# SViTT: Temporal Learning of Sparse Video-Text Transformers
# Supplemental Material

Yi Li[1] Kyle Min[2] Subarna Tripathi[2] Nuno Vasconcelos[1]

[1]UC San Diego [2]Intel Labs

The supplemental material is organized as follows: Appendix A provides implementation details of **SViTT**, metadata of all datasets, as well as training setups for pre-training and downstream tasks. Appendix B contains additional ablation studies and qualitative analysis. Discussion of limitations and future work is finally included in Appendix C.

## A. Implementation Details

### A.1. Model Architecture

**Sparse configurations.** The sparsity of **SViTT** is controlled by the following hyperparameters:

- Visual token keep rate $q_v^{(l)}$ and multimodal token keep rate $q_m^{(l)}$ per layer $l$ for *node* sparsity;
- Local attention blocks $K_l$, random attention blocks $K_r$ and block size $G$ shared across layers for *edge* sparsity.

Tab. A1 lists the configurations for each stage of pre-training and the corresponding sparsity $s$, computed as the percent of reduction in edges of sparsified attention graph $\mathcal{G}$ from that of a dense transformer. For the $l^{\text{th}}$ layer of visual encoder $f_v$, the number of edges is given by

$$|\mathcal{E}_v^{(l)}| = N_v^{(l)}(K_l + K_r)G \tag{1}$$

where input length $N_v^{(l)} = \lceil q_v^{(l-1)} N_v^{(l-1)} \rceil$. For multimodal layers $f_m$, the edge count is

$$|\mathcal{E}_m^{(l)}| = N_m^{(l)} N_t \tag{2}$$

where $N_t$ denotes text length and $N_m^{(l)} = \lceil q_m^{(l-1)} N_m^{(l-1)} \rceil$. Therefore an **SViTT** model with $L_v = 12$ visual layers and $L_m = 3$ multimodal layers has overall edge sparsity

$$S(q_v, q_m, K_l, K_r) = 1 - \frac{\sum_{l=1}^{L_v} |\mathcal{E}_v^{(l)}| + \sum_{l=1}^{L_m} |\mathcal{E}_m^{(l)}|}{L_v N_v^2 + L_m N_t N_v} \tag{3}$$

| Frames $T$ | Attn. blocks $K_l, K_r, G$ | Keep rate $q_v, q_m$ | Edges (M) $\lvert\mathcal{E}\rvert$ | Sparsity $S$ |
|---|---|---|---|---|
| 4 | | (0.7, 0.1) | 1.48 | 0.80 |
| 8 | (1, 3, 56) | (0.6, 0.1) | 2.60 | 0.91 |
| 16 | | (0.5, 0.1) | 4.61 | 0.96 |

Table A1. **SViTT Configurations.** We report hyperparameters controlling the edge and node sparsity for different clip lengths $T$, as well as the overall sparsity as computed by Eq. (3).

**Temporal expansion.** Transformer architectures do not require fixed input lengths as its operations are either pointwise (e.g. FFN) or permutation equivariant (e.g. MHSA). This makes the temporal expansion (Sect. 4) of input clips a mostly trivial process, except for the position embeddings, which does depend on spatiotemporal dimensions of inputs. Following prior work on training video transformers with image models, we *inflate* the 2D positional embedding

$$\mathbf{P} = [\mathbf{p}_{\text{cls}}, \mathbf{p}_{1,1}, \ldots, \mathbf{p}_{H,W}] \in \mathbb{R}^{(HW+1)\times d} \tag{4}$$

into a 3D embedding tensor

$$\mathbf{P}' = [\mathbf{p}_{\text{cls}}, \mathbf{p}'_{1,1,1}, \ldots, \mathbf{p}'_{T,H,W}] \in \mathbb{R}^{(THW+1)\times d} \tag{5}$$

for inputs of $T$ frames, by duplicating the local embeddings $\mathbf{p}_{hw}$ along the temporal dimension:

$$\mathbf{p}'_{t,h,w} = \mathbf{p}_{h,w}, \qquad \forall t, h, w \tag{6}$$

Likewise, expansion of clip length from $T_1$ to $T_2$ can be performed by temporally resizing the positional embedding, e.g. through nearest neighbors interpolation:

$$\mathbf{p}'_{t,h,w} = \mathbf{p}_{\lfloor t\cdot\frac{T_1}{T_2}+\frac{1}{2}\rfloor,h,w}, \qquad \forall t, h, w \tag{7}$$

