OMNIVORE: A Single Model for Many Visual Modalities

Rohit Girdhar*    Mannat Singh*    Nikhila Ravi*    Laurens van der Maaten    Armand Joulin    Ishan Misra*
Meta AI
https://facebookresearch.github.io/omnivore

Abstract

Prior work has studied different visual modalities in isolation and developed separate architectures for recognition of images, videos, and 3D data. Instead, in this paper, we propose a single model which excels at classifying images, videos, and single-view 3D data using exactly the same model parameters. Our ‘OMNIVORE’ model leverages the flexibility of transformer-based architectures and is trained jointly on classification tasks from different modalities. OMNIVORE is simple to train, uses off-the-shelf standard datasets, and performs at-par or better than modality-specific models of the same size. A single OMNIVORE model obtains 86.0% on ImageNet, 84.1% on Kinetics, and 67.1% on SUN RGB-D. After finetuning, our models outperform prior work on a variety of vision tasks and generalize across modalities. OMNIVORE’s shared visual representation naturally enables cross-modal recognition without access to correspondences between modalities. We hope our results motivate researchers to model visual modalities together.

1. Introduction

Computer vision research spans multiple modalities related to our perception of the visual world, such as images, videos, and depth. In general, we study each of these modalities in isolation, and tailor our computer vision models to learn the best features from their specificities. While these modality-specific models achieve impressive performance, sometimes even surpassing humans on their specific tasks, they do not possess the flexibility that a human-like vision system does—the ability to work across modalities. We argue that the first step towards a truly all-purpose vision system is to build models that work seamlessly across modalities, instead of being over-optimized for each modality.

Beyond their flexibility, such modality-agnostic models have several advantages over their traditional, modality-specific counterparts. First, a modality-agnostic model can perform cross-modal generalization: it can use what it has learned from one modality to perform recognition in other modalities. For example, it can recognize pumpkins in 3D images even if it has only seen labeled videos of pumpkins. In turn, this allows existing labeled datasets to be used more
effectively: it becomes possible to train models on the union of vision datasets with different input modalities. Second, it saves the research and engineering effort spent on optimizing models for a specific modality. For example, image and video models have followed a similar trajectory of evolution, from hand-crafted descriptors [45, 53] to convolutional networks [34, 87] and, eventually, vision transformers [5, 21]; however, each had to be developed and tuned individually. A common architecture would make scientific progress readily available to users of any visual modality. Finally, a model that operates on many visual modalities is naturally multi-modal and can easily leverage new visual sensors as they become available. For instance, a modality-agnostic recognition model running on a robot can readily exploit a new depth sensor when it is installed on that robot. Despite such clear advantages, modality-agnostic models have rarely been studied and their performance compared to their modality-specific counterparts has been disappointing. There are many reasons that explain this situation, such as the need for a flexible architecture with enough capacity to learn modality-specific cues from the different modalities; and enough compute to train it on video, images, and single-view 3D simultaneously.

This paper develops a modality-agnostic vision model that leverages recent advances in vision architectures [21, 49]. The model we develop is “omnivorous” in that it works on three different visual modalities: images, videos, and single-view 3D. Our OMNIVORE model does not use a custom architecture for each visual modality. It performs recognition on all three modalities using the same, shared model parameters. It works by converting each input modality into embeddings of spatio-temporal patches, which are processed by exactly the same Transformer [88] to produce a representation of the input. We train OMNIVORE on a collection of standard, off-the-shelf classification datasets that have different input modalities. Unlike prior work [33, 75], our training does not use explicit correspondences between different input modalities.

Our experiments demonstrate the advantages of our OMNIVORE models. Surprisingly, we find that OMNIVORE representations generalize well across visual modalities (see Figure 1) even though OMNIVORE was not explicitly trained to model cross-modal correspondences. These capabilities emerge without explicit cross-modal supervision simply due to the parameter sharing between models for different modalities. On standard image, video, and single-view 3D benchmarks, OMNIVORE performs at par with or better than modality-specific vision models with the same number of parameters. The same OMNIVORE model obtains 85.6% top-1 accuracy on ImageNet-1K, 83.4% top-1 on Kinetics-400, and 67.4% top-1 accuracy on SUN RGB-D. OMNIVORE’s strong generalization capabilities also extend to transfer learning experiments. OMNIVORE performs at par with recent large transformers on ImageNet-1K, sets a new state-of-the-art on action recognition benchmarks such as EPIC-Kitchens-100, Something Something-v2, and on single-view 3D classification and segmentation benchmarks. We believe our work presents a compelling argument for shifting towards the development of vision models that can operate on any visual modality.

