Just Dance with π!
A Poly-modal Inductor for Weakly-supervised Video Anomaly Detection

Supplementary Material

A. Overview

The supplementary material is divided in three sections. In Section B we explain the experimental setup used to produce the results in the paper and describe the backbones utilized to extract the modalities features. Section C presents the experimental results obtained on the XD-Violence and MSAD datasets, similarly to Section 6 of the main paper. Finally, Section D discusses the impact of the poly-modal representations learned by the early and late inductors.

B. Experimental Setup

For XD-Violence, the most important metrics are Average Precision (AP), which measure the frame-level precision on normal and abnormal videos, and Abnormal AP (AP_A), which measures the precision on abnormal videos. Similarly, for UCF-Crime we measure the Area-Under-the-Curve (AUC) for normal and abnormal videos and the AUC of abnormal videos (AUC_A). For the MSAD dataset, there are no established protocols at this moment, therefore we report both AUC and AP (and their counterparts on abnormal videos). We train our own version of the UR-DMU baseline that is used as a teacher in PI-VAD and report its performance as well.

The experiments are conducted on an NVIDIA Titan RTX GPU with 24 GB of memory. The batch size is set to 16, equally split between normal and abnormal videos. The student model and the early/late induction modules are trained for 50 epochs using the training objective in Equation (5), and 75 epochs with Equation (6) of the main paper. λ_1 and λ_2 are empirically set to 0.2 and 0.1 respectively, while the learning rate of the AdamW [2] optimizer is set to $1e-5$.

Pose Embedding Extraction: We first extract the frame-level 2D body joints (J) of k humans from YoloV7-Pose [4] and stack them along the temporal dimension (T) to obtain $H_J \in \mathbb{R}^{T\times k\times J}$. Further, to embed the action-representative features to human joints, H_J is fed to a pre-trained action-recognition pose backbone UNIK [5] to compute a local human attributed feature embedding $J_P \in \mathbb{R}^{T\times k\times d_p}$ for every 16-frames snippet joint, where d_p is the feature dimension for a given set of J joints. Again, the real-world setting may have numerous sets of anomaly-irrelevant humans and noise, thus we apply *max-pool* across the k dimension of F_H to obtain a salient human feature map $e_P \in \mathbb{R}^{T\times d_p}$, where $d_p = 256$.

Vision-language Embedding Extraction: We extract d_{txt} dimensional snippet-level text augmented spatial features from VifiCLIP [3] Image encoder ViT/B-16 and text encoder with a **text codebook**[1] stacks them along the T dimension to obtain a text embedding $e_{txt} \in \mathbb{R}^{T\times d_{txt}}$, where $d_{txt} = 512$.

Depth Embedding Extraction: We extract d_D dimensional frame-level depth features from DepthAnythingV2 for the first frame of every 16-frame-snippet and stack them along the T dimension to obtain a depth embedding $e_D \in \mathbb{R}^{T\times d_D}$, where $d_D = 512$.

Panoptic Mask Embedding Extraction: We extract d_M dimensional frame-level panoptic mask features from SegmentAnything for the first frame of every 16-frame-snippet and stack them along the T dimension to obtain a depth embedding $e_M \in \mathbb{R}^{T\times d_M}$, where $d_M = 1024$.

C. Additional Performance Analysis

C.1. XD-Violence

Modality						XD-Violence	
Audio	Pose	Depth	Text	Pan.	Motion	AP	AP_A
-	-	-	-	-	-	81.66	83.52
✓	-	-	-	-	-	82.49	83.52
-	✓	-	-	-	-	82.28	83.78
-	-	✓	-	-	-	82.16	83.71
-	-	-	✓	-	-	**83.27**	83.66
-	-	-	-	✓	-	81.82	**84.25**
-	-	-	-	-	✓	82.32	83.76
✓	✓	-	-	-	-	82.38	83.23
✓	✓	✓	-	-	-	82.47	83.73
✓	✓	✓	✓	-	-	83.73	84.15
✓	✓	✓	✓	✓	-	85.35	85.41
✓	✓	✓	✓	✓	✓	**85.37**	**85.79**

Table 1. Modality impact comparisons on XD-Violence. The best results are written in **bold**.