The BEiT backbone of visual encoder uses relative position bias [22] in every self-attention layer, which encodes a scalar added to each entry of the similarity matrix depending on the relative position between query and key patches:

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sigma(\mathbf{Q}\mathbf{K}^T + \mathbf{B})\mathbf{V}, \tag{8}$$

$$\mathbf{B}_{(h,w),(h',w')} = \mathbf{R}_{h'-h,w'-w} \tag{9}$$

| Dataset | Avg. Dur. | # Videos | # Sent. / Q. |
|---|---|---|---|
| *Video-Text Pre-Training* | | | |
| WebVid-2M [2] | 18s | 2.5M | 2.5M |
| *Text-to-Video Retrieval* | | | |
| MSR-VTT [27] | 15s | 10K | 200K |
| DiDeMo [1] | 28s | 10K | 40K |
| Charades [24] | 30s | 10K | 16K |
| SSv2-Label [12] | 4s | 171K | 112K |
| *Video Question Answering* | | | |
| MSRVTT-QA [25] | 15s | 10K | 244K |
| ActivityNet-QA [28] | 180s | 5.8K | 58K |
| AGQA 2.0 [10] | 30s | 10K | 2.27M |

Table A2. **Pre-Training and Downstream Datasets.**

where $\mathbf{R} \in \mathbb{R}^{(2H-1)\times(2W-1)}$ are learnable parameters. When expanding the input to multi-frame clips, we again inflate the relative position bias to the temporal dimension:

$$\mathbf{B}'_{(t,h,w),(t',h',w')} = \mathbf{R}'_{t'-t,h'-h,w'-w}, \tag{10}$$

$$\mathbf{R}' \in \mathbb{R}^{(2T-1)\times(2H-1)\times(2W-1)} \tag{11}$$

$\mathbf{R}'$ is initialized by interpolating $\mathbf{R}$ temporally, identical to the procedure for absolute positional embedding $\mathbf{P}'$.

## A.2. Datasets

**Pre-training.** `SViTT` is pre-trained on WebVid-2M [2] with 2.5 million video-text pairs scraped from the Internet. While alternative datasets exist for video-language pre-training such as HowTo100M [20] and YT-Temporal [30], we choose WebVid as it has higher caption quality, covers a wide range of scenes, and can be trained with a reasonable amount of resource.

**Text-to-video retrieval.** We evaluate text-to-video retrieval on 4 datasets: MSR-VTT [27], DiDeMo [1], Charades [24] and Something-Something v2 [8]. MSR-VTT and DiDeMo are video-text datasets commonly used in prior work; Charades and SSv2 were initially collected for video action recognition, with an emphasis on human-object interactions and temporal modeling, but also includes text descriptions for each video clip.

**Video question answering.** Video question answering is evaluated on MSRVTT-QA [25], ActivityNet-QA [28] and AGQA 2.0 [10], annotated on top of the videos from MSR-VTT [27], ActivityNet [5] and Charades [24] respectively. MSRVTT-QA consists of mostly descriptive questions which can be solved without intricate temporal reasoning. ActivityNet-QA focuses on human actions and spatiotemporal relation between objects, posing a greater challenge beyond frame-based reasoning. AGQA contains difficult questions involving the composition of actions, testing the generalization capacity of video-text models.

Tab. A2 summarizes the statistics of all aforementioned datasets.

