sqrt{\eta})^2 \tag{10}$$

Thus, with *Assumption 3*:

$$\|\Delta\|_F \leq \sqrt{|\mathcal{C}_{\text{low}}|}(\beta + \sqrt{\eta}) < \delta_k/2 \tag{11}$$

Since $\|\Delta\|_2 \leq \|\Delta\|_F$, we have:

$$\|\Delta\|_2 < \delta_k/2 \tag{12}$$

Step 2: Subspace stability. In this step, we use the perturbation bound from Step 1 to show that the dominant singular subspace of the data remains stable. This confirms that our fusion process does not corrupt the principal components of the original data. By Equation (12) and Lemma A.1:

$$\|\sin\Theta(\mathcal{U}_k, \mathcal{U}'_k)\| \leq \frac{\|\Delta\|_2}{\delta_k - \|\Delta\|_2} < \frac{\delta_k/2}{\delta_k - \delta_k/2} = 1 \tag{13}$$

Step 3: Bounding $\langle X, \Delta\rangle$. In this step, we aim to show that the inner product $\langle X, \Delta\rangle$ is minimal when η is sufficiently small and the injected signals y_c are sufficiently novel to the dominant subspace of X. First, from the definition of Δ:

$$\langle X, \Delta\rangle = \sum_{c\in\mathcal{C}_{\text{low}}} \langle x_c, \Delta_c\rangle = \sum_{c\in\mathcal{C}_{\text{low}}} (1-\alpha_c)\langle x_c, y_c - x_c\rangle \tag{14}$$

The absolute value of this expression can be bounded using the triangle inequality and the condition $\alpha_c \in [0,1]$ (which implies $|1-\alpha_c| \leq 1$):

$$|\langle X, \Delta\rangle| \leq \sum_{c\in\mathcal{C}_{\text{low}}} |\langle x_c, y_c - x_c\rangle| \leq \sum_{c\in\mathcal{C}_{\text{low}}} \left(|\langle x_c, y_c\rangle| + \|x_c\|_2^2\right) \tag{15}$$

Here, we know that $\|x_c\|_2^2 = I_c \leq \eta$.

Now, we aim to bound $|\langle x_c, y_c\rangle|$. First, let the Singular Value Decomposition (SVD) of X be $X = \sum_{i=1}^r \sigma_i u_i v_i^\top$. Its c-th column vector x_c can be expressed as:

$$x_c = \sum_{i=1}^r \sigma_i v_{i,c} u_i \tag{16}$$

Here, $v_{i,c}$ is the c-th component of the i-th right singular vector v_i. Then, we can substitute the decomposition (16) into the inner product:

$$\langle x_c, y_c\rangle = \left\langle \sum_{i=1}^r \sigma_i v_{i,c} u_i, y_c \right\rangle = \sum_{i=1}^r \sigma_i v_{i,c} \langle u_i, y_c\rangle \tag{17}$$

We now decompose y_c into its projections onto the dominant and residual subspaces:

$$y_c = \underbrace{\sum_{i=1}^k \langle y_c, u_i\rangle u_i}_{y_c^{\text{dominant}}} + \underbrace{\sum_{i=k+1}^r \langle y_c, u_i\rangle u_i}_{y_c^{\text{residual}}} \tag{18}$$

Substituting into Equation (17):

$$\langle x_c, y_c\rangle = \left\langle \sum_{i=1}^r \sigma_i v_{i,c} u_i, y_c^{\text{dominant}} + y_c^{\text{residual}} \right\rangle \tag{19}$$

$$= \sum_{i=1}^r \sigma_i v_{i,c} \langle u_i, y_c^{\text{dominant}}\rangle + \sum_{i=1}^r \sigma_i v_{i,c} \langle u_i, y_c^{\text{residual}}\rangle \tag{20}$$

Taking absolute values:

$$|\langle x_c, y_c\rangle| = \left| \sum_{i=1}^r \sigma_i v_{i,c} \langle u_i, y_c^{\text{dominant}}\rangle + \sum_{i=1}^r \sigma_i v_{i,c} \langle u_i, y_c^{\text{residual}}\rangle \right| \tag{21}$$

Applying the triangle inequality:

$$|\langle x_c, y_c\rangle| \leq \underbrace{\left| \sum_{i=1}^r \sigma_i v_{i,c} \langle u_i, y_c^{\text{dominant}}\rangle \right|}_{T_1} + \underbrace{\left| \sum_{i=1}^r \sigma_i v_{i,c} \langle u_i, y_c^{\text{residual}}\rangle \right|}_{T_2} \tag{22}$$

