Text-Guided Video Masked Autoencoder

David Fan[1], Jue Wang[1], Shuai Liao[2], Zhikang Zhang[1], Vimal Bhat[1], and Xinyu Li[1]

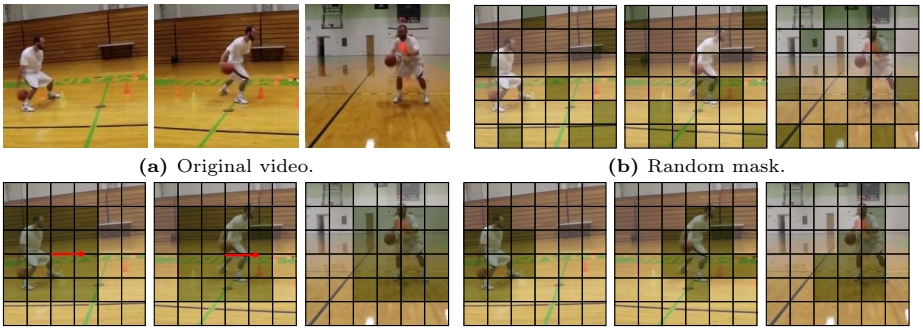
[1] Amazon Prime Video
[2] Amazon Fulfillment Technology

Abstract. Recent video masked autoencoder (MAE) works have designed improved masking algorithms focused on saliency. These works leverage visual cues such as motion to mask the most salient regions. However, the robustness of such visual cues depends on how often input videos match underlying assumptions. On the other hand, natural language description is an information dense representation of video that implicitly captures saliency without requiring modality-specific assumptions, and has not been explored yet for video MAE. To this end, we introduce a novel text-guided masking algorithm (TGM) that masks the video regions with highest correspondence to paired captions. Without leveraging any explicit visual cues for saliency, our TGM is competitive with state-of-the-art masking algorithms such as motion-guided masking. To further benefit from the semantics of natural language for masked reconstruction, we next introduce a unified framework for joint MAE and masked video-text contrastive learning. We show that across existing masking algorithms, unifying MAE and masked video-text contrastive learning improves downstream performance compared to pure MAE on a variety of video recognition tasks, especially for linear probe. Within this unified framework, our TGM achieves the best relative performance on five action recognition and one egocentric datasets, highlighting the complementary nature of natural language for masked video modeling.

1 Introduction

The success of masked language modeling [8, 29] has recently inspired the adoption of the masked autoencoder (MAE) for masked image and video modeling. Masking out random image patches [18] and reconstructing the missing image patches via an asymmetric encoder-decoder architecture achieves promising results in image recognition. In a similar fashion, works such as VideoMAE [39] and ST-MAE [15] achieve promising results in video recognition by extending random masking from 2D image patches to 3D video cubes.

These initial works demonstrate the strong potential of masked visual modeling. Subsequent works have further explored the question of "where to mask?" While simple and effective, random masking assumes that input information density is uniformly distributed. Several new masking strategies have been proposed which challenge this assumption and attempt to directly mask visual saliency. For



(a) Original video. **(b)** Random mask. **(c)** Motion-guided mask. **(d)** Text-guided mask. "A man dribbling a basketball"

Fig. 1: Illustration of different masking strategies. 1b: Random masking [15, 39] randomly masks patches independently of their contents. 1c: Motion-guided masking [11, 20] tracks the motion of patches over time to mask a moving volume. 1d: Our proposed text-guided masking masks the top video patch-to-text correspondence.

instance, in image domain, SemMAE [24] decomposes foreground into coarse semantic parts and then masks visual patches from each semantic segment according to a defined sampling probability. AutoMAE [3] utilizes adversarial training with bounding boxes to learn an object-centric mask. Both of these works intend to mask the foreground more often than background. These masking algorithms achieve better performance on image recognition benchmarks than the random masking baseline. However, there exists a trade-off point: masking information dense regions too aggressively degrades performance on downstream tasks [24].

In video domain, MGM [11] and MGMAE [20] leverage motion as a video-specific prior for saliency. Specifically, they mask the patches with highest motion over time, where motion is obtained either from motion vectors in the video codec [11] or optical flow [20]. Motion-guided masking achieves better performance than random masking in video domain, suggesting that masking saliency is also important for masked video modeling.