## A.3. Training Details

**Pre-training tasks.** `SViTT` is pre-trained on three losses following prior art in VLP [6, 7, 12, 14, 15].

- Video-text contrastive (VTC) applies InfoNCE loss between the video embeddings $\mathbf{Z}_v$ and text embeddings $\mathbf{Z}_t$ extracted at `[cls]` locations of their respective encoder $f_v$ and $f_t$:[1]

$$\mathcal{L}_{\text{VTC}} = \ell_c(\mathbf{Z}_v, \mathbf{Z}_t) + \ell_c(\mathbf{Z}_t, \mathbf{Z}_v), \tag{12}$$

$$\ell_c(\mathbf{X}, \mathbf{Y}) = -\sum_{i=1}^{B} \log \frac{e^{\langle \mathbf{x}_i, \mathbf{y}_i\rangle/\tau}}{\sum_{j=1}^{B} e^{\langle \mathbf{x}_i, \mathbf{y}_j\rangle/\tau}} \tag{13}$$

- Video-text matching (VTM) learns a binary classifier on top of the `[cls]` output of multimodal encoder $f_m$ to discriminate between paired and misaligned video-text pair, optimized by binary cross entropy:

$$\mathcal{L}_{\text{VTM}} = -\sum_{i=1}^{B} \Big( \log(f_m(\mathbf{z}_{v,i}, \mathbf{z}_{t,i})) + \log(1 - f_m(\mathbf{z}_{v,i}, \mathbf{z}_{t,i'}))\Big) \tag{14}$$

where $i' \neq i$ is a randomly selected negative sample.

- Masked language modeling (MLM) requires the multimodal encoder $f_m$ to predict randomly masked out text tokens conditioned on the rest of text and video sequence, through a cross-entropy loss:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i=1}^{B} \sum_{j\in\mathcal{J}} [\mathbf{x}_t]_{i,j}^T \log \mathbf{y}_{i,j} \tag{15}$$

where $[\mathbf{x}_t]_{i,j}$ is a one-hot vector denoting the word at location $j$ of example $i$, $\mathbf{y}_{i,j}$ is the classifier output predicting the word at the same location, and $\mathcal{J}$ is the set of masked indices.

We use equal weights for all three losses.

**Downstream tasks.** We follow the downstream evaluation setup of Singularity [12] for the most part. Text-to-video retrieval is performed by ranking all candidate videos $\mathbf{x}_v$ of the test set by their matching scores to text query $\mathbf{x}_t$. For video QA, a transformer decoder is applied on top of multimodal encoder $f_m$ to generate the answer.

[1] Linear projection on top of $\mathbf{z}_v, \mathbf{z}_t$ omitted.

| Method | PT | Frames | Sparsity | MSR-VTT R1 | MSR-VTT R5 | MSR-VTT R10 | MSR-VTT Mean | DiDeMo R1 | DiDeMo R5 | DiDeMo R10 | DiDeMo Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VideoCLIP [26] | 100M | — | — | 10.4 | 22.2 | 30.0 | **20.9** | 16.6 | 46.9 | — | **—** |
| Frozen [2] | 5M | 4 | | 23.2 | 44.6 | 56.6 | **41.5** | 21.1 | 46.0 | 56.2 | **41.1** |
| ALPRO [13] | 5M | 8 | | 24.1 | 44.7 | 55.4 | **41.4** | 23.8 | 47.3 | 57.9 | **43.0** |
| VIOLET [7] | 5M | 4 | | 25.9 | 49.5 | 59.7 | **45.0** | 23.5 | 49.8 | 59.8 | **44.4** |
| Singularity [12] | 5M | 1 | | 28.4 | 50.2 | 59.5 | **46.0** | 36.9 | 61.1 | 69.3 | **55.8** |
| Singularity* | 2M | 1 | — | 21.1 | 42.1 | 53.0 | **38.7** | 23.3 | 45.4 | 53.7 | **40.8** |
| | | 4 | | 24.4 | 43.8 | 51.7 | **40.0** | 26.4 | 48.7 | 57.3 | **44.1** |
| | | 8 | | 24.3 | 44.5 | 54.3 | **41.0** | 25.8 | 50.0 | 60.7 | **45.5** |
| **SViTT** | 2M | 8 | Dense | 26.0 | 47.7 | 57.1 | **43.6** | 29.6 | 54.1 | 64.1 | **49.3** |
| | | | Hybrid | 25.4 | 48.4 | 57.5 | **43.8** | 31.0 | 57.2 | 66.3 | **51.5** |