[1]Text attributes of UCF-Crime dataset: **`Category`**= {`Abuse, Arrest, Assault, Burglary, Fighting, Robbery, Shooting, Shoplifting, Stealing, Vandalism, Explosion, Arson, RoadAccidents`}

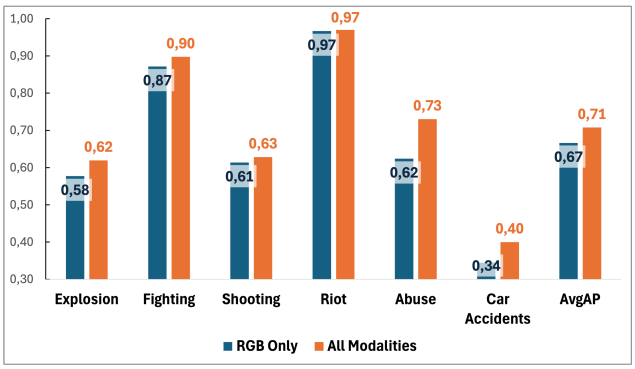

Figure 1. Comparison between the AP scores of different mixtures of modalities in the PI-VAD framework for the XD-Violence dataset.

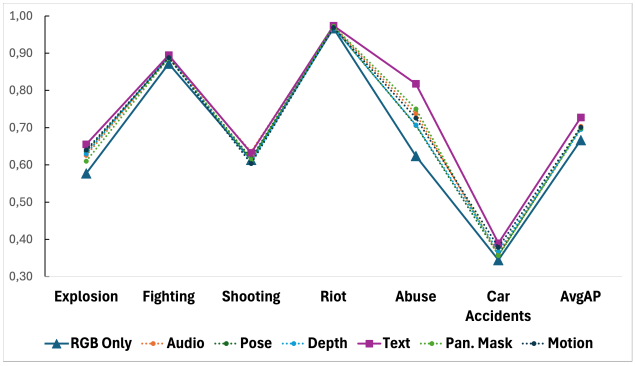

Figure 2. Comparison between the AP comparison between the baseline model with the addition of only one modality for the XD-Violence dataset.

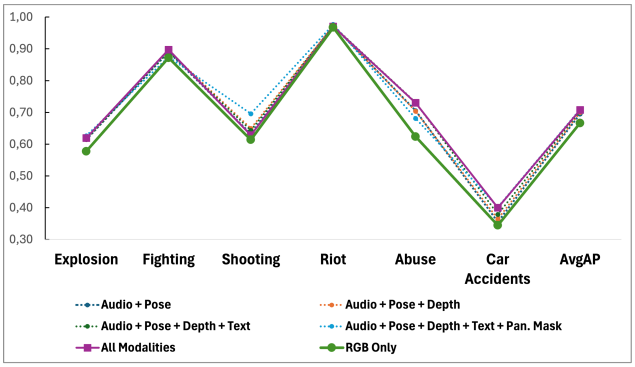

Figure 3. Comparison between the AP scores of different mixtures of modalities in the PI-VAD framework for the XD-Violence dataset.

Table 1 shows the AP and AP_A achieved by PI-VAD trained with individual modalities and mixtures of them. The text modality leads to much better performance on the AP, possibly due to the composition of the Kinetics-600 dataset used to train ViFiCLIP and the action recognition task on which it is trained. This performance increase does not fully translate to abnormal videos, where almost all other modalities outperform text. The opposite holds for the panoptic masks, which prove to be the most useful individual modality for this dataset. Given that XD-Violence is composed of movie clips, it is possible that this is due to the curated scene construction.

The performance increase for the class-wise AP metric on XD-Violence achieved by PI-VAD is shown in Figure 1. For most classes, the performance increments are moderate. The "Abuse" and "Car Accidents" classes benefit the most from the additional modalities by 11% and 6% respectively.

In Figure 2, we evaluate the audio features as a separate modality along with the others. Contrary to UCF-Crime, for this dataset, the motion modality does not exhibit large improvements on any of the anomalous classes. We conjecture that this effect is due to the fact that the videos in the XD-Violence dataset are clips of movies, containing fast camera movements and rapid changes in the viewpoint of the scene. This leads to suboptimal features for the modalities that would benefit more from the scene's temporal coherence, such as motion, pose or depth. Notably, the text modality is not affected by this issue and exhibits the largest performance increases per class on XD-Violence. The only class where a large increment in performance is observed is "Abuse", where all the modalities lead to sensibly better performance.