These previous works exploit various visual priors with the goal of masking objects and motion. However, their robustness depends on how often input videos match the underlying statistical assumptions. For instance, not all videos have higher foreground motion than background motion. On the other hand, natural language captions are an information dense representation of video that describe both "nouns" (e.g. humans and objects) and "verbs" (e.g. actions), without the need to make any prior assumptions. Provided a well-aligned vision-language model, it is possible to directly mask salient regions according to the input text. This makes natural language a promising source of saliency for masked video modeling that has not yet been explored. Thus, this work explores a new direction of improving masked video modeling with natural language, and presents a strong baseline composed a novel masking algorithm and additional loss.

First, we discard visual priors used in previous works and ask whether the content within captions can already capture the most salient regions of video.

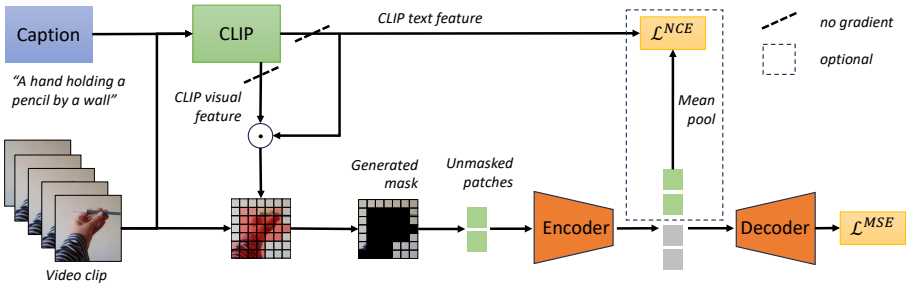


Fig. 2: For each video, we generate a caption using an off-shelf image captioning model such as BLIP [26]. We then leverage the aligned representation space of CLIP [36] to mask the patches with highest correspondence to the text. The MAE pipeline is identical to VideoMAE [39], where the encoder processes the visible patches and the decoder processes the union of encoded visible patches and mask tokens. We additionally introduce an *optional* contrastive loss to align the encoded visible patches with the text. This facilitates semantic-aware reconstruction. BLIP and CLIP receive no gradients.

To that end, we first introduce a novel text-guided masking (TGM) algorithm, which masks the video regions with highest correspondence to a given caption that is either machine-generated or human-annotated. We present the interesting insight that text-guided masking is competitive with state-of-the-art motion-guided masking, despite using no explicit motion guidance. This confirms the intuition that captions can capture video saliency without prior assumptions.

Second, to further leverage the semantics of natural language for masked video modeling, we introduce a general unified framework for MAE and masked video-text contrastive learning. For any given masking algorithm, introducing an optional contrastive loss aligns the masked encoder representation with the text. To our knowledge, we are the first to unify the generative nature of MAE pretraining and discriminative nature of masked contrastive learning for video and obtain benefits from both pretraining paradigms.

Training from scratch without any bells and whistles, our text-guided masking outperforms MGM by up to 1.3% on Kinetics-400 (K400) and 0.5% on Something-Something V2 (SSv2) in finetuning performance, and by up to 1.7% in linear evaluation. We show that text-guided masking generalizes to smaller action recognition datasets as well as egocentric action recognition. Lastly, we demonstrate that the synergistic nature of masked contrastive learning also applies to other masking algorithms such as random and motion-guided masking.

In summary, our contributions are:

1. Text-guided masking (TGM) – a simple yet effective masking algorithm.
2. Unifying masked video modeling and video-text contrastive learning.
3. Empirical evidence for the synergistic nature of masked contrastive learning and masked video modeling on five action recognition datasets and one egocentric understanding dataset.
4. Introducing a new area of research into language-guided masked video modeling and detailed insights that will help inspire future work.

2 Related Work

2.1 Masked Image Modeling

Early work such as iGPT [4] performs masked image modeling at the pixel-level. BEIT [1] elevates the reconstruction target from individual pixels to pre-trained dVAE tokens. MAE [18] reconstructs normalized pixel patches instead and demonstrates the efficacy of an asymmetric encoder-decoder design with high masking ratio. Recent work addresses the question of whether random masking is ideal. SemMAE [24] uses attention maps to obtain a coarse segmentation and masks patches based on a sampling distribution defined over each semantic class. AutoMAE [3] utilizes adversarial training with bounding boxes to learn an object-centric mask. Both SemMAE and AutoMAE find that masking higher proportion of foreground tokens improves image representations, up until a certain point where masking too aggressively degrades performance.

In contrast to these works, our work is designed for video and leverages textual information as the primary proxy for video saliency.

