Supplementary Material for PromptonomyViT

In this supplementary file, we provide additional information about our experimental results, qualitative examples, implementation details and datasets. Specifically, Section A provides more experiment results, Section B provides qualitative visualizations to illustrate our approach, Section C provides additional implementation details, and Section D provides additional datasets details.

A. Additional Experiment Results

We begin by presenting additional baseline results for all datasets and tasks in (Section A.1). Next, we present additional ablations (Section A.2) we performed in order to test the contribution of the different PViT components.

A.1. Baselines Comparison

Here, we evaluate several alternative ViT approaches (*MViTv2 MT* and *MViTv2 VPT*) to our task of using synthetic data towards improving action recognition models. Additionally, we report additional baselines that are comparable in compute and size to further compare to other approaches in (see Table 5), such as ORViT Mformer [11], UniFormer-S [20], SViT [2], VideoMAE [31], Video SWIN Transformer [26], STIN [27], and SAFCAR [17]. We can observe that our PViT approach improves upon MViTv2 and is competitive with other strong models. We note that even compared to VideoMAE, a recent self-supervised learning method, our results are similar in AVA (+1.3) and SSv2 (-0.3), although VideoMAE utilizes a larger backbone and more computing for training. Finally, PViT can be applied to any pretrained backbone, which gives it an advantage over other methods.

A.2. Additional Ablations

Next, we provide additional ablations that further illustrates the benefits of our PViT.

The importance of synthetic scene data. To examine how important the information provided by the synthetic scene data is, we test the PViT model, but provide it with "useless" synthetic label information. Specifically, we run an experiment in which the synthetic scene annotations are shuffled. As a result, the ground truth of the instance-level is shuffled for each synthetic scene task (e.g., for dense prediction tasks, the GT maps are shuffled). This ablation obtained 63.4%, similar to the baseline (63.3%). This is expected since wrong scene annotations are not likely to provide additional benefit beyond the baseline. Moreover, the model is capable of ignoring prompts if they are not required, so they should not have a negative impact beyond the baseline.

Prompts for real-world data. Even though real-world datasets are less reach in annotations compared to synthetic, PViT can still use them if available. To examine this, we added 2D hand-object boxes from SomethingElse as an additional auxiliary task along with its own prompt. This improved results by +1.2, suggesting that real data, if available, is beneficial. Clearly, the combination of synthetic and real data offers many promising and interesting directions, and we leave those to future work.

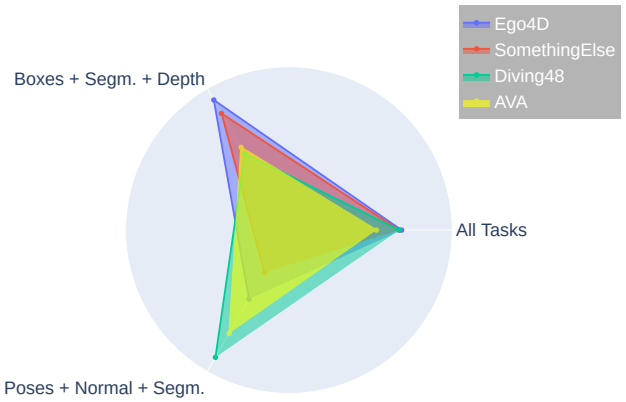

Figure 4. Dataset-Task Agreement. A polygon represents a real video dataset, and the closer a vertex is to the circle border, the greater the gain from using that synthetic task. The gains are scaled for comparison.

Comparison to a pretraining approach. Another approach for using synthetic datasets is first to pretrain on the synthetic data, and then finetune on the video-related task. Here, we demonstrate the effectiveness of our PViT approach as compared to this standard pretraining approach. To implement pretraining, we add prediction heads on top of MViTv2, and train them only on the synthetic datasets. Next, we remove these prediction heads and finetune the model by predicting using the CLS token. This approach achieved 61.9% compared to 63.3% for MViTv2 baseline and 65.5% for our PViT approach. This indicates that our PViT approach utilizes task information more effectively than a standard pretraining approach.