Table A3. **Zero-shot Text-to-video Retrieval.** Results reported in prior works marked in gray; * indicates our reproduced results.

| Task | Pre-training | | | Video-text Retrieval | | | Video QA |
|---|---|---|---|---|---|---|---|
| **Frames** $T$ | 4 | 8 | 16 | 4 | 8 | 16 | 8 |
| **Epochs** | | 10 | | | 15 | | 5 (1 for AGQA) |
| **Warm-up** | | 1 | | | 0 | | 0 |
| **Batch size** | 512 | 336 | 192 | 64 | 48 | 32 | 128 |
| **Learning rate** | $3 \times 10^{-5}$ | $1 \times 10^{-5}$ | $5 \times 10^{-6}$ | | $1 \times 10^{-5}$ | | $5 \times 10^{-5}$ |
| **Weight decay** | | 0.02 | | | 0.02 | | 0.02 |
| **Text length** | | 32 | | | 32 (64 for DiDeMo) | | 25 (Q), 5 (A) |
| **Attn. blocks** $(K_l, K_r, G)$ | | $(1, 3, 56)$ | | | $(1, 3, 56)$ | | $(1, 3, 56)$ |
| **Keep rate** $(q_v, q_m)$ | $(0.7, 0.1)$ | $(0.6, 0.1)$ | $(0.5, 0.1)$ | $(0.7, 0.1)$ | $(0.6, 0.1)$ | $(0.5, 0.1)$ | $(0.6, 0.5)$ |

Table A4. **Training Hyperparameters.**

**Training hyper-parameters.** We use a sparse frame sampling strategy following [2,7,12], splitting input videos into $T$ chunks and randomly selecting one frame from each during training. Video frames are preprocessed with random resized cropping into spatial resolution of $224 \times 224$, resulting in $14 \times 14$ spatial patches. All models are optimized using AdamW [17] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a cosine learning rate schedule and warm-up training. We use 10 epochs for pre-training and 15 for fine-tuning on all datasets other than AGQA, which uses 1 epoch due to its large size. Batch size $B$ and learning rate $\eta$ are adjusted depending on memory costs of sparse models. Tab. A4 summarizes the hyperparameters used for each task and model variant.

# B. Additional Results & Analysis

## B.1. Retrieval Metrics

We include full retrieval results with Recall@$\{1, 5, 10\}$ in Tab. A3 (zero-shot) and Tab. A5 (fine-tuned).

## B.2. Video-Text Backbone

In addition to the Singularity baseline with BEiT-B backbone used in the main paper, we also evaluate **SViTT** on a simpler structure from Frozen [2]. This is also a two-tower model with separate video and text encoders $f_v, f_t$, but unlike most vision-language transformers, does not contain a cross-modal encoder on top. Frozen is trained solely on the InfoNCE loss between video and text embeddings, and uses their cosine similarity to perform retrieval. While the cross-modal node sparsification does not apply to this framework, visual node sparsity and edge sparsity can still be applied to the visual encoder $f_v$ to enable temporal learning across frames.

The original Frozen model uses a divided space-time attention similar to TimeSformer [4], where temporal attention is added to a pre-trained ViT and initialized as identity mapping. During early experiments, however, we find that the temporal module with zero-init fails to learn meaningful attention across frames, with query and key matrices stuck at zero weights. We opted to remove the temporal attention modules and make the spatial attention global instead (i.e. each token attends to every token from the video clip, instead of just those from the same frame).