2.2 Masked Video Modeling

Some works use tokenization-based reconstruction targets. VIMPAC and BEVT use pretrained VQ-VAE [40] tokens. However, these works require extra pre-training. MaskFeat instead uses HoG features. VideoMAE [39] and ST-MAE [15] directly reconstruct randomly masked 3D video patches, achieving promising results on video benchmarks. Recent work explores whether motion-based priors can lead to improved masking algorithms for video. MGM [11] and MGMAE [20] mask the video patches with highest motion, under the assumption that higher motion co-occurs with higher saliency. These motion-guided masks enhance video representations compared to random masking. In contrast to these works, our work leverages textual information as a guide for where to mask, and also introduces a masked video-text contrastive loss that has not been explored yet by video MAE works to the best of our knowledge.

2.3 Contrastive Visual Pretraining

MOCO [6, 19] and SimCLR [5] introduce contrastive learning as an image representation learning paradigm. The encoder is trained by forcing invariance between the encoded representation for two views of the same image, which are typically generated through image augmentations. In video domain, CVRL [35] and ρ-MoCo [14] extend contrastive learning to video domain by sampling two subclips from the same video and applying video augmentation. BRAVE [37] and LSTCL [41] sample overlapping short and long clips to enforce temporal correspondence. These works all use intra-instance positive samples which are generated from the same image/video either through augmentation or re-sampling.

Other works go beyond intra-instance positives to explore inter-instance positive pair sampling. NNCLR [10] uses nearest-neighbor images as positive samples

to the anchor. Similarly, IIVCL [12] uses multiple nearest-neighbor videos as the positive samples to the anchor.

In contrast to these works, we do video-text contrastive learning rather than visual contrastive learning, and we apply masking on top of the video.

2.4 Vision-Language Pretraining

CLIP [36] and related works such as ALIGN [21] popularized image-to-text contrastive learning on hundreds of millions of image-text pairs. FLIP [27] scales CLIP to higher throughput by introducing image masking. A second line of works such as CoCa [46] and Florence [47] have explored captioning as a vision-language pretraining task. CoCa additionally introduces an optional image-text contrastive loss. Other works apply masked modeling on both image and text. For example, M3AE [16] combines image patches and text tokens as the input to a unified masked autoencoder. In contrast to these works, we unify video MAE and masked video-text contrastive learning, do not use captioning, and do not apply masking to text. We also pretrain from scratch on only ∼200K videos. We additionally propose our novel text-guided masking strategy.

Another line of work attempts to recreate the success of CLIP for video-text pretraining. CLIP4CLIP [31] takes pretrained CLIP and applies it frame-wise to video and explores different temporal aggregation strategies such as mean pooling and Transformer encoder to achieve a video-level embedding. CLIP4CLIP only explores retrieval tasks. ViCLIP [44] upgrades the ViT image encoder from CLIP with spatiotemporal attention to make it a video encoder, and trains on a self-curated high-quality dataset with 200M video-text pairs. InternVideo [45] alternates between video MAE pretraining and video-text contrastive learning using a different visual backbone for the contrastive learning. In contrast, our approach jointly optimizes the video MAE and video-text contrastive loss with the same visual backbone and the contrastive loss operates on the masked MAE encoder output. Our approach does not use any cross-attention layers. We additionally introduce our novel text-guided masking algorithm.

3 Method

3.1 Revisiting Video Masked Autoencoders

Given a video $V \in \mathbb{R}^{T \times H \times W \times C}$, where T, H, W, C denote the number of frames, height, width, and RGB-channels, the video is typically first split into cubes of size $t \times h \times w \times C$ and processed with a patch embedding layer $\mathcal{P}$ to obtain a sequence of cube embeddings V_p. Typically $t = 2$ and $h = w = 16$.

$$V_p = \mathcal{P}(V); V_p \in \mathbb{R}^{\frac{T}{t} \times \frac{H}{h} \times \frac{W}{w} \times D} \tag{1}$$

Next, a masking function η (e.g. random [15], tube [39], or motion-guided [11, 20]) generates a binary mask M to select a set of visible patches with mask ratio γ.

$$\begin{aligned} M &= \eta(V_p, \gamma) \\ V_{\text{p\_visible}} &= V_p \odot (\sim M) \\ V_{\text{p\_masked}} &= V_p \odot M \end{aligned} \tag{2}$$

The encoder ϕ then processes only the visible patches $V_{\text{p\_visible}}$ while the decoder ξ processes the full set of encoded patches and masked tokens $\phi(V_{\text{p\_visible}}) \cup V_{\text{p\_masked}}$ to reconstruct the video. This work uses the same asymmetric encoder-decoder design as [15, 18, 39].

$$E = \phi(V_{\text{p\_visible}}), \quad V' = \xi(E \cup V_{\text{p\_masked}}) \tag{3}$$

Finally, the model is trained with the MSE reconstruction loss $\mathcal{L}_{\text{MSE}}$, which is computed between V and V'. In this work, we propose a novel mask generator $\eta(\gamma)$ which is guided by text.