2. Related Work

We build on prior work in ConvNet architectures, Transformers, multi-modal learning, and multi-task learning. ConvNet architectures in vision. ConvNet architectures [26, 46] have been popular for many computer vision tasks in images, video, and 3D recognition. 2D convolutions are the main building block in ConvNets for images [34, 44, 75, 80], whereas 3D convolutions are used on 3D data [18, 32] or are combined with 2D convolutions for recognition of videos [13, 86, 87]. I3D [13] introduced a way to “inflate” 2D image convolutions into 3D convolutions, which allows 3D ConvNets for videos and 3D data to leverage image data indirectly via initialization from pre-trained image models. Since video and 3D datasets are relatively small, they benefit from inflated pretrained image networks. However, while the inflation technique is applicable only to model finetuning, OMNIVORE models are pretrained jointly on images, videos, and single-view 3D data.
Transformers in vision. The Transformer architecture [88] originally proposed for NLP tasks has been successfully applied in computer vision on images [11, 21, 68, 84, 89, 90], video [5, 8, 28, 29, 50, 64], and 3D data [58, 66, 98]. Models such as ViT [21], Swin [49], and MViT [24] perform competitively on benchmark tasks such as image classification, detection, and video recognition. For example, Swin [49, 50] and MViT [24] require minimal changes to be used in image or video recognition tasks. Similarly, the Perceiver [38] can model image, point cloud, audio, and video inputs. However, all these studies train separate models for each visual modality. Instead, we train a single model on multiple input modalities simultaneously, which equips our model with cross-modal generalization capabilities.

Multi-modal learning. Our work uses multiple visual modalities to train the model. Multi-modal learning architectures may involve training separate encoders for each type of input modality. For example, a range of tasks require training separate encoders for images and text [15, 30, 41, 55, 57], for video and audio [3, 4, 60, 61, 65, 69], or for video and optical flow [75]. Recently, Transformers have been used to fuse multiple modalities: Transformers have been used to fuse features in vision-and-language tasks [2, 17, 37, 40, 47, 54, 79, 82] and video-and-audio tasks [62], video-and-image tasks [7], and even tasks that involve video, audio, and text [1]. Unlike our work, most prior work assumes that all input modalities are in correspondence and available simultaneously, which restricts them to using only multi-modal datasets. In our work, we train a single model on different visual modalities without assuming simultaneous access to all modalities. This allows us to leverage standard off-the-shelf single-modality vision datasets and show that using a single shared encoder naturally leads to cross-modal generalization.

Multi-task learning. Our work is also related to studies on multi-task learning [14], which develop models that output predictions for multiple tasks on the same input [23, 27, 43, 56, 59, 97]. Such multi-task learners are known to work well when the target tasks exhibit strong similarities [59, 95]. They differ from OMNIVORE in that they operate on a single input modality but are trained to perform multiple tasks. By contrast, our models are trained to perform a single task (i.e., classification) on a variety of input modalities. Other multi-task learners operate on multi-modal inputs [39], but they use hand-designed model components for each modality.

3. Approach

Our goal is to learn a single model that can operate on three major visual modalities: images, videos, and single-view 3D. Because the model’s input modalities have different sizes and layouts—videos have a temporal axis and single-view 3D has an extra depth channel—this poses a challenge in designing the model. To overcome this challenge, we adopt the Transformer [88] architecture because the self-attention mechanism gracefully handles variable-sized inputs. Figure 2 presents an overview of our approach.

3.1. The Omnivore Model

We convert all visual modalities into a common format by representing them via embeddings. Our model then uses a series of spatio-temporal attention operations to construct a unified representation of the different visual modalities.

Input patches. We represent the different types of visual input as a 4D tensor $X \in \mathbb{R}^{T \times H \times W \times C}$, where $T$ is the size of the temporal dimension, $H$ and $W$ of the spatial dimensions, and $C$ of the channel dimension. Thus, RGB images $I \in \mathbb{R}^{1 \times H \times W \times 3}$ have $T = 1$ frame with $C = 3$ channels, RGB videos $V \in \mathbb{R}^{T \times H \times W \times 3}$ have $T > 1$ frames, and single-view 3D images $D \in \mathbb{R}^{1 \times H \times W \times 3}$ have $T = 1$ frame with three RGB channels and one depth channel.

We follow [21, 49, 50] and split the input into a collection of patches. We illustrate this process in Figure 2. Specifically, we convert the visual input $X$ into a set of 4D sub-tensors $x$ of size $t \times h \times w \times c$. Images $I$ are split into a set of non-overlapping image patches of size $1 \times h \times w \times 3$. Similarly, videos $V$ are split into a set of non-overlapping spatio-temporal patches of shape $t \times h \times w \times 3$. For single-view 3D images $D$, the image (RGB) and depth (D) channels are converted separately into patches of size $1 \times h \times w \times 3$ and $1 \times h \times w \times 1$, respectively.