Number of task prompts. This ablation tests whether adding more prompts per task will improve the results compared to PViT, which uses one prompt per task. We add a total of 20 prompts to each task, which results in 65.4%, demonstrating that the addition of more prompts does not necessarily improve its performance. Clearly, there are many possible design choices, such as selecting a number of prompts per task, their dimension, integrating into different depths, etc., and we leave those to future work.

Dataset-Task Agreement. In Figure 4, we aim to explore how a different synthetic task combination helps real datasets. Since there are multiple possible subsets, we simplify and focus on only two subsets: S_1 ={Boxes, Segmentation, Depth} and S_2 ={Poses, Normal, Segmentation}. The former relates to hand-object interaction (HOI), and the latter to human action (HA). The figure shows the accuracy for real datasets when trained on either S_1, S_2, or all five tasks. This confirms our original

(a) **Something–Something V2**

Model	Pretrain	Top-1	Top-5
SlowFast [7], R101	K400	63.1	87.6
MViTv1 [6]	K400	64.7	89.2
ViViT-L [1]	IN+K400	65.4	89.8
UniFormer-S [20]	IN+K600	67.9	92.1
ORViT Mformer [11]	K400	67.9	90.5
VideoMAE (ViT-S)	K400	66.8	90.3
MViTv2 [22]	K400	68.2	91.4
MViTv2 MT	K400	68.4	91.3
MViTv2 VPT	K400	61.5	87.5
PViT (Ours)	K400	**69.6** (+1.2)	**91.6** (+0.2)

(b) **Diving48**

Model	Pretrain	Frames	Top-1
SlowFast [7], R101	K400	16	77.6
TimeSformer [3]	IN	16	74.9
TimeSformer-L [3]	IN	96	81.0
SViT [2]	K400	16	79.8
MViTv2 [22]	K400	16	73.1
MViTv2 MT	K400	16	75.6
MViTv2 VPT	K400	16	69.8
PViT (Ours)	K400	16	**85.8** (+6.0)

(c) **AVA-V2.2**

Model	Pretrain	mAP
SlowFast [7], R50	K400	22.7
SlowFast [7], R101	K400	23.8
ORViT MViT-B [11]	K400	26.6
VideoMAE (ViT-S) [31]	K400	22.5
VideoMAE (ViT-B) [31]	K400	26.7
MViTv1 [6]	K400	25.5
MViTv2 [22]	K400	26.8
MViTv2 MT	K400	27.2
MViTv2 VPT	K400	19.0
PViT (Ours)	K400	**28.4** (+1.6)

(d) **SomethingElse**

Model	Compositional		Base		Few-Shot	
	Top-1	Top-5	Top-1	Top-5	5-Shot	10-Shot
I3D [4]	42.8	71.3	73.6	92.2	21.8	26.7
SlowFast [7]	45.2	73.4	76.1	93.4	22.4	29.2
TimeSformer [3]	44.2	76.8	79.5	95.6	24.6	33.8
STIN [27]	48.2	72.6	-	-	-	-
TSM [23]	52.3	78.0	-	-	-	-
Mformer [28]	60.2	85.8	82.8	96.2	28.9	33.8
SAFCAR [17]	60.7	84.2	-	-	-	-
MViTv2 [22]	63.3	87.5	83.7	96.8	32.7	40.2
MViTv2 MT	62.7	87.6	81.4	96.2	34.0	40.9
MViTv2 VPT	53.0	81.8	76.8	94.8	31.8	39.0
PViT (Ours)	**65.5** (+2.2)	**89.0** (+2.5)	**85.0** (+1.3)	**97.4** (+0.6)	**34.3** (+1.6)	**41.3** (+1.1)