3.2 Caption Generation

Because Kinetics-400 and Something-Something v2 do not have human annotated captions, we utilize BLIP-2 [25, 26] offline to generate video-text pairs for pretraining. For each pretraining video, we uniformly sample 3 keyframes and inference a caption per frame, for a total of 3 captions per video. During pretraining, we randomly sample from these 3 captions per video to form the text-video pair. Note that the captioning model is only used offline to obtain captions and does not receive any gradients during training. Note that using off-shelf captioning models for video is convenient but leads to noisier captions than human annotation. Other caption sources are ablated in Table 6b.

3.3 Text-Guided Masking

We leverage the aligned representation space of CLIP [36] to compute a text-guided mask. First, we compute the visual features. For each frame f_t, we compute the feature map $V_t \in \mathbb{R}^{\frac{H}{h} \times \frac{W}{w} \times D}$ using ViT-B/32. This is done by patchifying each frame and resizing each patch to full input resolution. We then take the CLS token as the embedding per patch. To get the text embedding $w \in \mathbb{R}^D$, we follow CLIP and take the activation map from the last layer of the transformer at the `[EOS]` token. We then compute the cosine similarity between V_t and w and take the top k patches by text-video cosine similarity to form the binary mask $\mathcal{M}_t$, where $k = \frac{H}{h} \cdot \frac{W}{w} \cdot \gamma$ and $\sum \mathcal{M}_t / \ell(\mathcal{M}_t) = \gamma$ to satisfy the masking ratio γ. We then apply $\mathcal{M}_t$ per frame f_t to obtain the visible $V_{\text{p\_visible}}$ and masked $V_{\text{p\_masked}}$.

3.4 Video-Text Alignment

The video-text contrastive loss is a standalone module that can be optionally applied on top of the MAE pipeline for improved performance. The MAE encoder ϕ already processes only the visible patches $V_{\text{p\_visible}}$, so no additional

computation from the encoder is required for the video-text contrastive loss, as shown in Figure 2. Let i index the mini-batch. The global video embedding v_i for video i is computed by mean pooling $\phi(V_{\text{p\_visible}})$ across all patches. We then compute $(\mathcal{L}^{\text{NCE}}(v_i, t_i, t_{j\neq i}) + \mathcal{L}^{\text{NCE}}(t_i, v_i, v_{j\neq i}))/2$ over the mini-batch of size N, where the negative samples are all other text embeddings $t_{j\neq i}$ and video embeddings $v_{j\neq i}$ respectively. Similar to SimCLR [5,6], we use a prediction head with global batch norm. We note that one convenience of masking is that we get large batch size by design, which is beneficial for contrastive learning.

The InfoNCE loss [33] $\mathcal{L}^{\text{NCE}}$ maximizes the similarity of a given sample q with its positive key k^+, while minimizing similarity to negative samples $\mathcal{N}^-$:

$$\mathcal{L}^{\text{NCE}}(q, k^+, \mathcal{N}^-) = -\log \frac{\exp(sim(q, k^+)/\tau)}{\sum_{k\in\{k^+\}\cup\mathcal{N}^-} \exp(sim(q, k)/\tau)} \tag{4}$$

where $\tau > 0$ is a temperature hyper-parameter and $sim(\cdot)$ denotes the similarity function — which in this work is the dot product (cosine) similarity between two ℓ_2 normalized vectors: $sim(q, k) = q \cdot k = q^T k/(\|q\|\|k\|)$.

3.5 Text-Guided MAE

The final loss is either $\mathcal{L}_{\text{MSE}}$ in the case of pure MAE, or $\mathcal{L}_{\text{MSE}} + \mathcal{L}^{\text{NCE}}$ when MAE and video-text contrastive loss are combined. After this self-supervised pretraining, the model is then transferred to downstream tasks such as classification via finetuning and linear probe with cross-entropy loss.