Model architecture. Our model $f$ maps the resulting spatio-temporal visual patches into a shared representation $\Phi$ for images, videos, and single-view 3D. We design the model to enable maximal parameter sharing across visual modalities. The input layer of the model processes each patch $x$ independently, and projects the patches into an embedding $e$ using a linear layer followed by a LayerNorm [6] (linear+LN). Each patch $x$ of shape $t \times h \times w \times c$ is converted into an embedding of size $d$. We use the same layers to embed all the three-channel RGB patches, i.e., for image patches, video patches, and patches of the first three channels of a single-view 3D image. We zero-pad the single-frame patches on one side to ensure all patches have the same shape, $t \times h \times w \times 3$. We use a separate linear+LN layer to embed the depth-channel patches and add its output to the embedding of the corresponding RGB patch.

We use the same model (parameters) to process all the resulting embeddings. While OMNIVORE can use any vision transformer architecture [21, 24] to process the patch embeddings, we use the Swin transformer architecture [49] as our base model given its strong performance on image and video tasks. We rely on the self-attention [88] operation for spatio-temporal modeling across the patch embeddings, $e$. Akin to [49], the self-attention involves patch embeddings from spatially and temporally nearby patches. We also use
two sets of relative positional encodings: one for the spatial dimension and the other for the temporal dimension.

3.2. Training the OMNIVORE Model

The OMNIVORE model \( f \) creates a single embedding \( f(X) = \Phi \) for multiple types of visual inputs. We train our model using a collection of classification tasks that provide inputs \( \{(X_i, y_i)\} \) with a visual input, \( X_i \), and a label, \( y_i \). For example, we train most OMNIVORE models jointly on the ImageNet-1K dataset for image classification, the Kinetics-400 dataset for action recognition, and the SUN RGB-D dataset for single-view 3D scene classification.

This approach is similar to multi-task learning [14] and cross-modal alignment [15], but there important differences. In particular, we neither assume that the input observations are aligned (i.e., we do not assume access to correspondences between images, videos, and 3D data) nor do we assume that these datasets share the same label space. To achieve this, we employ dataset-specific linear classification layers on top of the final representation, \( \Phi \), produced by the model. The training loss of a sample is computed based solely on the output of the classification layer that corresponds to that sample’s source dataset.

Loss and optimization. We train OMNIVORE to minimize the cross-entropy loss on the training datasets using mini-batch SGD. We experiment with two different mini-batch construction strategies for SGD. In our first strategy, we construct mini-batches from each dataset (modality) separately. This strategy is easy to implement but alternating between datasets may potentially lead to training instabilities. Hence, we experiment with a second strategy that constructs mini-batches that mix samples from all datasets. We evaluate both mini-batch construction strategies in § 4.3.

4. Experiments

We perform a series of experiments to assess the effectiveness of OMNIVORE. Specifically, we compare OMNIVORE models to their modality-specific counterparts and to state-of-the-art models on a variety of recognition tasks. We also ablate several design choices we made in OMNIVORE.

Pre-training datasets. We train OMNIVORE on images from the ImageNet-1K dataset [73], videos from the Kinetics dataset [42], and single-view 3D images from the SUN RGB-D dataset [76]. We measure the top-1 and top-5 classification accuracy of our models on the respective validation sets. We note that the three datasets have negligible overlap in their visual concepts: ImageNet-1K focuses on object-centric classes, Kinetics-400 on action classes, and SUN RGB-D on indoor scene classes.

Images. The ImageNet-1K (IN1K) dataset has \( \sim 1.2M \) training and 50K validation images that comprise 1,000 classes.

Videos. The Kinetics-400 (K400) dataset consists of \( \sim 240K \) training and 20K validation video clips that are 10 seconds long, and are labeled into one of 400 action classes.

Single-view 3D. The SUN RGB-D dataset has \( \sim 5K \) training and \( \sim 5K \) validation RGBD images with 19 scene classes. Following [72], we convert the depth maps into disparity maps.

Implementation details. We use the Swin transformer [49, 50] architecture as the backbone for OMNIVORE, and attach linear heads for each target dataset. At training time, we use a resolution of \( 224 \times 224 \) and train using standard image augmentations [84] on ImageNet. For kinetics, we sample 32 frames at stride 2. SUN RGB-D is processed similarly to ImageNet but we randomly drop the RGB channels with a probability of 0.5 in order to encourage the model to use the depth channel for recognition as well. We provide complete implementation details in Appendix A. Our models are optimized using AdamW [52] for 500 epochs where a single epoch consists of one epoch each for ImageNet-1K and Kinetics, and 10 epochs for SUN RGB-D.

Transfer datasets and metrics. We evaluate OMNIVORE in transfer learning experiments on a diverse set of image, video, and single-view 3D tasks; see Table 1 for a summary. We present details on the experimental setup in Appendix B.

Images. We evaluate OMNIVORE on fine-grained object recognition on the iNaturalist-2018 dataset [36], fine-grained classification on the Oxford-IIIT Pets dataset [67], and in scene classification on the Places-365 dataset [100].