The analysis of the poly-modal mixtures in Figure 3 confirms the observation that no modality is more important than the others for this dataset.

C.2. MSAD

Modality					MSAD			
Pose	Depth	Text	Pan.	Motion	AUC	AUC_A	AP	AP_A
-	-	-	-	-	85.78	67.95	67.35	75.30
✓	-	-	-	-	87.21	69.21	68.26	76.75
-	✓	-	-	-	87.23	69.82	66.37	73.38
-	-	✓	-	-	87.56	69.09	**69.65**	**76.93**
-	-	-	✓	-	87.19	69.29	69.45	76.10
-	-	-	-	✓	**87.94**	**70.27**	69.25	76.84
✓	✓	-	-	-	87.94	69.04	68.80	75.42
✓	✓	✓	-	-	88.27	69.75	68.69	74.88
✓	✓	✓	✓	-	88.47	70.68	70.44	77.58
✓	✓	✓	✓	✓	**88.68**	**71.25**	**71.26**	**77.86**

Table 2. Modality impact comparisons on MSAD. The best results are written in **bold**.

For the MSAD dataset, there are no established evaluation protocols yet. Therefore, we report AUC, AUC_A, AP and AP_A in order to provide a baseline comparison for future research on the dataset. Table 2 reports the performance

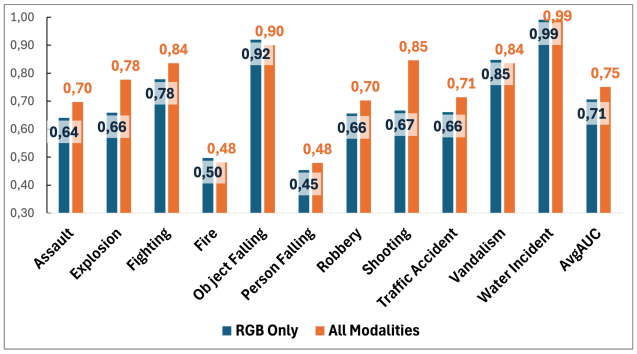

Figure 4. Comparison between the AUC scores of the baseline model and PI-VAD for the MSAD dataset.

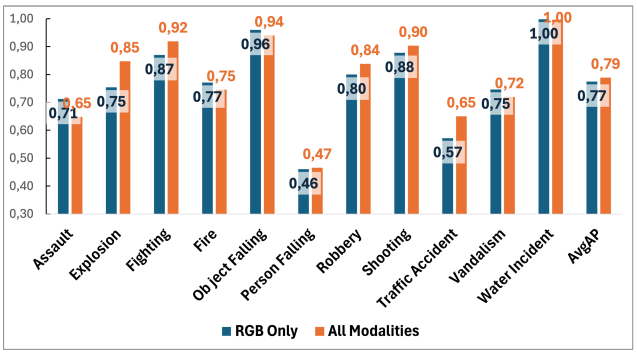

Figure 5. Comparison between the AP scores of the baseline model and PI-VAD for the MSAD dataset.

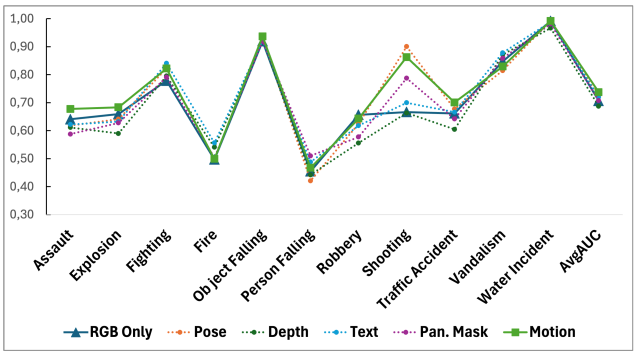

Figure 6. Classwise AUC comparison between the baseline model with the addition of only one modality for the MSAD dataset.

achieved by PI-VAD with the individual modalities and with their mixture. The proposed PI-VAD framework, trained on all five modalities, achieves a 2.9% improvement on AUC and 3.91% on AP. Figure 4 shows that PI-VAD obtains the largest performance gains wrt the baseline AUC score on the "Shooting" and "Explosion" classes. For what concerns the AP score, the largest improvements are observed in the "Explosion" and "Traffic Accident" classes. We also observe a drop in performance in the "Assault", "Fire" and