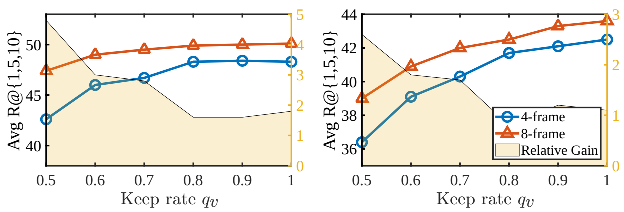
Tab. A6 shows the performance of **SViTT** applied to the Frozen model. Similar to the results in the main paper, our dense spatiotemporal transformer with the above modifica-

| Method | PT | Frames | Sparsity | Charades | | | | SSv2-Label | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | R1 | R5 | R10 | **Mean** | R1 | R5 | R10 | **Mean** |
| Frozen [2] | 5M | 32 | | 11.9 | 28.3 | 35.1 | **25.1** | | — | | |
| CLIP4Clip [18] | 400M | 12 | | 13.9 | 30.4 | 37.1 | **27.1** | 43.1 | 71.4 | 80.7 | **65.1** |
| ECLIPSE [16] | 400M | 32 | | 15.7 | 32.9 | 42.4 | **30.3** | | — | | |
| MKTVR† [19] | 400M | 42 | — | 16.6 | 37.5 | 50.0 | **34.7** | | — | | |
| Singularity [12] | 5M | 1 | | | — | | | 36.4 | 64.9 | 75.4 | **58.9** |
| | | 4 | | | — | | | 44.1 | 73.5 | 82.2 | **66.6** |
| **SViTT** | 2M | 8 | Dense | 16.0 | 34.9 | 47.2 | **32.7** | 43.6 | 72.6 | 82.2 | **66.1** |
| | | | Hybrid | 17.7 | 39.5 | 49.8 | **35.7** | 47.5 | 76.3 | 84.2 | **69.3** |

Table A5. **Text-to-video Retrieval with Fine-tuning.** † denotes concurrent work.

| Method | | PT | Frames | DiDeMo | | | |
|---|---|---|---|---|---|---|---|
| | | | | R1 | R5 | R10 | **Mean** |
| Frozen [2] | | 5M | 4 | 21.1 | 46.0 | 56.2 | **41.1** |
| **SViTT** | Dense | 2M | 8 | 21.9 | 45.6 | 56.6 | **41.4** |
| | Hybrid | | | 22.9 | 47.7 | 58.1 | **42.9** |

Table A6. **Zero-shot Retrieval with SViTT on Frozen Baseline.**




(a) DiDeMo zero-shot. (b) MSR-VTT zero-shot.

Figure A1. **Node Sparsity for 4- and 8-frame Models.** Model of longer clip length is more robust to node sparsification.

tions outperformed the original implementation of [2], despite being trained without image-text data (CC3M [23]). **SViTT** with hybrid sparsity again outperforms the dense version while using less computation and training memory.

## B.3. Video Sparsity vs. Clip Length

To demonstrate the claim that video sparsity increases with clip length, we evaluate dense models trained with clip length 4 and 8 under different levels of sparsity. As shown in Fig. A1, the 8-frame model is more robust to token pruning with lower keep rates. On DiDeMo, it outperforms 4-frame model by 4% at $q_v = 0.5$, while the two models differ by under 2% under dense evaluation. This reveals that longer clips contain greater level of redundancy, and should be modeled with higher sparsity (as done in this work).

## B.4. Chunking Strategy

In edge sparsification, the flattened video sequence $\mathbf{z}_{1:N}$ is chunked into subsequences of length $G$. While this strategy is straightforward and common in language transformers [3, 29], it breaks the spatiotemporal continuity of video data. We investigate an alternative to naïve chunking, by reordering the input tokens using space-filling curves such as Morton [21] and Hilbert [11] curves. This ensures that neighboring tokens in the flattened sequence are close to each other in the original multidimensional space, leading to more localized chunks.

| Order | MSR-VTT | | | | DiDeMo | | | |
|---|---|---|---|---|---|---|---|---|
| | R1 | R5 | R10 | **Mean** | R1 | R5 | R10 | **Mean** |
| Standard | 21.0 | 43.0 | 51.5 | **38.5** | 29.1 | 53.5 | 63.1 | **48.6** |
| Morton | 20.6 | 40.6 | 49.4 | **36.9** | 27.3 | 51.9 | 61.9 | **47.1** |
| Hilbert | 20.3 | 40.9 | 49.6 | **36.9** | 27.9 | 52.5 | 62.5 | **47.7** |