Videos. We use the Something Something-v2 dataset, which has a special emphasis on temporal modeling for action recognition. We also evaluate the EPIC-Kitchens-100 dataset, which has 100 hours of unscripted egocentric video. Each clip is labeled with a verb and a noun that together form an action. Our model is trained to recognize all 3,806 actions, i.e., verb-noun pairs in the dataset. We marginalize over verbs to obtain noun predictions and vice versa.

Single-view 3D. We use the NYU-v2 dataset for single-view 3D scene classification and segmentation. We follow the setup from [33] for scene classification and [10, 33] for segmentation. For segmentation, we follow [49] and use the UPerNet [91] head with the Swin trunk.

Table 1. Transfer datasets used to evaluate OMNIVORE on image, video and single-view 3D modalities. The table reports the task, number of classes (#cls), number of training samples (#train), and number of validation samples (#val) for each dataset.
Table 2. O MNIVORE vs. modality-specific models that have the same model architecture and number of parameters. OMNIVORE is a single model trained from scratch jointly on the IN1K, K400 and SUN datasets whereas the modality-specific models are trained specifically for each dataset (modality). The ImageSwin model is trained from scratch while the VideoSwin and DepthSwin models are finetuned from the ImageSwin model. OMNIVORE performs at-par or outperforms modality-specific models.

Table 3. Comparing OMNIVORE with modality-specific models after finetuning the models on seven downstream tasks. Results are presented for three different model sizes: T, S, and B. Our image specific model is pretrained on IN1K. The video specific and single-view 3D specific models are both initialized using inflation from the pretrained image-specific model and finetuned on K400 and SUN RGB-D respectively. OMNIVORE models are at-par with or outperform modality-specific models on nearly all downstream tasks.

4.1. Comparison with Modality-Specific Models

We compare OMNIVORE to models trained on a specific visual modality. We train OMNIVORE from scratch jointly on the IN1K, K400, and SUN datasets. Our modality-specific baseline models use the same Swin transformer architecture as OMNIVORE; we refer to them as ImageSwin, VideoSwin, and DepthSwin. Excluding the patch-embedding linear layers, these models have the same number of parameters as OMNIVORE. Following standard practice [49, 50], the ImageSwin model is trained on IN1K, whereas VideoSwin and DepthSwin models are finetuned by inflating the ImageSwin model. We experiment with three model sizes: \textit{viz.}, Swin-T, Swin-S, and Swin-B.\footnote{We refer to [49] for details on these model sizes.}

Pretraining performance. In Table 2, we compare OMNIVORE to modality-specific models on the pretraining datasets. The results in the table show that across model sizes, OMNIVORE models match or exceed the performance of their modality-specific counterparts. This observation supports our hypothesis that it is possible to learn a single visual representation that works across visual modalities. OMNIVORE learns representations that are as good as modality-specific representations using the same training data, same model parameters and same model capacity. This implies that OMNIVORE provides a viable alternative to the pretrain-then-finetune paradigm commonly used to deploy modality-specific models: it can deliver the same or better recognition accuracy with a third of the parameters.

From our results, we also observe that higher-capacity models benefit more from omnivorous training. OMNIVORE models using the larger Swin-B architecture improve over their modality-specific counterparts on both IN1K and K400, whereas the smallest Swin-T model does not.

Figure 3 presents a detailed analysis of the improvements of OMNIVORE over the VideoSwin baseline (both using the Swin-B architecture) on the K400 dataset. Here VideoSwin is pre-trained on IN1K and finetuned on K400, whereas OMNIVORE is trained jointly on IN1K, K400, and SUN RGB-D. Both models use the the Swin-B architecture. OMNIVORE particularly improves the recognition of classes that require reasoning about parts of the human body such as the hands, arms, head, mouth, hair \textit{etc}. We surmise this is because joint training on images helps OMNIVORE to learn a better model of the spatial configuration of parts.

Transfer learning performance. We compare OMNIVORE to modality-specific models by finetuning on various downstream tasks. Table 3 presents the results of these experiments. We observe that OMNIVORE transfers better than modality-specific models on nearly all downstream tasks. In particular, OMNIVORE provides significant gains on video-recognition tasks, even though it does not get any additional video supervision during pre-training compared to the baseline. We reiterate that OMNIVORE has the same model ca-
Table 4. Comparing OMNIVORE with state-of-the-art models on the image, video, and single-view 3D classification datasets used to pre-train OMNIVORE. OMNIVORE performs on par with or better than state-of-the-art models on all three pre-training tasks, including modality-specific models of similar size.

Table 5. Comparing OMNIVORE with state-of-the-art models in image classification finetuning experiments on three datasets. OMNIVORE representations generalize well to scene classification (P365) and fine-grained classification (iNat18, Pets). ↑ indicates finetuning on a higher resolution image (384×384px; see [85]).