(e) **Ego4D**

Model	Temporal Localization Error	PNR Classification Top-1
Bi-LSTM	0.790	65.3
BMN [25]	0.780	-
I3D ResNet-50 [4]	0.739	68.7
EgoVLP (TimeSformer) [24]	0.666	73.9
Video Swin Transformer [26]	0.660	69.5
MViTv2 [22]	0.702	71.6
MViTv2 MT	0.640	73.6
MViTv2 OP	0.652	73.7
PViT (Ours)	0.637 (-0.065)	**74.8** (+3.2)

Table 5. **Results on SSv2, Diving48, AVA-V2.2, SomethingElse, and Ego4D datasets.** We report top-1 and top-5 accuracy on SSv2 and SomethingElse. On AVA, we report the mAP metric. On Diving48, we report top-1. On Ego4D we report classification error. IN refers to ImageNet-21K.

hypothesis from the main paper that the datasets are roughly clustered into two categories: (i) SomthingElse and Ego4D benefit more from the HOI set. These datasets indeed usually contain hands interacting with objects, often in first person and with a low field of view. (ii) AVA and Diving48 belong benefit more from the HA group. These datasets generally consist of zoomed-out frames with mostly full human bodies.

Contribution from Datasets and Tasks. In order to quantify the impact of each dataset and task, we conducted a comprehensive analysis in Table 6. At the top of the table we display the contribution of each synthetic dataset to the downstream task, and at the bottom we display the contribution of each synthetic task (namely, we use all existing annotations from across all of our synthetic datasets). We observe that EHOI achieves the highest gains. This is similar to our observation in the main paper that hand-object interaction videos (HOI) benefit more from bounding box supervision. For more details, see the Dataset-Task Agreement ablation (Figure 4d) in the main paper. In the bottom portion of the table, we examined in the auxiliary tasks contribute to performance individually, as well as the most effective combinations of auxiliary tasks. As can be seen, we find that performing PViT on auxiliary tasks individually does improve performance (see also *Dataset Task Agreement* below). However, using all tasks (last line) improves more than any individual task, and is also close to the optimal combination. This reinforces our strategy of simply training on all tasks. For a visualization of the datasets, see Section B in supplementary.

B. Qualitative Visualizations

Figure 5 and Figure 3 in the main paper show visualizations of "task prompts" predictions on examples of real videos from SSv2, Diving48, Ego4D, and AVA. It can be seen that predictions are reasonable, despite the model not being trained on these labels for the real videos. For better illustration, we show in Figure 6 the different auxiliary synthetic datasets we used in the main paper, as described in Section 4.1 and further elaborated upon in Section D.1.

C. Additional Implementation Details

Our PViT model can be used on top of the most common video transformers (MViT [6], TimeSformer [3], Mformer [28],

Dataset	Depth	Segm.	Normal	3D Poses	2D Boxes	Top-1	Top-5
-	✗	✗	✗	✗	✗	63.3	87.5
PHAV	✓	✓	✗	✗	✗	64.2	87.6
SUR	✓	✗	✗	✓	✗	63.9	88.4
ES	✗	✗	✗	✓	✗	63.9	88.1
HS	✓	✗	✓	✗	✗	64.1	87.4
EHOI	✗	✓	✗	✗	✓	65.0	88.5
PHAV+HS+SURR	✓	✗	✗	✗	✗	64.8	88.7
SUR+EHOI	✗	✓	✗	✗	✗	65.0	88.7
HS	✗	✗	✓	✗	✗	63.9	88.2
SUR+ES	✗	✗	✗	✓	✗	64.1	88.4
EHOI	✗	✗	✗	✗	✓	64.7	88.6
best combination	✓	✓	✗	✗	✓	**65.5**	**89.0**
All	✓	✓	✓	✓	✓	65.1	88.8

Table 6. **Compositional Action Recognition task on the SomethingElse dataset.** The contribution of every synthetic auxiliary dataset (top) and task (bottom).