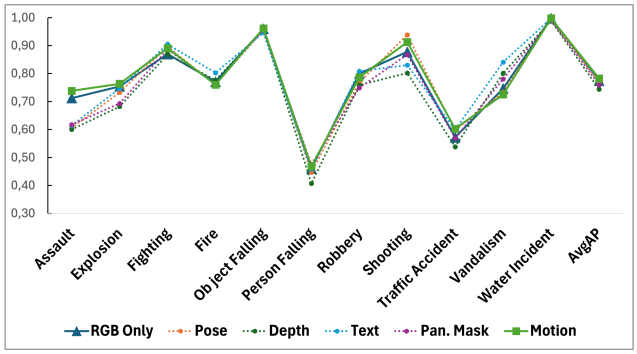

Figure 7. Classwise AP comparison between the baseline model with the addition of only one modality for the MSAD dataset.

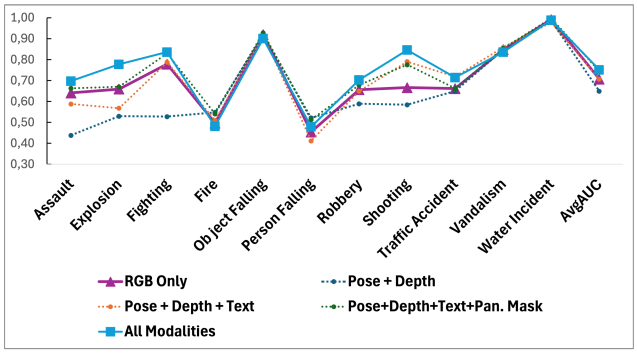

Figure 8. Comparison between the AUC scores of different mixtures of modalities in the PI-VAD framework for the MSAD dataset.

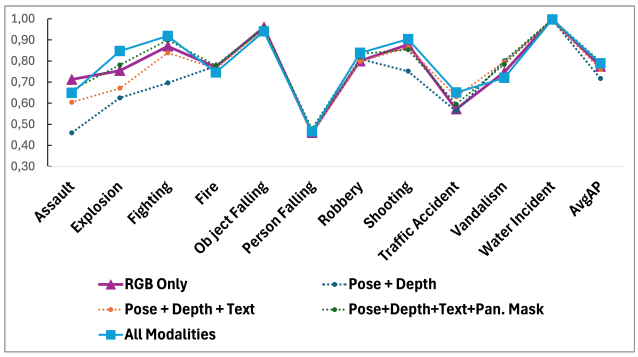

Figure 9. Comparison between the AP scores of different mixtures of modalities in the PI-VAD framework for the MSAD dataset.

"Vandalism" classes.

In this dataset, we observed that the text modality leads to the best AP performance, while motion leads to the best AUC. Both observations are coherent with the results on UCF-Crime, where the motion modality leads to the best AUC performance, and XD-Violence, on which the best AP performance is achieved by the text modality. These results suggest a correlation between the individual modal-

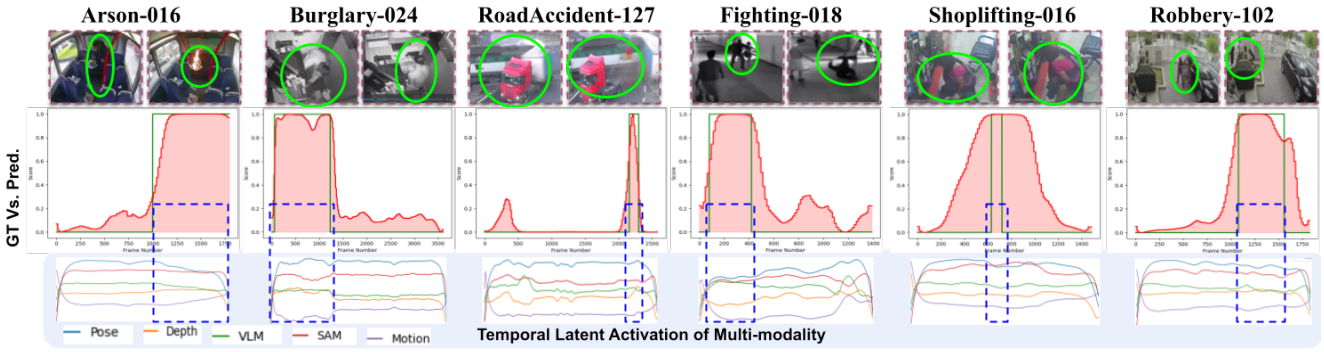

Figure 10. Qualitative results and modality activations from the **early** inductor.