Table A7. **Ablation on Token Ordering.** We compare the standard **SViTT** trained with flattened video tokens and reordering using space-filling curves.

| Model | Sp. | SSv2-Label | | | ActivityNet-QA | | |
|---|---|---|---|---|---|---|---|
| | | N | S | Δ | N | S | Δ |
| Singularity [12] | — | 66.6 | 66.3 | **0.3** | 41.8 | 41.8 | **0.0** |
| **SViTT** | D | 66.1 | 64.9 | **1.2** | 42.5 | 42.3 | **0.2** |
| | H | 69.3 | 65.8 | **3.5** | 43.2 | 42.3 | **0.9** |

Table A8. **Temporal Probing.** Video-text transformers are evaluated using Normal and Shuffled frame order.

However, early experiments showed no benefit of space-filling token order over naïve flattening, as shown in Tab. A7. This is possibly because video encoders are initialized from image transformers, and block attention with reordering prevents video tokens from attending to other spatial locations from the same frame. We leave the study of an optimal chunking strategy for 3D inputs for future work.

## B.5. Temporal Probing

To measure the sensitivity of the learned video-text model to temporal cues, we perform an evaluation with shuffled input frames. Tab. A8 shows a performance drop of **SViTT** models on retrieval (SSv2) and video QA (ActivityNet) tasks, indicating that the video-text models have learned to reason about the temporal dynamics of video clips. The difference $\Delta$ between normal and shuffled inputs is more prominent on hybrid sparse models, possibly because they attend more to the foreground which contains more temporal variations. Notably, this behavior does not hold for the Singularity model, whose performance is unaffected by frame order. This suggests that late temporal aggregation after spatial global pooing is insufficient to capture spatiotemporal relations across video frames.

### B.6. Qualitative Results

Figs. A2 and A3 visualizes the node sparsification patterns generated by visual encoder $f_v$ and multimodal encoder $f_m$. While visual sparsification alone can significantly reduce the number of tokens during forward pass, we find that the cross-modal attention map aligns better with regions of interest in each clip, enabling greater node sparsity in video-text modeling.

## C. Limitations & Future Work

While **SViTT** shows great potential towards building long-term video-text models, we recognize that learning temporal relationships from videos would not be possible without high-quality pre-training data. We find that WebVid-2M exists a strong tendency towards spatial appearances: Many videos consist of only simple motions (running, talking etc.), and captions are often highly correlated to the static background. Given this, we suspect that further increasing the clip length beyond 16 frames per video is unlikely to make a significant difference in modeling performance. Building on top of the sparse video-text architecture in this work, future studies can focus on pre-training on video-language datasets and tasks that require aggregating information over a longer period of time span, e.g. narrated egocentric videos over long episodes [9], where **SViTT** may provide larger gains over frame-based approaches and dense spatiotemporal transformers.

## References


[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 2, 3, 4

[3] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 3

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[6] 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 *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[7] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2, 3

[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 *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 5

[10] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa 2.0: An updated benchmark for compositional spatio-temporal reasoning. *arXiv preprint arXiv:2204.06105*, 2022. 2

[11] D Hilbert. Über die stetige abbildung einer linie aufein flächenstück. *Mathematische Annalen*, 38:459–460, 1891. 4

[12] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 2, 3, 4

[13] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 3

[14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[15] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 2