Video Swin [26]). For our experiments, we choose the MViTv2 [22] model because it performs well empirically. These are all implemented based on the MViTv2 [22] library (available at https://github.com/facebookresearch/mvit), and we implement PViT based on this repository. Furthermore, we set the λ parameters (see Equation 9) for the $\mathcal{L}_{Depth}$, $\mathcal{L}_{Normal}$, $\mathcal{L}_{Segm}$, $\mathcal{L}_{3DPose}$, $\mathcal{L}_{Boxes}$, and $\mathcal{L}_{DT}$ losses, to 0.5, 0.5, 0.1, 3.0, 0.1 and 1 respectively (across all datasets). We choose these lambda components such that all loss components have the same scale. We elaborate next on the additional implementation details for each dataset, including information about optimization, and training and inference.

Dense Prediction Heads. In order to preserve the spatio-temporal information in dense prediction tasks, we use patch tokens in addition to task the tokens. First, we upsample patch tokens from layers 2, 12, 15 (out of 16) using a 3D convolution layer, followed by Dropout and concatenation. We then concatenate them with relevant task tokens and forward them to an MLP for a final prediction.

C.1. Diving48

Dataset. Diving48 [21] contains 16K training and 3K testing videos spanning 48 fine-grained diving categories of diving activities. For all of these datasets, we use standard classification accuracy as our main performance metric.

Optimization details. We train using 16 frames with sample rate 4 and batch-size 128 (comprising 64 videos and 64 auxiliary synthetic datasets) on 8 RTX 3090 GPUs. We train our network for 10 epochs with Adam optimizer [18] with a momentum of $9e{-}1$ and Gamma $1e{-}1$. Following [22], we use $lr{=}1.5e{-}4$ with half-period cosine decay.

Training details. We use crops of size 224 for the standard model and jitter scales between $256-320$. together with RandomFlip augmentation. Finally, we sample T frames from the start and end annotation times, following [33].

Inference details. We take 3 spatial crops per single clip to form predictions over a single video in testing, as in [3].

C.2. SomethingElse

Dataset. The SomethingElse dataset [27] contains 174 action categories with 54,919 training and 57,876 validation samples. The compositional [27] split in this dataset provides disjoint combinations of a verb (action) and noun (object) in the training and testing set, defining two disjoint groups of nouns $\{\mathcal{A},\mathcal{B}\}$ and verbs $\{1,2\}$. Given the splits of groups, they combine the training set as $1\mathcal{A}{+}2\mathcal{B}$, while the validation set is constructed by flipping the combination into $1\mathcal{B}{+}2\mathcal{A}$. In this way, different combinations of verbs and nouns are divided into training or testing splits.

Few Shot Compositional Action Recognition. As mentioned in Section 4.4, we also evaluate on the few-shot compositional action recognition task in [27]. For this setting, we use 88 *base* action categories and 86 *novel* action categories. We train on the base categories (113K/12K for training/validation) and fine-tune on few-shot samples from the novel categories (for 5-shot, 430/50K for training/validation; for 10-shot, 860/44K for training/validation). We use the same training recapie as in C.2.

Optimization details. We train using 16 frames with sample rate 4 and batch-size 128 (comprising 64 videos and 64 auxiliary synthetic datasets) on 8 RTX 3090 GPUs. We train our network for 100 epochs with Adam optimizer [18] with a momentum of $9e{-}1$ and Gamma $1e{-}1$. Following [22], we use $lr{=}7e{-}5$ with half-period cosine decay.

Regularization details. We use weight decay of $1e{-}4$, and a dropout [12] of $5e{-}1$ before the final perdition.

Training details. We use standard crop size of 224, and we jitter scales from 256 to 320.

Inference details. We take 3 spatial crops per single clip to form predictions over a single video in testing.

C.3. Something-Something v2

Dataset. The SSv2 [27] is a $\sim$160K-video dataset contains 174 action categories of common human-object interactions. We follow the official splits from [8].