Since we define y_c^{dominant} as $y_c^{\text{dominant}} = \sum_{j=1}^{k} \langle y_c, u_j \rangle u_j$ in Equation (18):

$$T_1 = \left| \sum_{i=1}^{r} \sigma_i v_{i,c} \left\langle u_i, \sum_{j=1}^{k} \langle y_c, u_j \rangle u_j \right\rangle \right| \tag{23}$$

$$= \left| \sum_{j=1}^{k} \langle y_c, u_j \rangle \sum_{i=1}^{r} \sigma_i v_{i,c} \langle u_i, u_j \rangle \right| \tag{24}$$

$$= \left| \sum_{j=1}^{k} \langle y_c, u_j \rangle (\sigma_j v_{j,c}) \right| \quad \text{(by orthonormality of } u_i\text{)} \tag{25}$$

Now we can bound T_1 by applying the Cauchy-Schwarz inequality:

$$T_1 \leq \left(\sum_{j=1}^{k} |\langle y_c, u_j \rangle|^2 \right)^{1/2} \left(\sum_{j=1}^{k} |\sigma_j v_{j,c}|^2 \right)^{1/2} \tag{26}$$

$$\leq \left(\sum_{j=1}^{k} |\langle y_c, u_j \rangle|^2 \right)^{1/2} \sqrt{I_c} \quad (\text{since } I_c = \sum_{j=1}^{r} \sigma_j^2 v_{j,c}^2) \tag{27}$$

Now let us bound T_2. Since we define y_c^{residual} in Equation (18) as $\langle u_i, y_c^{\text{residual}} \rangle = 0$ for $i \leq k$:

$$T_2 = \left| \sum_{i=1}^{r} \sigma_i v_{i,c} \langle u_i, y_c^{\text{residual}} \rangle \right| \tag{28}$$

$$= \left| \sum_{i=k+1}^{r} \sigma_i v_{i,c} \langle u_i, y_c^{\text{residual}} \rangle \right| \tag{29}$$

Applying the Cauchy-Schwarz inequality two times:

$$T_2 \leq \left(\sum_{i=k+1}^{r} |\sigma_i v_{i,c}|^2 \right)^{1/2} \left(\sum_{i=k+1}^{r} |\langle u_i, y_c^{\text{residual}} \rangle|^2 \right)^{1/2} \tag{30}$$

$$\leq \sqrt{I_c} \cdot \left(\sum_{i=k+1}^{r} |\langle u_i, y_c^{\text{residual}} \rangle|^2 \right)^{1/2} \tag{31}$$

Since $y_c^{\text{residual}} = \sum_{i=k+1}^{r} \langle y_c, u_i \rangle u_i$:

$$(31) = \sqrt{I_c} \cdot \|y_c^{\text{residual}}\|_2 \tag{32}$$

Now we combine both bounds from Equation (27) and Equation (32). Recalling from Equation (22):

$$|\langle x_c, y_c \rangle| \leq T_1 + T_2 \tag{33}$$

By applying *Assumption 4*, where $\sum_{i=1}^{k} |\langle y_c, u_i \rangle|^2 \leq \gamma \sum_{i=k+1}^{r} |\langle y_c, u_i \rangle|^2 = \gamma \|y_c^{\text{residual}}\|_2^2$ to both Equation (27) and Equation (32):

$$T_1 + T_2 \leq \sqrt{\gamma} \|y_c^{\text{residual}}\|_2 \sqrt{I_c} + \sqrt{I_c} \|y_c^{\text{residual}}\|_2 \tag{34}$$

$$= \sqrt{I_c} \|y_c^{\text{residual}}\|_2 (\sqrt{\gamma} + 1) \tag{35}$$

Since $\|y_c^{\text{residual}}\|_2 \leq \|y_c\|_2 \leq \beta$ (by *Assumption 1*) and $I_c \leq \eta$ for $c \in \mathcal{C}_{\text{low}}$, by combining Equation (33) and Equation (35), we finally get:

$$|\langle x_c, y_c \rangle| \leq \beta \sqrt{\eta} (\sqrt{\gamma} + 1) \tag{36}$$