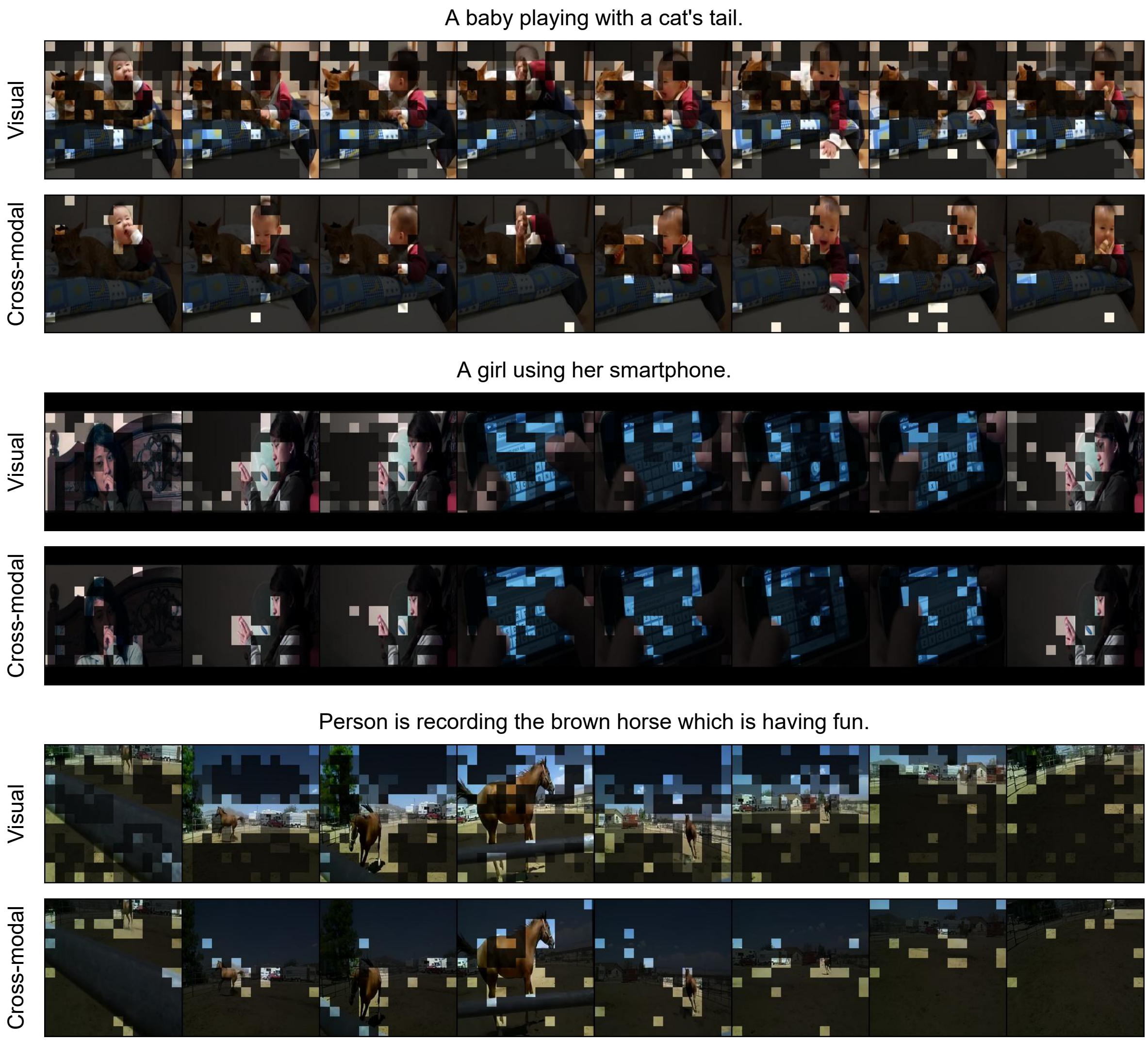



Figure A2. **Qualitative Results.** We visualize node sparsity patterns generated by visual ($q_v = 0.6$) and cross-modal encoder ($q_m = 0.1$).