Optimization details. For the standard SSv2 [27] dataset, we train using 16 frames with sample rate 4 and batch-size 128 (comprising 64 videos and 64 auxiliary synthetic datasets) on 8 RTX 3090 GPUs. We train our network for 100 epochs with Adam optimizer [18] with a momentum of $9e{-}1$ and Gamma $1e-1$. Following [22], we use $lr = 7e-5$ with half-period cosine decay.

Regularization details. We use weight decay of $1e{-}4$, and a dropout [12] of $5e{-}1$ before the final classification.

Training details. We use a standard crop size of 224, and we jitter the scales from 256 to 320 along with RandomFlip.

Inference details. We take 3 spatial crops per single clip to form predictions over a single video in testing as in [22].

C.4. Ego4D

Dataset. Ego4D [9] is a new large-scale dataset of more than 3,670 hours of video data, capturing the daily-life scenarios of more than 900 unique individuals from nine different countries around the world. The videos contain audio, 3D meshes of the environment, eye gaze, stereo and/or synchronized videos from multiple egocentric cameras.
Metrics. In the Object State Change Temporal Localization task, the absolute error (in seconds) is used for evaluation. In the Object State Change Classification task, the top-1 accuracy is used for evaluation, following [9] protocol.
Optimization details. We train using 16 frames with sample rate 4 and batch-size 128 (comprising 64 videos and 64 auxiliary synthetic datasets) on 8 RTX 3090 GPUs. We train our network for 10 epochs with Adam optimizer [18] with a momentum of $9e{-}1$ and Gamma $1e{-}1$. Following [22], we use $lr = 1.5e{-}5$ with half-period cosine decay. Additionally, we used Automatic Mixed Precision, which is implemented by PyTorch.
Training details. We use a standard crop size of 224, and we jitter the scales from 256 to 320.
Inference details. We follow the official evaluation, both for the state change temporal localization and the state change classification tasks, available at `https://github.com/EGO4D/hands-and-objects`.

C.5. AVA-2.2

Dataset. AVA-2.2 (Atomic Visual Action) dataset [10] contains bounding box annotations for spatio-temporal localization of human actions. There are 211K training videos and 57K validation videos in the dataset. We report mean Average Precision (mAP) on 60 classes [10] on AVA v2.2 according to the standard evaluation protocol.
Architecture. SlowFast [7] and MViTv2 [22] use a detection architecture with a RoI Align head on top of the spatio-temporal features. We follow their implementation to allow a direct comparison, elaborating on the RoI Align head proposed in SlowFast [7]. First, we extract the feature maps from our PViT model by using the RoIAlign layer. Next, we take the 2D proposal at a frame into a 3D RoI by replicating it along the temporal axis, followed by a temporal global average pooling. Then, we max-pooled the RoI features and fed them to an MLP classifier for prediction.
Optimization details. To allow a direct comparison, we used the same configuration as in MViTv2 [22]. We trained 16 frames with sample rate 4, depth of 16 layers and batch-size 32 (comprising 16 videos and 16 auxiliary synthetic datasets) on 8 RTX 3090 GPUs. We train our network for 30 epochs with an SGD optimizer. We use $lr = 0.03$ with a weight decay of $1e{-}8$ together with early-stopping and a half-period cosine schedule of learning rate decaying.

Dataset	Available Annots.	#Training Samples ($\times 10^3$)	Real/Synt.
PHAV	D+S	39.9	Synt.
SURREACT	D+S+P3D	108.3	Synt.
ElderSim	P3D	48.8	Synt.
HyperSim	N+D	31.1	Synt.
EHOI	B+S	20.0	Synt.
SomethingElse	-	54.91	Real
SSv2	-	157.4	Real
AVA-2.2	-	193.3	Real
Ego4D	-	41.1	Real
Diving48	-	15.0	Real

Table 7. **Real and synthetic dataset details**. We show (a) Top: the auxiliary synthetic datasets, and (b) Bottom: downstream real datasets. The available annotations are depth maps (D), segmentation (S), 3D poses (P3D), normal maps (N) and boxes (B).