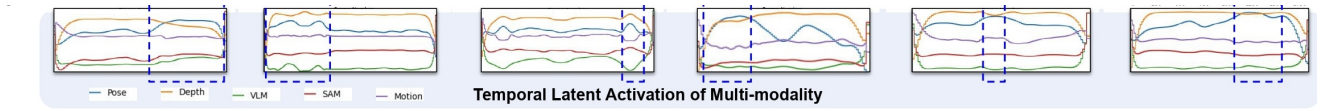

Figure 11. Modality activations from the **late** inductor

ities and the metric used for evaluation. Figure 6 illustrates the class-wise AUC performance of each individual modality. The most evident improvement is obtained by the pose, panoptic masks, and motion modalities in the "Shooting" class. For what concerns the AP, Figure 7 shows that the motion modality has a positive influence on the "Assault" class, where every other modality performs worse than the RGB-only baseline. Notably, we observe a positive influence of the pose modality on the "Shooting" class and of the text modality on the "Vandalism" class.

The class-wise evaluation of the poly-modal mixtures highlights the complementary effects of the modalities on the MSAD dataset. Figures 8 and 9 show that, for the "Assault", "Explosion", "Fighting" and "Robbery" classes, the modalities have a complementary effect on both the AUC and AP metric. In the "Fire" class, we can observe that the performance of PI-VAD decreases on both AUC and AP as the modalities are added to the mixture. This suggests that, in this class, the modalities have a contrastive effect, similar to what can be observed in the "Arson" class of UCF-Crime. The videos of the "Fire" and "Arson" classes of MSAD and UCF-Crime have similar visual features that can be more salient than the additional modalities.

D. Early/Late Inductors

In our framework, we integrated two inductors within the UR-DMU architecture. As shown in the main paper's ablation study, using both inductors together significantly boosts performance on the UCF-Crime dataset. Further analysis of modality activations reveals distinct poly-modal representations learned by each inductor. The late inductor shows peak activations corresponding to high abnormal scores, particularly emphasizing the depth modality, as discussed in Section 5.3 and illustrated in Figure 11.

In the early inductor, the pose modality generally has higher activations, though the depth modality aligns more clearly with high abnormal scores in certain cases. For the "Road Accident-127" example in Figure 10, the depth activation peak aligns with the abnormal score peak. In "Burglary-024," similar peaks appear in both text and depth modalities, indicating the early inductor's ability to construct a poly-modal, anomaly-aware representation. Additionally, examples such as "Fighting-018" and "Robbery-102" display overlapping activations across multiple modalities within the anomaly time region, suggesting that the early inductor has effectively learned to model inter-modality interactions. However, the modality activations of the early inductor do not exhibit such clear correlations in every example, as shown by "Arson-016" and "Shoplifting-016". This suggests that, in the early inductor, the anomaly-aware poly-modal representation is not as accurate as the one learned by the late inductor. In fact, the modality activations of the late inductor in Figure 11 display much sharper peaks in correspondence to the anomaly regions of these two examples. It is worth noting that the "Arson-016" and "Shoplifting-016" videos are two instances of subtle anomalies, which suggests that the more nuanced latent representation learned in later stages of the model is more effective in processing subtle poly-modal cues. This highlights the experimental results presented in Table **??**, further supporting the hypothesis that the two inductors have a complementary effect on PI-VAD.