4 Results

4.1 Datasets

We conduct experiments on six commonly used datasets:
Something-Something V2 (SSv2) [17] contains 220K videos with 174 action classes. SSv2 is considered a motion heavy dataset, as most of the labels are defined by the motion and directionality of the actual action. **Kinetics-400 (K400)** [22] is the de-facto standard dataset used to evaluate video recognition. It contains 240K Internet videos with 400 action classes. **UCF101** [38] is a dataset containing 13K Internet short videos with 101 action classes. **HMDB51** [23] is a dataset containing 5K short movie clips from 51 action classes. **Diving48** [28] contains 18K untrimmed video clips from 48 action classes, all of which are types of dives. We report the top-1 accuracy on the evaluation set for all datasets following standard practices [13]. Only UCF101 and HMDB51 have multiple split versions; we use split 1. **Epic-Kitchens 55** [7] contains around 30K egocentric first-person video clips from nearly 3K action classes. Egocentric videos feature heavy occlusion, camera motion, and jitter.

			SSv2			K400		
Mask	**Backbone**	**Epochs**	**Pretrain**	**FT**	**LP**	**Pretrain**	**FT**	**LP**
Tube§ [39]	ViT-B	200	SSv2	66.6	25.7	K400	78.4	38.1
MGM§ [11]	ViT-B	200	SSv2	67.3	33.0	K400	79.9	32.1
TGM	ViT-B	200	SSv2	67.1	26.2	K400	79.9	33.8

Table 1: Comparison between our text-guided masking to (random) tube and motion-guided masking in pure MAE on Something-Something v2 (SSv2) and Kinetics-400 (K400). FT = finetune, LP = linear probe. § = our reproduction.

4.2 Implementation Details

Model Configuration: The default backbone is ViT-Base [9] with global joint space-time attention. For fair comparison, we use the same input patch size of $2 \times 16 \times 16$ for all models following [39].

Pre-Processing: We pretrain with clips of 16 frames sampled at a temporal stride of 4 for K400 and stride of 2 for SSv2 respectively following [39]. We use a fixed spatial resolution of 224×224 for all experiments. We apply multi-scale-crop and horizontal flip augmentation by default (flip is not applied to SSv2). We follow [39] to use AdamW [30] optimizer with a base learning rate $1.5e-4$, weight decay of 0.05, $\beta = [0.9, 0.95]$, and cosine learning rate decay.

Finetuning and Linear Probe: We use the same 16-frame clip for finetuning and multi-view evaluation protocol following standard practice [13]. We use TSN-style sampling [42, 43] on SSv2 dataset with 2 temporal $\times$ 3 spatial views during test-time following [39] for fair comparison. For Kinetics-400, UCF101, HMDB51, Diving48, and Epic-Kitchens 55, we use 5 temporal $\times$ 3 spatial views during test-time following [39] for fair comparison. See the supplementary material for hyperparameter details which are mostly the same as [39].

4.3 Text-Guided Masking

We first apply our TGM to pure MAE (no contrastive learning) on SSv2 and K400 in Table 1 and compare to random tube masking [39] and motion-guided masking [11, 20]. We chose MGM [11] over MGMAE [20] due to the higher scalability of motion vectors than optical flow.

On SSv2, our TGM achieves better finetune and linear probe performance than tube masking. On K400, our TGM achieves better finetune performance than both motion-guided masking and tube masking, and better linear probe performance than motion-guided masking. We do not claim state-of-the-art results, but instead emphasize the surprising and useful insight that text-guided masking is already competitive with other state-of-the-art masking algorithms — without leveraging any explicit visual cues for saliency such as motion vectors.

	Components		SSv2	
Mask	**MAE**	**V→T**	**FT**	**Linear**
$\text{Tube}_{p=0.75}$		✓	44.9	12.9
$\text{Tube}_{p=0.75}$	✓		64.9	20.8
$\text{Tube}_{p=0.75}$	✓	✓	65.5	33.3
$\text{MGM}_{p=0.75}$		✓	47.2	6.6
$\text{MGM}_{p=0.75}$	✓		67.3	33.0
$\text{MGM}_{p=0.75}$	✓	✓	67.0	**37.1**
$\text{TGM}_{p=0.6}$	✓		67.1	26.2
$\text{TGM}_{p=0.6}$	✓	✓	**67.5**	33.4

Table 2: Systematic performance breakdown of pure MAE, pure masked video-text contrastive loss, and unified MAE + video-text contrastive loss on Something-Something v2 (SSv2). FT = finetune, LP = linear probe.