Training details. We use a standard crop size of 224 and we jitter the scales from 256 to 320. We use the same ground-truth boxes and proposals that overlap with ground-truth boxes by $IoU > 0.9$ as in [7].
Inference details. We perform inference on a single clip with 16 frames. For each sample, the evaluation frame is centered in frame 8. We take 1 spatial crop of 224 with 10 different randomly sampled clips to aggregate predictions over a single video in testing.

D. Additional Synthetic Datasets Details

Here we provide additional information about the "auxiliary synthetic datasets" (Section D.1), as well as the licenses and privacy policies for these datasets (Section D.2). Figure 6 shows examples of the synthetic videos we used to train on, while Table 7 presents the size of training samples across all synthetic and real datasets.

D.1. Auxiliary Synthetic Datasets

Synthetic datasets. There has been recent interest in learning video understanding from synthetic data, including several popular synthetic datasets that have been proposed to improve video understanding. More specifically, a novel approach to data generation has been proposed by SURREACT [32] and UESTC [15] for synthesizing humans for actions. KIST SynADL [13] is a large-scale synthetic dataset of elders' activities generated by the ElderSim engine [14]. The PHAV [5] dataset is a human action dataset that relies on the procedural generation of modern game engines. NTU RGB+D [30] and UESTC RGB-D [15] are large-scale synthetic datasets that was proposed in order to allow the training of large video models for video understanding. HyperSim [29] is a photo-realistic synthetic dataset for holistic indoor scene understanding.

Egocentric Human-Object Interactions (EHOI) [19] explores hand-object interaction in an industrial environment involving different objects, e.g. power supply, electrical panels, sockets, and more. In spite of the fact that these datasets contain different dataset styles, our approach is able to enhance video understanding models by utilizing synthetic data from various sources with multiple types of scene annotations. Next, we provide more details for each dataset separately.

SURREACT [32]. The SURREACT dataset, which stands for Synthetic hUmans foR REal ACTions, renders video sequences from 3D skeleton joints by using a Skinned Multi-Person Linear Model (SMPL). The ground truth joints are extracted either by Kinect v2, or HMML [16]. SURREACT consists of **(1) NTU RGB+D**, which is a large-scale dataset for RGB-D human action recognition. It consists of 56,880 samples of 60 action classes collected from 40 subjects. The actions are generally categorized into three categories: 40 daily actions (e.g., drinking, eating, reading), nine health-related actions (e.g., sneezing, staggering, falling down), and 11 mutual actions (e.g., punching, kicking, hugging). These actions take place under 17 different scene conditions corresponding to 17 video sequences (i.e., S001–S017). The actions were captured using three cameras with different horizontal imaging viewpoints, namely, $-45°$,$0°$, and $+45°$ degrees. Last, multi-modality information is provided for action characterization, including depth maps, 3D skeleton joint position, RGB frames, and infrared sequences; and **(2) UESTC RGB-D**, which contains 40 categories of aerobic exercise. The authors utilized two KinectV2 cameras in 8 fixed directions and 1 round direction to capture these actions with the data modalities of RGB video, 3D skeleton and depth map sequences.

HyperSim [29]. The HyperSim dataset is a high-resolution dataset consisting of 77,400 images from 461 indoor scenes with detailed per-pixel labels and corresponding ground truth geometry. It contains material and lighting information for every scene as well as dense per-pixel semantic instance segmentation, as well as complete camera information for every image. HyperSim was originally designed to handle the challenging per-pixel annotation of real data.

KIST SynADL [13]. KIST SynADL is a synthetic dataset that focuses on elders' daily activities, which differ from other natural actions due to their high degree of variety. The activities of elders, such as *sitting down* or *washing face*, are more consistent psychically, shorter, and often rely on body position. Last, the dataset is generated using ElderSim [14], a synthetic action simulation platform aimed at generating synthetic data on elders' daily activities. Throughout the paper, we refer to ElderSim as KIST SynADL.