Now, we substitute this bound and $\|x_c\|_2^2 = I_c \leq \eta$ into Equation (15):

$$|\langle X, \Delta \rangle| \leq \sum_{c \in \mathcal{C}_{\text{low}}} \left(|\langle x_c, y_c \rangle| + \|x_c\|_2^2 \right) \tag{37}$$

$$\leq \sum_{c \in \mathcal{C}_{\text{low}}} (\beta \sqrt{\eta} (\sqrt{\gamma} + 1) + \eta) \tag{38}$$

$$= |\mathcal{C}_{\text{low}}| \left(\beta (\sqrt{\gamma} + 1) \sqrt{\eta} + \eta \right) \tag{39}$$

This final bound shows that the inner product between X and Δ becomes very small when η is sufficiently small and γ is bounded. The bounded alignment assumption (small γ, *Assumption 4*) ensures that the injected signals y_c contribute primarily to the residual subspace rather than interfering with the dominant patterns of X.

Step 4: Spectral flattening. In this step, we combine the previous results to analyze the change in the representation's overall magnitude, measured by the squared Frobenius norm $\|X'\|_F^2$. The squared Frobenius norm is defined as the inner product of a matrix with itself as below.

$$\|X'\|_F^2 = \langle X', X' \rangle \tag{40}$$

Since $X' = X + \Delta$:

$$\|X'\|_F^2 = \langle X + \Delta, X + \Delta \rangle \tag{41}$$

Using the distributive property (bilinearity) of the inner product:

$$\langle X + \Delta, X + \Delta \rangle = \langle X, X \rangle + \langle X, \Delta \rangle + \langle \Delta, X \rangle + \langle \Delta, \Delta \rangle \tag{42}$$

Since $\langle X, \Delta \rangle = \langle \Delta, X \rangle$ for the Frobenius inner product, the expression simplifies to:

$$\|X'\|_F^2 = \|X\|_F^2 + 2\langle X, \Delta \rangle + \|\Delta\|_F^2 \tag{43}$$

Rearranging the equation,

$$\|X'\|_F^2 - \|X\|_F^2 = 2\langle X, \Delta \rangle + \|\Delta\|_F^2 \tag{44}$$

Combining the result of **Step 3** (Equation (39)) and *Assumption 2* ($\|\Delta\|_F^2 \geq \epsilon$):

$$\|X'\|_F^2 - \|X\|_F^2 = 2\langle X, \Delta\rangle + \|\Delta\|_F^2 \tag{45}$$

$$\geq -2|\mathcal{C}_{\text{low}}|\,(\beta(\sqrt{\gamma}+1)\sqrt{\eta}+\eta) + \epsilon \tag{46}$$

$$> \epsilon/2 > 0 \tag{47}$$

for sufficiently small η and γ. This positive change is injected primarily into the tail subspace, as ensured by the bounded alignment assumption, leading to a flatter singular value spectrum.

Step 5: Effective rank increase. From **Step 2**, the dominant subspace remains stable. Also, from **Step 4**, the squared Frobenius norm of the representation increases, i.e., $\|X'\|_F^2 > \|X\|_F^2 + \epsilon/2$. The bounded alignment condition (*Assumption 4*) ensures that the modification Δ is poorly aligned with the dominant subspace of X. This, combined with the fact that the update is restricted to low-informativeness channels (which are inherently less aligned with the dominant subspace), suggests that the increase in squared Frobenius norm is primarily concentrated in the residual singular values.

Therefore, we have:

$$\sum_{j\leq k}\sigma_j'^2 \approx \sum_{j\leq k}\sigma_j^2 \quad \text{(dominant spectrum preserved)} \tag{48}$$

$$\sum_{j>k}\sigma_j'^2 > \sum_{j>k}\sigma_j^2 + \epsilon/2 \quad \text{(residual spectrum amplified)} \tag{49}$$

Let p and p' be the normalized singular value spectra of X and X'. This redistribution of squared singular values, where the total sum increases while the dominant part remains roughly constant, leads to a *flattening* of the normalized singular value distribution p' compared to p. Specifically, the increase in total mass ($\|X'\|_F^2 > \|X\|_F^2$) is concentrated in the tail, which reduces the probability mass of the dominant components. Since the Shannon entropy $H(p)$ is Schur-concave, it increases when the probability mass is transferred from larger components to smaller ones [51]. Therefore:

$$H(p') > H(p) \implies \text{ERank}(X') > \text{ERank}(X) \tag{50}$$

□

B. Analysis

B.1. Additional Analysis on Spectrum Flattening

In Figure 2, we show that the extent of improvement varies across datasets. For instance, in UTKinect (Figure 2 (b)), Depth shows more pronounced spectral expansion than RGB, whereas in DARai (Figure 2 (a),(c)) and NTURGBD (Figure 2 (d)), RGB shows more expansion than Depth. To explain this, Table 6 quantifies the complementarity between modalities via principal angles (Assumption 4 in Theorem 3.1). In UTKinect, the angle from RGB to Depth (65.57°) is larger than the reverse (63.80°), suggesting that RGB contributes more novel information to Depth, consistent with the stronger spectral expansion observed in Depth shown in Figure 2 (b). In contrast, the smaller and more symmetric angles in DARai and NTURGBD correspond to more prominent gains in RGB. These findings validate our theoretical insight that the degree of complementarity between modalities affects how much each can benefit from fusion.

Dataset	$\mathbf{X} \leftarrow \mathbf{Y}$ (Fusion Direction)	**Angle between subspaces** (°)
DARai (Coarse)	RGB ← Depth	48.74
	Depth ← RGB	47.73
UTKinect	RGB ← Depth	63.80
	Depth ← RGB	65.57
DARai (Fine-grained)	RGB ← Depth	48.68
	Depth ← RGB	47.20
NTURGBD	RGB ← Depth	34.25
	Depth ← RGB	31.95

Table 6. This shows an complementarity (in degrees) between the principal subspaces $\mathbf{X}$ and the injected channels $\mathbf{Y}_{\mathbf{C}_{\text{low}}}$. Higher values (closer to 90°) indicate stronger complementarity, which contributes more effectively to rank increase.

C. Architecture

C.1. Temporal Fuser Architecture

MHSA. To understand how different frames relate to each other and provide contextual meaning to each scene, we employ a MHSA within each temporal fusion layer. Unlike traditional sequential models (e.g., RNNs), self-attention enables the model to dynamically attend to the most relevant frames, ensuring that each time step is interpreted in the correct context. Given an input of the stacked multi-modal feature sequence $\mathbf{F}_{\text{stacked}} = \begin{bmatrix}\mathbf{F}^{\text{RGB}} \\ \mathbf{F}^{\text{Depth}}\end{bmatrix} \in \mathbb{R}^{B\times T\times 2C}$, where B is the batch size, T is the video sequence length, and $2C$ represents the concatenated feature dimensions from both modalities, we define the MHSA operation as:

$$\text{MHSA}(\mathbf{F}^{(l)}) = \text{Concat}(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_h)\mathbf{W}_o, \tag{51}$$

where $\mathbf{F}^{(l)}$ is the input sequence representation in l^{th} layer, Concat$(\cdot)$ denotes a concatenation function, h is the number of attention heads and $\mathbf{W}_o \in \mathbb{R}^{(h\cdot d_h)\times d}$ is the output projection matrix.

Each attention head computes:

$$\mathbf{A}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_h}}\right)\mathbf{V}_i, \tag{52}$$

where $\mathbf{Q}_i = \mathbf{F}^{(l)}\mathbf{W}_i^Q$, $\mathbf{K}_i = \mathbf{F}^{(l)}\mathbf{W}_i^K$, $\mathbf{V}_i = \mathbf{F}^{(l)}\mathbf{W}_i^V$, are the query, key, and value projection matrices, respectively, $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d \times d_h}$, and $d_h = \frac{d}{h}$ is the feature dimension per head.

To ensure stable training, we apply Layer Normalization $\text{LN}(\cdot)$ and residual connections:

$$Z_t = \text{LN}(F_{\text{stacked}}) + \text{MHSA}(\text{LN}(F_{\text{stacked}})). \tag{53}$$

MLP. While MHSA models dependencies across time, it does not directly enhance the expressiveness of individual frame features. To refine the representation of each frame, we apply an $\text{MLP}(\cdot)$ to introduce non-linearity and feature transformation. The MLP module operates as:

$$\mathbf{F}^{(l+1)} = \text{LN}(\mathbf{Z}^{(l)} + \text{MLP}(\mathbf{Z}^{(l)})), \tag{54}$$

where $\text{LN}(\cdot)$ and $\text{MLP}(\cdot)$ are applied independently at each time step.