[16] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. *arXiv preprint arXiv:2204.02874*, 2022. 4

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[18] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 4

[19] Avinash Madasu, Estelle Aflalo, Gabriel Ben Melech Stan, Shao-Yen Tseng, Gedas Bertasius, and Vasudev Lal. Improving video retrieval using multilingual knowledge transfer. *arXiv preprint arXiv:2208.11553*, 2022. 4

[20] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[21] Guy M Morton. A computer oriented geodetic data base and a new technique in file sequencing. 1966. 4

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 1

[23] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 4

[24] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in

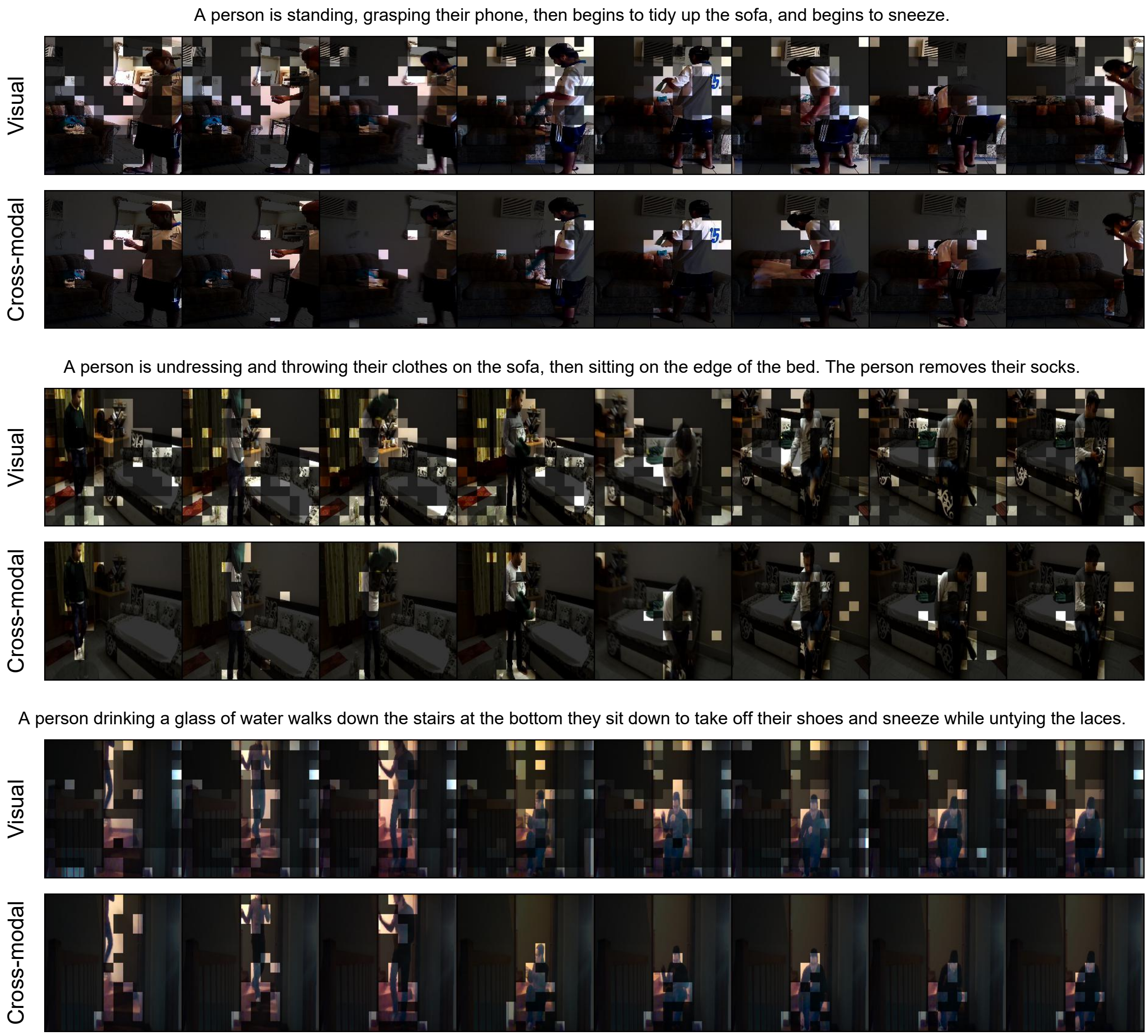



Figure A3. **Qualitative Results (continued).**

homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2

[25] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 2

[26] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[27] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[28] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 2

[29] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020. 4

[30] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 2