Transfer learning performance. We compare OMNIVORE models to modality-specific models by finetuning on downstream tasks. In Table 5, we report results on image classification. OMNIVORE models outperform prior state-of-the-art in scene classification on Places-365, and in fine-grained classification on iNaturalist-2018 and Oxford-IIIT Pets.

We finetune OMNIVORE on video-classification and report the results in Table 6. On the EPIC-Kitchens-100 dataset, the OMNIVORE Swin-B model achieves the absolute best performance across verb, noun, and verb-noun pair (action) classification. Similarly, on the SSv2 dataset, which requires temporal reasoning, OMNIVORE outperforms all prior work. This suggests that OMNIVORE representations transfer well to temporal-reasoning tasks – OMNIVORE sets a new state-of-the-art while outperforming architectures specialized for these video tasks.

Finally, in Table 7, we report finetuning results for RGD3 scene classification and segmentation. While prior work relies on specialized 3D operators [10], fusion techniques [93], or depth encoding schemes [33], OMNIVORE uses a generic architecture and operates directly on disparity. OMNIVORE achieves state-of-the-art performance on both the scene classification and segmentation tasks.

4.3. Ablation Study

We ablate some of OMNIVORE’s key design choices in Table 8. Together, the results suggest OMNIVORE’s performance is relatively stable under different design choices. For a faster turnaround time in the ablations, we train the model for 300 epochs.

Training from scratch or finetuning. We compare training OMNIVORE models from scratch on different modalities (top row) with initializing the model via image classification followed by finetuning on all modalities (second row). For the finetuning result, we initialize OMNIVORE (Swin-B) using a pretrained ImageNet-21K model followed by joint finetuning on IN1K, K400, and SUN for 100 epochs. The
Table 6. Comparing Omnivore with state-of-the-art models in video classification finetuning experiments on two datasets. We highlighted columns that show the two primary classification metrics used in prior work. Omnivore models obtain state-of-the-art results on both datasets.

| Method                  | EK100 verb | EK100 noun | EK100 action | SSv2 top-1 | SSv2 top-5 |
|-------------------------|------------|------------|--------------|------------|------------|
| SlowFast [25]           | 65.6       | 50.0       | 38.5         | 63.0       | 88.5       |
| TimeSformer [8]         | –          | –          | –            | –          | –          |
| MViT-B-24 [24]          | –          | –          | –            | 62.4       | –          |
| TAR [74]                | 66.0       | 53.4       | 45.3         | –          | –          |
| VIMPAC [83]             | –          | –          | –            | 68.1       | –          |
| ViViT-L [5]             | 66.4       | 56.8       | 44.0         | 65.9       | 89.9       |
| MFormer-L [70]          | 67.1       | 57.6       | 44.1         | 68.1       | 91.2       |
| ORViT [35]              | 68.4       | 58.7       | 45.7         | 69.5       | 91.5       |
| VideoSwin-B [50]        | 67.8       | 57.0       | 46.1         | 69.6       | 92.7       |
| Omnivore (Swin-B)       | 69.5       | 61.7       | 49.9         | 71.4       | 93.5       |

Table 7. Comparing Omnivore with state-of-the-art models in RGBD finetuning experiments on the NYU-v2 dataset. The left column shows the scene classification accuracy while the right column shows the mean intersection-over-union of semantic segmentation. Omnivore outperforms prior art in RGBD classification and segmentation.

| Method                  | Classification | Segmentation |
|-------------------------|----------------|--------------|
| DF²Net [48]             | 65.4           | ×            |
| TRecgNet [22]           | 69.2           | ×            |
| ShapeConv [10]          | ×              | 51.3         |
| BCMFP + SA-Gate [16]    | ×              | 52.4         |
| TCD [93]                | ×              | 53.1         |
| Omnivore (Swin-B)       | 79.4           | 55.1         |
| Omnivore (Swin-L)       | 79.8           | 56.8         |

5. Cross-Modal Generalization

A key advantage of Omnivore over modality-specific models is that it can generalize across visual modalities. This generalization emerges naturally because we use the same model for all modalities. Our model is neither trained with corresponding data across modalities nor with any cross-modal consistency losses.

![Graph showing cross-modal generalization](Image)

Figure 4. Retrieving depth maps given RGB images on the ImageNet-1K dataset. We show retrieved depth maps from the IN1K training set (right) for RGB image queries from the IN1K validation set (left). Although Omnivore was not trained on IN1K depth maps, the shared visual representation enables it to retrieve depth maps that are semantically similar to the query.
Retrieval across images and depth. We use the OMNIVORE representation to retrieve depth maps given an RGB image. To create a database of depth maps, we run a monocular depth-prediction model [72] on the ImageNet-1K train set. We note that OMNIVORE was not trained on ImageNet-1K depth maps nor on predicted depth. We use the ImageNet-1K val set (RGB) images as queries. Figure 4 shows five examples of retrieved maps. These results illustrate that OMNIVORE constructs good depth-map representations, even though it had not previously observed ImageNet-1K depth maps during training. We emphasize that this cross-modal generalization ability is not the result of explicitly learning correspondences between visual modalities [33, 75]. Instead, it emerges due to the use of an almost entirely shared encoder for those modalities.