After L layers of temporal fusion, we obtain the final output representation:

$$\mathbf{F}_{\text{final}} = \text{LN}(\mathbf{F}^{(L)}) \in \mathbb{R}^{B \times T \times d}. \tag{55}$$

Segmentation head. The ultimate goal of the Temporal Fuser is to determine what each frame represents by assigning an action label to every time step. To accomplish this, we employ a segmentation head, which consists of a fully connected layer $\text{FC}(\cdot)$:

$$\hat{\mathbf{Y}} = \text{FC}(\mathbf{F}_{\text{final}}), \tag{56}$$

where $\hat{y}_t$ the predicted action class at time step t.

C.2. Action anticipation Module Architecture

Future queries as learnable representations. Instead of relying on a single deterministic output for action anticipation, we introduce learnable queries—denoted as future queries ($\mathbf{Q}_{\text{future}}$) to predict multiple action possibilities [22]. These queries are randomly initialized and learned during training. Given $\mathbf{N_q}$ future queries, we define: $\mathbf{Q}_{\text{future}} \in \mathbb{R}^{N_q \times C}$ where C is the hidden dimension. These queries are independent of any specific frame and instead act as a learnable representation that extracts relevant information from the past observations.

Multi-head cross-attention (MHCA). To effectively anticipate future actions, the model must attend to relevant moments in the past. To achieve this, we employ MHCA, where the future queries ($\mathbf{Q}_{\text{future}}$) interact with the Temporal Fuser's output ($\mathbf{F}_{\text{temporal}}$).

$$\mathbf{H}_{\text{attn}}^{(h)} = \text{Attention}(\mathbf{Q}_{\text{future}}\mathbf{W}_Q^{(h)}, \mathbf{F}_{\text{temporal}}\mathbf{W}_K^{(h)}, \mathbf{F}_{\text{temporal}}\mathbf{W}_V^{(h)}), \tag{57}$$

where h denotes the head index, and $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)}$ are learnable projection matrices for queries, keys, and values, respectively.

The outputs from all attention heads are concatenated and projected through a learnable weight matrix $\mathbf{W}_O$:

$$\mathbf{H}_{\text{MHCA}} = \text{Concat}(\mathbf{H}_{\text{attn}}^{(1)}, \mathbf{H}_{\text{attn}}^{(2)}, \ldots, \mathbf{H}_{\text{attn}}^{(H)})\mathbf{W}_O, \tag{58}$$

where H denotes the number of attention head and $\mathbf{W}_O$ ensures that the output dimensionality remains consistent.

This formulation allows the model to refine future action representations by selectively attending to relevant past temporal features.

Feed-forward network (FFN). The output of $\text{MHCA}(\cdot)$ is further processed through a $\text{FFN}(\cdot)$ to capture non-linear dependencies in the learned future queries: $\text{FFN}(\mathbf{X}) = \sigma(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$, where $\mathbf{W}_1 \in \mathbb{R}^{d \times d_f}$, $\mathbf{W}_2 \in \mathbb{R}^{d_f \times d}$, and $\sigma(\cdot)$ is a non-linear activation.

Action anticipation head. The refined representation from the $\text{FFN}(\cdot)$ is used to anticipate the future action label using a $\text{FC}(\cdot)$ layer:

$$\hat{y}_{\text{future}} = \text{FC}(\text{FFN}(\mathbf{X})), \tag{59}$$

where $\hat{y}_{\text{future}}$ denotes the predicted future action class.

D. Experimental Details

D.1. Datasets

We aim to evaluate R3D's ability to anticipate human actions in diverse and realistic scenarios, particularly in settings where spatial reasoning and multi-modal fusion play a critical role. To this end, we utilize three action anticipation datasets: NTURGBD [59], UTKinect-Action3D [65], and DARai [32]. The UTKinect-Action3D dataset was collected by the University of Texas at Austin and features 10 types of human actions, each performed twice by 10 subjects from multiple viewpoints. The dataset poses additional challenges due to actor-dependent variations, occlusions caused by human-object interactions, and body parts moving out of the field of view. The NTURGBD dataset is one of the largest and most diverse action recognition datasets, designed for 3D human action analysis. It contains 56,880 video samples covering 60 action classes performed by 40 subjects from 80 viewpoints. The DARai dataset offers highly realistic scenarios that closely resemble real-world human behavior. Unlike UTKinect and NTURGBD, the videos in DARai are untrimmed, showcasing raw, continuous human activity in real-world contexts without artificial segmentation. It comprises 150 action classes, 50 participants, two distinct exocentric views, and three levels of hierarchical labels.