4.4 MAE with Masked Video-Text Contrastive Learning

Next, we introduce the masked video-text contrastive loss as a mask-agnostic module that can be used both by itself, or optionally combined with MAE. We evaluate downstream finetune and linear probe performance for three masking algorithms on SSv2 in Table 2. Recall that previous video MAE works do not explore masked video-text contrastive learning, so our results and insights for random tube masking and motion-guided masking are new.

Comparing within same masking algorithm. In this section, we focus on each masking algorithm individually for an apple-to-apple comparison. First, note that pure masked video-text contrastive learning does not achieve competitive results for any masking algorithm. For instance, with random tube masking, pure masked video-text contrastive learning trails pure MAE with random tube masking by over 20% in finetune and 8% in linear probe performance. Second, we note that when MAE and contrastive loss are combined, linear probe performance improves notably compared to pure MAE with the same masking algorithm. For example, this boost is 12.5% for tube masking, 4.1% for motion-guided masking, and 7.2% for our TGM. Finetune performance also improves by 0.6% for tube masking and by 0.4% for TGM. In the case of motion-guided masking, a small drop of 0.3% in finetune performance is counterbalanced by a 4.1% improvement in linear probe performance. Overall, the results indicate that the masked video-text contrastive algorithm is synergistic with MAE pretraining, and that this benefit is general across multiple masking algorithms.

Comparing across different masking algorithms. With the combined MAE and contrastive loss, we see that TGM achieves the highest finetune performance (+0.5% over MGM and +2.0% over tube masking), while MGM achieves the highest linear probe performance. However, TGM still achieves a reasonable trade-off which is competitive with both tube and motion-guided masking. Thus, in subsequent experiments, we primarily focus on our TGM to evaluate its generalizability to other downstream tasks and yield additional insights.

		UCF	HMDB	Diving48	UCF		HMDB		Diving48	
Mask	**V → T**	**Linear Probe**			**R@{1,5}**					
Tube [39]		70.1	46.1	11.1	89.2	94.5	62.0	78.1	15.0	42.1
MGM [11]		70.9	47.0	10.1	85.1	92.2	56.8	73.6	14.9	39.3
TGM		67.7	41.6	11.3	85.1	91.7	57.8	73.9	13.8	39.9
TGM	✓	87.1	64.3	19.9	97.6	99.1	75.7	87.4	18.0	44.8

Table 3: TGM generalizes to downstream datasets in linear probe and zero-shot retrieval settings. All results are pretrained for 200 epochs on K400.

4.5 Transfer Learning

Small Action Recognition Datasets. We next evaluate TGM pretrained on K400 when transferred to smaller action recognition datasets: UCF101 [38] (13K videos), HDMB51 [23] (5K videos), and Diving48 [28] (18K videos). These datasets range greatly in content diversity; UCF101 contains Internet videos, HDMB51 contains cinematic clips, and Diving48 contains sports videos. Motivated by the poor linear probe performance of previous MAE works in both image and video domain [11, 18, 39], we focus on evaluation methods that do not finetune the backbone. Specifically, we use both linear evaluation and zero-shot retrieval to see whether TGM has learned sufficient semantics to generalize even when limited labeled data is available. First, we compare TGM without contrastive learning to tube masking and motion-guided masking, and see that TGM is competitive in both the linear probe and zero-shot retrieval settings. When combined with contrastive learning, TGM achieves a notable performance boost of up to 22.7% in linear probe over TGM without contrastive learning, as well as up to a 18% boost in recall@1. Thus, we see that the findings from Table 1 and Table 2 still hold true in this small dataset setting.

Egocentric Action Recognition. To further challenge TGM, we shift to a new task of egocentric action recognition which is uniquely challenging due to featuring first-person perspectives. Egocentric video contains high occlusion, camera motion, and jitter, so it is an interesting setting for evaluating whether our text-guided masking can still be competitive with motion-guided masking — despite not leveraging any visual cues. In Table 4, we see that without contrastive learning, our TGM is competitive with both tube and motion-guided masking and only suffers a minor performance drop. With contrastive learning, TGM improves by 2.5% in finetune and 5.7% in linear probe performance, and also outperforms MGM with contrastive learning. This is surprising since MGM explicitly models motion while our TGM does not leverage explicit visual cues for where to mask. These results indicate that the semantic alignment from video-text contrastive learning is still synergistic with MAE in this high-motion setting.