Classifying based on different modalities.

| Method         | RGB  | D    | RGBD |
|----------------|------|------|------|
| OMNIVORE (Swin-B) | 84.3 | 63.1 | 83.7 |

To quantitatively measure OMNIVORE’s generalization performance across different modalities, we perform k-nearest neighbor (k-NN, \( k = 20 \)) classification experiments on the ImageNet-1K dataset using the predicted depth maps. We extract OMNIVORE representations from the RGB images on the val set and measure the model’s ability to retrieve images, RGBD images, and depth-only images from the train set. We observe that OMNIVORE produces a representation that allows for successful k-NN classification, which demonstrates its strong generalization performance. Surprisingly, we observe a high accuracy is attained even when retrieving depth-images, which provide less information about the object class than RGB images.

Retrieval across all modalities. We further probe the OMNIVORE visual representations in retrieval experiments across images, videos, and depth maps. We use the RGB images from the ImageNet-1K val set as queries and use them to retrieve similar depth maps from ImageNet-1K (predicted depth) and videos from Kinetics-400. Figure 1 shows examples of the resulting retrievals. The results illustrate how OMNIVORE supports retrieval of visual concepts across images (RGB), single-view 3D (RGBD), and videos (RGBT) using its shared representation space.

Bridging frame-based and clip-based video models. OMNIVORE’s cross-modality generalization capabilities also make it more robust to changes in lengths of videos to be classified. We demonstrate this in in Figure 5, where we classify videos using different length clips at inference time. The model is trained with 32 frames at stride 2, and by default uses 4 clips of the same length and stride to cover the full 10 second video at inference time. In this experiment, we vary the clip length from 1 to 32, increasing the number of clips proportionally to still cover the full video in each case. The results show that OMNIVORE’s performance degrades more gracefully as the video length decreases. Notably, OMNIVORE outperforms the baseline by 18.5% at a clip length of 1 frame (frame-level inference). This suggests that joint training on images and videos enables the model to use both temporal and spatial cues effectively.

6. Discussion and Limitations

Although OMNIVORE presents an advance over traditional modality-specific models, it has several limitations. Our current implementation of OMNIVORE only works on single-view 3D images and does not generalize to other 3D representations such as voxels, point clouds, etc. A simple approach to deal with such inputs may be to render multiple single-view 3D images from such inputs and average OMNIVORE predictions over those images, but such an approach would not effectively leverage multi-view information. Another caveat is that depth inputs are not scale-invariant; we used normalizations to alleviate this issue [72]. Also, OMNIVORE focuses only on visual modalities, so co-occurring modalities such as audio are not used. OMNIVORE was pretrained using only classification and using using structured prediction tasks such as segmentation might yield richer representations. We leave such extensions to future work.

Ethical Considerations. Our study focuses on technical innovations in training models for visual recognition. These innovations themselves appear to be neutral from an ethics point-of-view. However, all ethical considerations that apply to other visual-recognition models apply equally to OMNIVORE. Any real-world deployment of a model like OMNIVORE is best preceded by a careful analysis of that model for ethical problems, including but not limited to: performance disparities between different user groups, associations that may be harmful to some users, and predictions that may propagate stereotypes.

Figure 5. Accuracy as a function of clip length on the K400 dataset. Models are trained on 32-frame clips but evaluated on clips of different length (with the same fps used for frame sampling). The performance of OMNIVORE degrades more gracefully than that of the VideoSwin-B model, and is still effective when doing frame-level inference (i.e., when the clip length is 1).
References

[1] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. arXiv preprint arXiv:2104.11178, 2021.

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. NeurIPS, 2020.

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In ICCV, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In ECCV, 2018.

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A video vision transformer. In ICCV, 2021.

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.

[7] Max Bain, Arsha Nagrani, Gil Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. arXiv preprint arXiv:2104.00650, 2021.

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In ICML, 2021.

[9] Ali Caglayan, Nevrez Imamoglu, Ahmet Burak Can, and Ryosuke Nakamura. When cnns meet random rnns: Towards multi-level analysis for rgb-d object and scene recognition. arXiv preprint arXiv:2004.12349, 2020.

[10] Jimming Cao, Hanchao Deng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconvl: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In ICCV, 2021.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégo, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In ICCV, 2021.

[13] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017.

[14] Rich Caruana. Multitask learning. Machine Learning, 1997.

[15] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In CVPR, 2016.

[16] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In ECCV, 2020.

[17] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In ECCV, 2020.

[18] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In CVPR, 2019.

[19] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In CVPR, 2020.