Procedural Human Action Videos (PHAV). The PHAV dataset is a diverse, realistic, and physically plausible dataset of human action videos. It contains a total of 39,982 videos, with more than 1,000 examples of each action in 35 categories across 7 different environments and 4 types of weather. The data is generated based on the existing motion-based real database CMU MOCAP database, for basic human animations. One of its key components is the use of Ragdoll physics to animate a human model while respecting basic physics properties such as connected joint limits, angular limits, weight, and strength. The videos are generated at 30fps and a resolution of 340x256.

Egocentric Human-Object Interactions (EHOI) [19]. EHOI is a synthetic image dataset that explores hand-object interaction in an industrial environment involving different objects, such as a power supply, electrical panels, sockets, etc. To create 3D models, several 3D scanners are applied, then using Blender, the authors generate the following: (1) photo-realistic RGB images; (2) depth maps; (3) semantic segmentation masks, objects, and hand-bounding boxes with contact states; and (4) distance between hands and objects in 3D space. The generated synthetic dataset contains a total of 20,000 images, 29,034 hands (of which 14,589 are involved in an interaction), 123,827 object instances (14,589 of which are active objects), and 19 object categories including portable industrial tools (e.g., screwdrivers, electrical boards) and instruments.

D.2. Licenses and Privacy

The license, PII, and consent details of each dataset are in the respective papers. In addition, we wish to emphasize that the datasets we use do not contain any harmful or offensive content, as many other papers in the field also use them. Thus, we do not anticipate a specific negative impact, but, as with any Machine Learning method, we recommend to exercise caution.