Mask	V → T	Epic-Kitchens FT	Epic-Kitchens Linear
Tube [39]		35.3	16.3
MGM [11]		35.2	15.3
MGM	✓	36.9	20.1
TGM		34.7	14.4
TGM	✓	**37.2**	**20.1**

Table 4: Egocentric action recognition performance on Epic-Kitchens 55 [7]. Models are pretrained for 200 epochs on K400.

Ratio	Finetune
0.55	67.1
0.60	**67.5**
0.75	66.4

(a) Mask ratio

Masking	Finetune	Linear
Bottom-K	67.2	33.0
Tube [39]	65.5	33.3
Top-K	**67.5**	**33.4**

(b) Top vs. bottom-K sampling.

# Cap.	Finetune	Linear
1	66.5	31.3
3	**67.5**	**33.4**

(c) # of captions per video.

Table 5: Ablations with TGM pretrained for 200 epochs on SSv2 and evaluated on SSv2. 5a: mask ratio, 5b: masking bottom vs. top patches by textual similarity, 5c: number of captioned frames.

4.6 Ablations

Mask Ratio. In Table 5a, we ablate the mask ratio used for TGM with contrastive learning when pretrained on SSv2 for 200 epochs. The optimal masking ratio for our text-guided mask is 0.6, which is significantly lower than other video MAE works. For example, the optimal mask ratio in VideoMAE [39], ST-MAE [15], and MGMAE [20] is 0.9, while the optimal mask ratio in MGM [11] is 0.75. This suggests that TGM masks more information dense regions.

Top vs. Bottom-K Textual Similarity. In Table 5b, we ablate the choice to mask the top patches by visual-to-text similarity. When we instead mask the bottom patches with lowest visual-to-text similarity, there is a slight drop in performance in both finetune and linear evaluation. However, even with bottom-K sampling, finetune performance is still better than random masking, while linear probe performance is slightly better. We hypothesis that textual guidance is more structured than random masking, but sampling patches with lower CLIP similarity makes the text-to-video alignment noisier.

Number of Captions. In Table 5c, we ablate the number of captions per video. We sample either the center frame or 3 uniformly spaced frames and observe that both finetune and linear probe performance is better with more diverse captions, however the center frame caption already provides strong performance.

4.7 Visualizations

We offer visualizations from three perspectives. First, we visualize the masking algorithm and see that the text-guided mask masks the most salient regions of

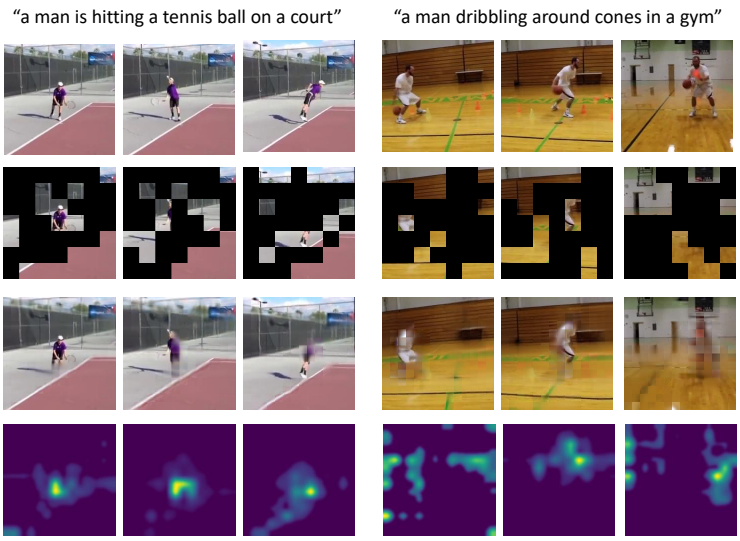


Fig. 3: Visualizations from three perspectives: the visualized mask (row 2), reconstructed RGB output (row 3), and encoder attention map (row 4). Our TGM learns the reconstruction task reasonably well and attends to the salient video regions.

the video matching the natural language description generated by BLIP for the video. Second, we see that the reconstruction quality is decent, meaning the model has learned to solve the reconstruction task. Third, we plot the encoder attention map using the center patch of the center frame as the query. The encoder roughly attends the salient regions of the video. Overall, this suggests that our model has achieved good alignment with the text while solving the MAE reconstruction task. We emphasize that visualizations are not intended to provide a formal explanation for model behavior. Our intention is to provide additional insights into the model to complement our quantitative results.