Method	Time / frame (ms)	FLOPs (per frame)
GTAN [72]	5.920	49 GFLOPs
Ours	0.119	0.58 GFLOPs
without RTF	0.020	0.12GFLOPs
128 → 64 channels	0.090	0.24 GFLOPs

Table 7. Comparison of computational cost and scalability on an NVIDIA A40 GPU.

D.2. Experimental setups

We use pretrained ResNet features as input visual features for the NTURGBD, UTKinects, and DARai. To align with the temporal resolution of each dataset, the sampling rates are set to 15 for DARai, and 1 for UTKinects and NTURGBD. We use the number of Future Queries $\mathbf{N}_q$ fixed at 8. Based on the density of each dataset, the hidden dimension size D is set to 128 for both NTURGBD, UTKinects, and DARai. During training, the observation rate α is set to $\alpha \in \{0.2, 0.3, 0.5\}$, while the prediction rate β is fixed at 0.5. Yet when training UTKinects, to further augment the training dataset, we use the observation rate $\alpha \in \{0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55\}$. The model is trained for 60 epochs using the AdamW optimizer [49] with a learning rate of $1e-3$ and a batch size of 8. A cosine annealing warm-up scheduler [50] is applied, with warm-up stages spanning 10 epochs. For evaluation setup, we follow the long-term action anticipation framework protocol [1, 2, 34, 58]. We set the observation rate $\alpha \in \{0.2, 0.3\}$ and prediction rate $\beta \in \{0.1, 0.2, 0.3, 0.5\}$, and measure mean over classes (MoC) accuracy for evaluation metrics. To ensure consistency, we report average performance across 3 number of iteration, each with fixed seeds $1, 10, 13452$.

D.3. Computational cost and Scalabilty

Computational cost. We evaluate the computational cost and scalability of R3D on an NVIDIA A40 GPU. For comparison, we chose GTAN [72], the current state-of-the-art model reported in Table 1. As shown in Table 7, R3D processes a single RGB-Depth frame in 0.119 ms, requiring 61.77MB of memory and 0.58 GFLOPs. In contrast, GTAN requires 5.92ms per RGB frame and approximately 49 GFLOPs. This gap arises because GTAN is a diffusion-based model whose computational cost scales linearly with the number of sampling steps, while R3D avoids such overhead.

Scalability. As shown in Table 7 without RTF, inference on a single RGB-Depth frame costs 0.02ms per frame. This is because RTF involves SVD operations, which is the main contributor to runtime. To mitigate this overhead, we experiment with lowering the channel dimension from 128 to 64 before the RTF stage on the UTKinects dataset. This adjustment reduced the per-frame cost from 0.119ms to 0.09ms and from 0.58 GFLOPs to 0.24 GFLOPs, while still achieving state-of-the-art performance. In fact, the results reported in Table 1 for UTKinects are based on this scaled configuration, showing that hyperparameter tuning enables us to maintain accuracy while significantly improving computational efficiency.

Dataset	Methods	$\beta(\alpha=0.2)$				$\beta(\alpha=0.3)$			
		0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5
UTKinects	W/O	31.31	23.02	22.11	17.18	13.40	17.31	15.78	11.34
	W/	**38.96**	**37.68**	**37.16**	**35.95**	**34.80**	**32.51**	**36.03**	**26.98**
NTURGBD	W/O	16.77	16.59	16.59	16.62	20.64	20.73	20.76	20.85
	W/	**21.98**	**20.18**	**19.72**	**18.90**	**23.17**	**21.27**	**20.51**	**19.89**
DARai	W/O	20.92	20.43	19.37	18.87	37.88	31.04	30.51	29.20
	W/	**33.44**	**32.14**	**30.56**	**29.59**	**46.29**	**42.05**	**43.41**	**40.25**

Table 8. Ablation Study across three datasets evaluating the impact of Temporal Fuser. "W" denotes the use of Temporal Fuser, while "W/O" indicates its absence.

Dataset	Methods	$\beta(\alpha=0.2)$				$\beta(\alpha=0.3)$			
		0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5
UTKinects	Shared	34.07	**40.43**	35.06	27.60	27.11	26.82	34.62	25.30
	Separated	**38.96**	37.68	**37.16**	**35.95**	**34.80**	**32.51**	**36.03**	**26.98**
NTURGBD	Shared	10.39	10.21	10.23	10.23	9.09	9.28	9.26	9.32
	Separated	**21.98**	**20.18**	**19.72**	**18.90**	**23.17**	**21.27**	**20.51**	**19.89**
DARai (Coarse)	Shared	27.83	28.16	26.25	24.75	38.90	31.18	32.16	30.66
	Separated	**33.44**	**32.14**	**30.56**	**29.59**	**46.29**	**42.05**	**43.41**	**40.25**
DARai (Fine)	Shared	29.60	25.44	23.70	16.13	**27.46**	26.12	22.17	17.65
	Separated	**32.57**	**25.92**	**24.02**	**16.68**	26.97	**28.43**	**24.81**	**18.02**

Table 9. Ablation study on the effect of projecting each modality into a shared latent space prior to the Token Fuser. "Shared" denotes the setting where modality-specific features are projected into a common latent space before fusion, while "Separated" refers to the setting where modalities are fused without this projection step.