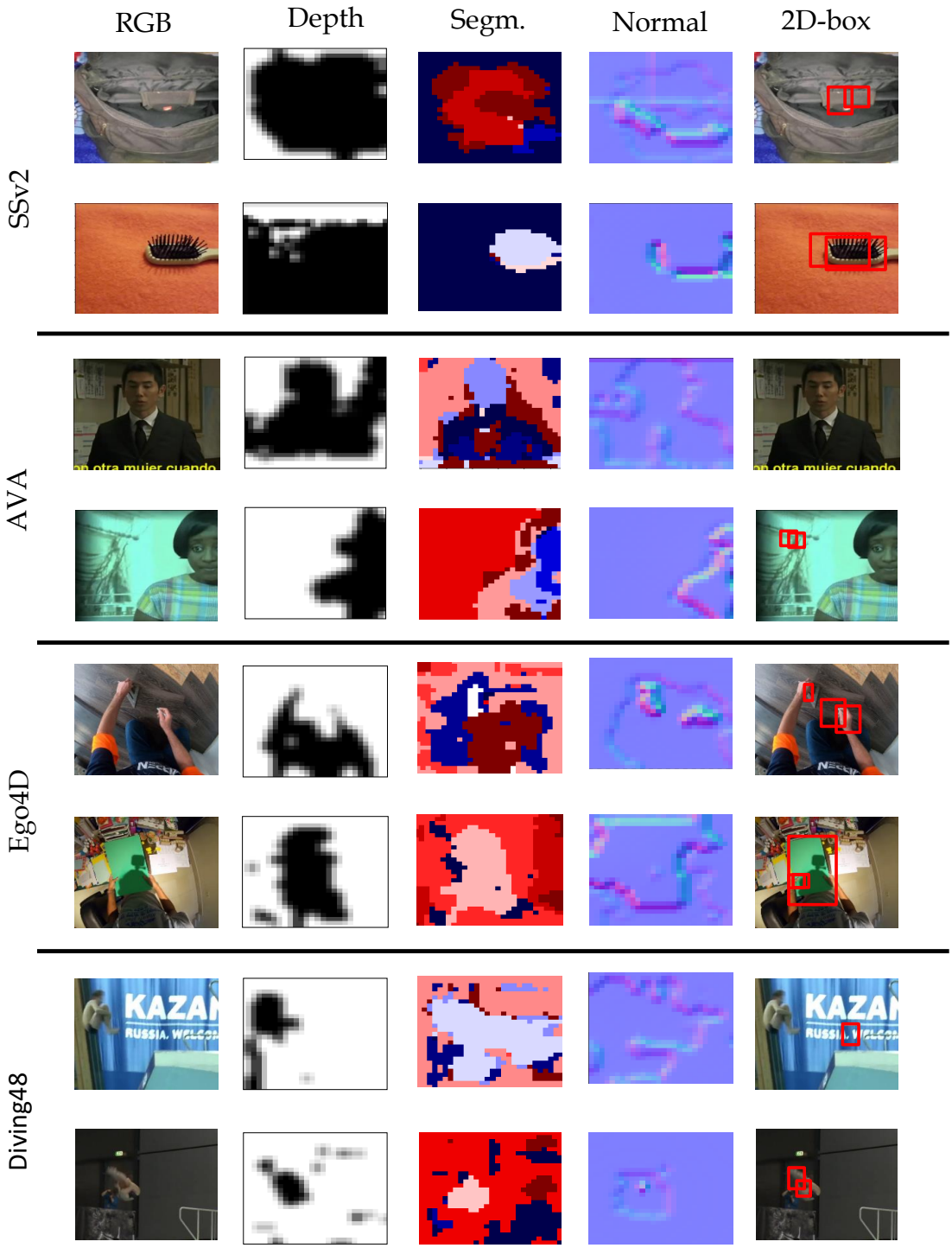

Figure 5. **Qualitative visualization of the "Task Prompts"**. Visualization of the output of the "task prompts" prediction heads on frames from the SSv2, Diving48, Ego4D, and AVA datasets. The model was trained on the SomethingElse dataset for action recognition. The predictions are the head outputs, H_i, for depth, normal, part-semantic segmentation and hand-object 2D boxes. It can be observed that the task prompts produce meaningful maps, despite not receiving labels for the real videos.

Figure 6. **Synthetic Datasets Visualization.** Our training datasets for PViT consist of several synthetic datasets that each emphasize different topics, including multi-views, static objects, hand-object interaction, and human motion activities.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021. 7, 2

[2] Elad Ben Avraham, Roei Herzig, Karttikeya Mangalam, Amir Bar, Anna Rohrbach, Leonid Karlinsky, Trevor Darrell, and Amir Globerson. Bringing image scene structure to video via frame-clip consistency of object tokens. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 7, 1, 2

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 6, 7, 2, 3

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6, 2

[5] CR De Souza, A Gaidon, Y Cabon, and AM Lopez Pena. Procedural generation of videos to train deep action recognition networks. In *CVPR*, 2017. 4

[6] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 7, 2

[7] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019. 6, 7, 2, 4

[8] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, page 5, 2017. 3

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. *CoRR*, abs/2110.07058, 2021. 4

[10] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6047–6056. IEEE Computer Society, 2018. 4

[11] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[12] Geoffrey E. Hinton, Nitish Srivastava, A. Krizhevsky, Ilya Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv*, abs/1207.0580, 2012. 3

[13] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications, 2020. 4, 5

[14] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. *ArXiv*, abs/2010.14742, 2020. 4, 5

[15] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition, 2019. 4

[16] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video, 2018. 5

[17] Tae Soo Kim and Gregory D. Hager. Safcar: Structured attention fusion for compositional action recognition, 2020. 1, 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3, 4

[19] Rosario Leonardi, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Egocentric human-object interaction detection exploiting synthetic data, 2022. 5

[20] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning, 2022. 1, 2

[21] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[22] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 6, 7, 8, 2, 3, 4

[23] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding, 2018. 2

[24] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 2

[25] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3888–3897, 2019. 6, 2

[26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 1, 2, 3

[27] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 3

[28] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers, 2021. 6, 2

[29] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 4, 5

[30] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 4

[31] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. 1, 2

[32] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. In *IJCV*, 2021. 4, 5

[33] Chuhan Zhang, Ankush Gputa, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3