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. IJCV, 2021.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

[22] Kapeng Du, Limin Wang, Huiling Wang, Kai Zhao, and Gangshan Wu. Translate-to-recognize networks for rgb-d scene recognition. In CVPR, 2019.

[23] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In ICCV, 2015.

[24] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In ICCV, 2021.

[25] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In ICCV, 2019.

[26] Kunihiko Fukushima. A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern., 1980.

[27] Golnaz Ghiasi, Barret Zoph, Ekin D Cubuk, Quoc V Le, and Tsung-Yi Lin. Multi-task self-training for learning general representations. In ICCV, 2021.

[28] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In CVPR, 2019.

[29] Rohit Girdhar and Kristen Grauman. Anticipative Video representations from weakly aligned data. In ICCV, 2016.

[30] Raghav Goyal, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In CVPR, 2019.

[31] Christoph Grgoire Parreira, Radu Smeulders, and Antonio Torralba. The “something something” video database for learning and evaluating visual common sense. In ICCV, 2017.

[32] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In CVPR, 2018.

[33] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In CVPR, 2013.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR,
Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

David G Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265, 2019.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In CVPR, 2020.

Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In CVPR, 2019.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In CVPR, 2020.

Ishan Misra, Rohit Giridhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In ICCV, 2021.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In CVPR, 2016.

Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In CVPR, 2021.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In CVPR, 2021.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In NeurIPS, 2021.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.

Daniel Naimark, Omri Bar, Maya Zohar, and Amir Globerson. Object-region video transformers. In ICML, 2021.

Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-plus video transformers. arXiv preprint arXiv:2110.01155, 2021.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Sheikh, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In CVPR, 2018.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In ICCV, 2021.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. ICML, 2021.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, Llion Jones, and Jakob Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Imago: Vision transformer networks. arXiv preprint arXiv:2102.00719, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

David G Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265, 2019.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In CVPR, 2020.

Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In CVPR, 2019.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In CVPR, 2020.

Ishan Misra, Rohit Giridhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In ICCV, 2021.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In CVPR, 2016.

Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In CVPR, 2021.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In CVPR, 2021.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In NeurIPS, 2021.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.

Daniel Naimark, Omri Bar, Maya Zohar, and Amir Globerson. Object-region video transformers. In ICML, 2021.

Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-plus video transformers. arXiv preprint arXiv:2110.01155, 2021.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Sheikh, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In CVPR, 2018.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In ICCV, 2021.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. ICML, 2021.

Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Jakob Uszkoreit, Llion Jones, and Jakob Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Imago: Vision transformer networks. arXiv preprint arXiv:2102.00719, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

David G Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265, 2019.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In CVPR, 2020.

Kevis-Kokitsi Maninis, Ilija Radosavovic, and Iasonas Kokkinos. Attentive single-tasking of multiple tasks. In CVPR, 2019.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In CVPR, 2020.

Ishan Misra, Rohit Giridhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In ICCV, 2021.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In CVPR, 2016.

Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In CVPR, 2021.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In CVPR, 2021.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In NeurIPS, 2021.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In ECCV, 2012.

Daniel Naimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. arXiv preprint arXiv:2102.00719, 2021.

Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In ECCV, 2018.

Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In CVPR, 2021.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In CVPR, 2012.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In ICML, 2018.

Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multimodal self-supervision from generalized data transformations. arXiv preprint arXiv:2003.04298, 2020.

Mandela Patrick, Dylan Campbell, Yuki M Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In NeurIPS, 2021.
A. Implementation details for Pretraining

We train using AdamW with a batch size of 4096 for each dataset, and use a cosine learning rate (LR) schedule with linear warm up and cool down phases for the first and last 10% of training, respectively. We train for 500 epochs with a peak LR of $2 \cdot 10^{-3}$ and a weight decay of $5 \cdot 10^{-2}$. Swin-T, Swin-S and Swin-L use a window size of $8 \times 7 \times 7$, whereas Swin-B uses a window size of $16 \times 7 \times 7$. The models are trained with stochastic depth with a drop rate of 0.1 for Swin-T, 0.2 for Swin-S, and 0.3 for Swin-B, and Swin-L. We use exponential moving average (EMA) [71] with a decay of $10^{-4}$ and report the best results during training since EMA results peak before the end of training.

For IN1K and IN21K we use RandAugment [19], mixup [96], CutMix [94], label smoothing [81], and Random Erasing [99] with the same settings as used in [84], and color jittering of 0.4. For SUN RGB-D we clamp and normalize the disparity channel, drop the RGB channels with a probability of 0.5, and we also apply 0.5 Dropout [78] before the linear head when pre-training with ImageNet-21K. For Kinetics-400 we use mixup, CutMix and label smoothing, and Dropout of 0.5 before the linear head.