E. More Ablation Studies

E.1. Impact of Temporal Fuser

We conduct an ablation study on the Temporal Fuser, which models temporal dependencies across RGB and depth features. As observed in Table 8, removing the Temporal Fuser leads to a noticeable performance drop, highlighting its role in leveraging temporal context for action anticipation.

E.2. Ablation on Latent Space Projection

To evaluate the impact of aligning modalities in a shared feature space prior to fusion, we conduct an ablation study comparing two settings: Shared, where modality-specific features are projected into a common latent space before entering the Rank-enhancing Token Fuser, and Separated, where fusion is performed without this projection step. As shown in Table 9, the Separated setting consistently outperforms the Shared one. We hypothesize that this performance gap arises because projecting both modalities into the same latent space reduces the structural differences between them. This highlights the importance of maintaining modality-specific representations before fusion in order to

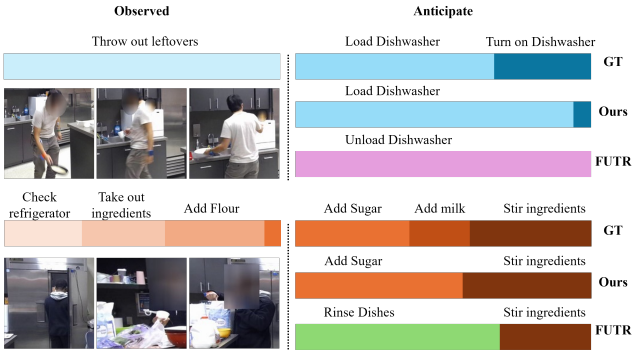

Figure 8. Qualitative comparison of action anticipation results between the Ground Truth (GT), the RGB-based baseline model (GTAN), and R3D.

fully leverage richer and more complementary information from both modalities.

F. Qualitative Analysis

To assess the ability to distinguish fine-grained human actions, we compare R3D against the strongest baseline from Table 1, using the most detailed label set—L3 (procedure) level from the DARai dataset. Figure 8 presents a qualitative comparison between the ground truth (GT), the RGB-based GTAN model, and R3D. In the first example of Figure 8, the task requires differentiating between loading and unloading a dish into a dishwasher. The RGB-only GTAN model struggles with this distinction as it lacks explicit spatial cues to determine whether the dish is moving toward or away from the dishwasher. This results in misclassification due to the absence of motion directionality cues. In contrast, by allowing the model to infer that the dish is moving toward the dishwasher, R3D correctly predicts 'loading' by leveraging depth-based directionality. In the second example of Figure 8, the person first adds flour, then quickly transitions to adding sugar. Due to the brevity of this transition, only a small portion of the add sugar action is observed before the model must anticipate the next step. R3D successfully distinguishes between adding sugar and adding flour, demonstrating its ability to capture subtle procedural differences. This is achieved as R3D effectively utilizes the strengths of RGB features for recognizing sugar and flour.

G. Discussion: why Depth?

To support our choice of depth as the complementary modality in Section 3.2, we conduct a qualitative analysis to better understand why depth is particularly effective for multimodal fusion.

G.1. Depth Mitigates Modality Collapse

We investigate whether using depth indeed mitigates modality collapse, as hypothesized in our modality selection strategy. In Table 1, the largest performance gap between our model and the SOTA baseline occurs on the DARai dataset, whereas the smallest gap is observed on the NTURGB+D dataset. We analyze both cases by plotting the eigenvalue spectra of the fused features. As shown in Figure 9, We observe that the existing method (AFFT) exhibits modality

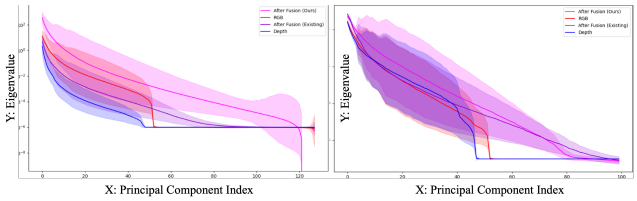

Figure 9. This figure compares the eigenvalue distribution of RGB (red), Depth (blue), existing fusion method (purple), and our proposed method (pink) in DARai (Left) and NTURGBD (Right) Dataset. Unlike the existing fusion method, where the eigenvalues fall below those of the RGB modality, our method exhibits a broader and higher eigenvalue spectrum than both RGB and Depth.