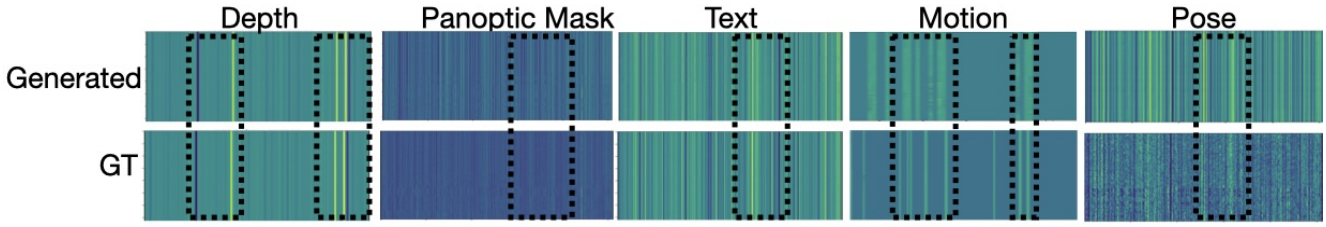

Figure 12. Visualization of the activation maps of ground truth vs. learned representations for all 5 modalities, highlighting key accurately reconstructed areas.

E. Additional Experiments and Discussion from Rebuttal

Analysis of Teacher Model Dependency: A weak teacher can suffice with the fully-supervised methods. In contrast, with *weakly-supervised* learning, where precise temporal labels are unavailable, a strong teacher is essential to generate high-quality, clean pseudo-labels. This approach has proven effective in WSVAD works such as ECU(CVPR'23), MIST (CVPR'21), and OE-CTST (WACV'24). Therefore, we have adopted this in our work, proposing a novel multi-modal method. **Impact of Teacher Model Selection on Performance:** Additional experiments in below Table reveal that even with a weaker teacher (column 4), PI-VAD has only a minor performance drop, while still surpassing several SoTA methods. Notably, a weaker teacher-student pairing (column 3) still yields significant gains. Thus, pre-training the teacher on the specific WS-VAD dataset followed by PI-VAD can achieve significant accuracy gain.

	MIL	UR-DMU	PI-VAD Combination T(MIL), S(MIL)	T(MIL), S(UR-DMU)	T(UR-DMU), S(UR-DMU)
AUC(%) on UCF-C	75.0	86.9	83.49	87.56	**90.33**

Experiments on robustness and standard deviation : The following Table shows the standard deviation (σ) for individual modalities, highlighting PI-VAD's improved robustness with lower standard deviation compared to [1]. These are calculated at every 5 epochs on the test set from epochs 75 to 125. PI-VAD training includes a 50-epoch warmup phase (Eq. (5) **MP**), followed by 75 epochs of Eq. (6) in **MP**.

Modality						UCF-Crime	XD-Violence	MSAD
RGB	Pose	Depth	Text	Pano.Mask	Motion	σ(AUC)	σ(AP)	σ(AUC)
✓	✓	-	-	-	-	0.44	0.43	0.51
✓	-	✓	-	-	-	0.38	0.39	1.35
✓	-	-	✓	-	-	0.43	0.31	0.48
✓	-	-	-	✓	-	2.06	0.51	1.29
✓	-	-	-	-	✓	0.53	1.87	1.05
✓	✓	✓	✓	✓	✓	1.85	2.45	1.47

Why comparing Uni- and multi-modality performances : We agree that multi-modal training often outperforms RGB-only methods, but experiments on the UCF-Crime (shown in the Table below) reveal a different trend. UR-DMU with multi-modality (late fusion) performs worse than RGB-only UR-DMU due to multi-modality challenges like noise, redundancy, and conflicting cues. These challenges arise because all modalities are estimated from RGB frames, without any specialized sensors. For fair SoTA comparison in Table 6 of **MP**, we focus on inference conditions, noting that "RGB-modality at inference" is lightweight and efficient, while "multi-modality at inference" incurs higher computational costs. PI-VAD, using only RGB at inference, achieves a strong trade-off between performance and computational efficiency, enabling a reasonably fair comparison with uni-modal SoTA.

	UR-DMU (RGB)	UR-DMU (RGB+MM)	PI-VAD (RGB)	PI-VAD (RGB+MM)
AUC(%) on UCF-C	86.9	83.6	90.33	90.58
Parameters	6.16	2329.85	82.81	2406.50
GFLOPs	1.54	2543.06	19.88	2561.40

References

[1] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5559, 2023.

[2] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

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

[4] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023.

[5] Di Yang, Yaohui Wang, Antitza Dantcheva, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Unik: A unified framework for real-world skeleton-based action recognition. *arXiv preprint arXiv:2107.08580*, 2021.