B. Details on the Transfer Tasks

B.1. Image Classification

We finetune all models on the downstream tasks for 100 epochs and optimize the models with mini-batch SGD. We use a half-wave cosine learning rate [51] and set the weight decay to zero. For all models, including the modality-specific models, we perform a grid search for the best learning rate in the range $[5e-3, 1e-2, 2e-2, 4e-2, 8e-2, 1e-1, 2e-1, 3e-1, 4e-1, 5e-1, 6e-1]$ and drop path in $[0.1, 0.3]$. We use the strong augmentations from [84] for finetuning. For the evaluations in Tables 3 and 5, we resize the images to shortest side of $224px$ and evaluate the models on the center crop of $224 \times 224$. For higher resolution ($384 \times 384$) evaluations in Table 5, we follow [49] to resize the images to shortest side of $384px$ and evaluate the models on the center crop of $384 \times 384$. We also increase the spatial window size for all the Swin models from 7 to 12.

B.2. Video Classification

In Table 3, we finetune video models using hyperparameters as described in [50]. For Something Something-v2, we finetune for 60 epochs with AdamW optimizer. We use half-wave cosine learning rate with warmup. We start the learning rate from $10^{-6}$ and linearly warmup to a peak learning rate of $6 \cdot 10^{-3}$ over 5% of the training, and rest 95% we use half-wave cosine schedule to decay the learning rate back to $10^{-6}$. We train the classification head with this learning rate, and the backbone with $0.1 \times$ the above learning rate. Throughout we use a weight decay of 0.05. We use a batch size of $4 \times 64$ distributed over 64 32GB GPUs. For EPIC-Kitchens-100, we use similar hyperparameters with only difference being that we use a peak learning rate of $1.8 \cdot 10^{-2}$ and we train for 150 epochs. These settings provided better performance for the modality-specific baseline, and we use it for finetuning both the baseline and OMNIVORE models.

In terms of preprocessing, at train time we sample a 32 frame video clip at stride 2 from the full video using temporal segment sampling as in [50]. We scale the short side of the video to $256px$, take a $224px$ random resized crop, followed by RandAugment and Random Erasing. At test time, we again sample a 32 frame clip with stride 2, scale the short side to $224px$ and take 3 spatial crops along the longer axis to get $224 \times 224$ crops. The final predictions are averaged over these crops.

For comparison to the state-of-the-art in Table 6, when finetuning OMNIVORE models trained with IN21K, we found slightly different hyperparameters to perform better. For Something Something-v2, we used peak learning rate of $1.2 \cdot 10^{-3}$ over 150 epochs. For EPIC-Kitchens-100, we used weight decay of 0.004, over 100 epochs, peak learning rate of $4 \cdot 10^{-4}$, with the same learning rate schedule for backbone and head. We also used cutmix augmentation and label smoothing. All other hyperparameters in both cases were as described earlier. We also use EMA with similar settings as used during pretraining.

B.3. Single-view 3D Tasks

NYU Scene classification. We follow the setup from [33] for scene classification and use 10 classes derived from the original 19 classes. In Table 7 (classification) the best Swin B and Swin L models were trained for 200 epochs with
starting learning rate of $5 \times 10^{-3}$, weight decay of 0 for Swin B and $1 \times 10^{-4}$ for Swin L. All other hyperparameters were as described earlier.

**NYU RGBD Segmentation.** We follow the training and evaluation setup from [10]. We follow the Swin segmentation architecture which uses an UperNet [91] head with the Swin trunk. All models are finetuned with AdamW [52] with a weight decay of 0.01. The learning rate follows a Polynomial Decay (power 1) schedule and starts at 0.00006. We warmup the learning rate for 1500 iterations and train the model with a batchsize of 32. All the depth maps in NYU are converted into disparity maps by using the camera baseline and focal length of the Kinect sensor.

**B.4. k-NN experiments**

**Extracting depth on ImageNet-1K.** We ran a monocular depth-prediction model [72] on the IN1K train set. We used the pretrained dpt_large model and followed the input image preprocessing steps as provided in [72].

**Classifying ImageNet-1K using different modalities.** For the experiments involving classification using different modalities, we extract features from the IN1K train set using the RGB, RGBD or just Depth (D) modalities, and on IN1K validation set using the RGB modality. We follow the k-NN protocol from [12] for evaluation and briefly describe it next. We extract the stage 3 [49] features and $L_2$ normalize them. For each validation feature as the query, we retrieve the nearest neighbors from the train set using euclidean distance, and take the top-$k$ closest matches. For each match we create a one-hot vector using its ground truth label, and scale it by $e^{s/\tau}$, where $s$ is the dot product between the feature of the matched image the query image, and $\tau$ is a temperature hyperparameter (set to 0.07). We compute an effective prediction for the query by summing the top-$k$ one-hot vectors. Similar processing is used for the visualizations in Figure 1 and Figure 4.

**C. Other Results**

We present the gain of OMNIVORE over the VideoSwin baseline ($\S$ 4.1 of the main paper) in Figs. 6 and 7.