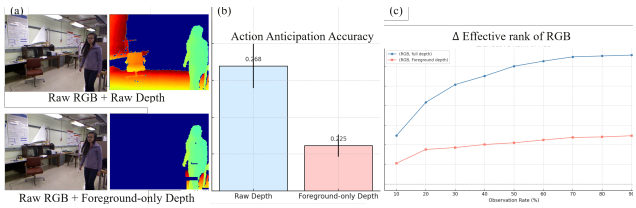

Figure 10. Analysis of the impact of background depth in RGB-depth fusion for action anticipation. (a) Two types of inputs from the UTKinect dataset: RGB images paired with either raw depth (top) or foreground-only depth (bottom), where foreground regions are segmented using joint annotations. (b) Average action anticipation performance (across 3 seeds) when fusing RGB with raw depth (blue) versus foreground-only depth (red). (c) Change in the effective rank of RGB representations as a function of observation rate, highlighting that raw depth significantly enhances the expressive capacity of RGB features compared to foreground-only depth.

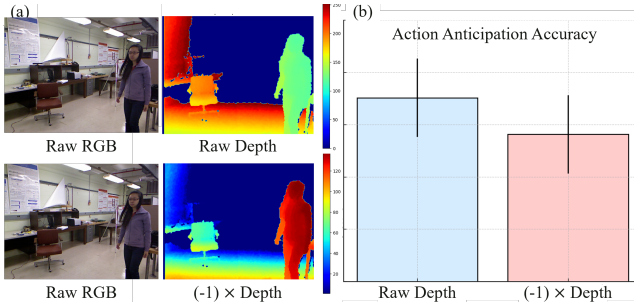

Figure 11. Analysis of the impact of background depth in RGB-depth fusion for action anticipation. (a) shows two types of inputs from the UTKinect dataset: RGB images paired with either raw depth (top) or foreground-only depth (bottom), where foreground regions are segmented using joint annotations. (b) Average action anticipation performance (across 3 seeds) when fusing RGB with raw depth (blue) versus foreground-only depth (red).

collapse: in the DARai dataset (left), its fusion spectrum lies below that of RGB, indicating a collapse of RGB features; in contrast, on the NTURGB+D dataset (right), the fusion spectrum falls below that of depth. In both cases, our method maintains a broader and higher-magnitude spectrum. Notably, this trend holds on the both datasets, even when the performance gap between models is relatively small, reinforcing the effectiveness of our fusion strategy in preserving modality diversity.

G.2. Depth Delivers Contextual Information

To demonstrate the importance of depth information, as shown in Figure 10, we conduct experiments on the UTKinect dataset, where most actions occur primarily in the foreground. Using the provided joint annotations, we segment the depth maps into foreground-only depth and raw depth as shown in Figure 10 (a). We then fuse each version with RGB and compare their performance on the action anticipation task. Figure 10 (b) shows that even in a foreground-centric dataset where foreground-only information might appear sufficient, raw depth consistently yields better performance. To better understand this phenomenon, we measure the change in Effective Rank of the RGB stream. As shown in Figure 10 (c), when raw depth which includes background is used, RGB representations exhibit a more substantial increase in expressive capacity. In contrast, with foreground-only depth, the increase in RGB representation is notably smaller. In conclusion, depth provides more than just motion cues from the foreground—it also delivers critical spatial and contextual information from the background. Thus, even background depth plays a vital role in structuring scene understanding, reinforcing depth as a key modality.

G.3. Directionality in Depth Features Matters

As shown in Figure 11, to test the importance of directionality in depth features, we negate the depth features (i.e., multiply by -1) before fusion as shown in Figure 11 (a). This operation preserves the magnitude of feature variation but flips their geometric interpretation. The performance drop in Figure 11 (b) suggests that directionality of depth information matters in multi-modal fusion. This observation motivates the need for modality-aware and direction-sensitive fusion.