5 Discussion and Limitations

Unifying MAE and Contrastive Learning. The performance of previous contrastive learning works such as CLIP [36] may make the boost from contrastive learning observed in Table 2, 3, 4 seem obvious. We offer two perspectives for why these results are insightful.

First, previous contrastive learning works leverage pretrained weights and/or are trained on hundreds of millions of image/video-text pairs – which is several orders of magnitude more data than what we use in this paper (roughly 200K video-text pairs). For instance, CLIP [36] and FLIP [27] train on 400 million image-text pairs and ViCLIP [44] is trained on 200 million video-text pairs. We showed in Table 2 that on Kinetics-400 and Something-Something, pure video-text contrastive learning achieves much lower performance than pure MAE. The

benefits of video-text contrastive learning when training from scratch on our scale of data are only realized when combined with MAE.

Second, careful design is required to unify the generative nature of MAE and discriminative nature of contrastive learning. Previous work even suggests that MAE and video-text contrastive learning are antagonistic, which conflicts with our findings. For instance, FLIP [27] reports degraded finetuning performance when combining image MAE with masked contrastive image-text learning. It is not obvious that MAE and video-text contrastive learning are synergistic for video. Our empirical results contribute the useful insight that this is not only the case, but also that this benefit can be realized across multiple masking algorithms – even with noisy machine-generated captions on "regular" sized datasets.

We leave better theoretical understanding of why these paradigms are complementary to future work. To provide an additional perspective, we plot the contrastive loss with and without optimization of the contrastive loss for all three mask algorithms in Figure 4. We observe that even for pure MAE, the contrastive loss naturally decreases. This suggests that the MAE encoder already learns semantics that somewhat align with text, even when there is no textual supervision. This further motivates the benefit of unifying MAE and contrastive learning to facilitate semantic-aware reconstruction.

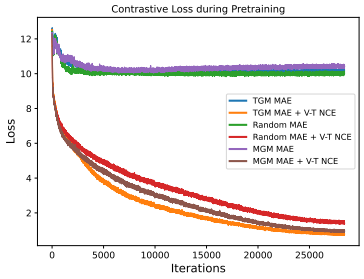


Fig. 4: Contrastive loss for each mask alg. with and without optimization.

Choice of Captions. Our work utilizes frame-wise captions from BLIP-2 [25, 26] which is an off-shelf image captioning model. An interesting question is whether our performance is dependent on this particular captioning model, and whether higher model capacity helps. In Table 6b, we first test an oracle captioner that directly outputs the action label. We then use GPT3.5 [2] which is a large language model with several orders of magnitude more parameters than BLIP. GPT3.5 is a pure language model so it requires a textual prompt. Previous works [32, 34] demonstrate that GPT-3 is capable of generating detailed descriptions about specific object and action categories with only textual inputs. Following CuPL [34], we utilize three prompt templates listed in Table 6a to generate "vision-free" captions per action class. We then utilize these vision-free captions for pretraining to assess the effect of large-language model generated captions in our combined framework.

We see that linear probe performance for both the oracle and GPT3.5 is much higher than BLIP, while finetune performance drops. We posit that this is primarily because the oracle and GPT3.5 are vision-free, so the generated caption is not guaranteed to capture the granular visual details of any video. However, the textual caption generated by GPT3.5 may abstractly match the action class better since it is asked to describe each action class in general, and GPT3.5 has several orders of magnitude more parameters and training data. In sum-

GPT Prompts
"Describe the action {}."
"What does a person {} look like?"
"What does the act of {} look like?"

(a)

Text Source	FT	Linear
Oracle	64.1	51.5
GPT3.5 [2]	66.2	**54.0**
BLIP [27]	**67.5**	33.4
BLIP + GPT3.5	65.8	51.4

(b)

Table 6: 6a: Prompts provided to GPT3.5 to generate text-based captions per action. Inspired by CuPL [34]. 6b: Effect of different text sources when pretraining on SSv2 for 200 epochs and evaluated on SSv2.

mary, more model capacity and external domain knowledge does not necessarily translate to better performance in our framework, however vision-free captions are still surprisingly useful. Combining BLIP and GPT3.5 captions degrades performance probably due to the conflicting nature of vision-free captions.

Computational Efficiency. Although mask generation requires CLIP inference, the wall clock time for training is not significantly higher. For example, at mask ratio 0.65, training on SSv2 for 200 epochs takes 15 hours for TGM vs. 13 hours for MGM [11] on our setup. This is because the high mask ratio and asymmetric encoder-decoder design of the MAE pipeline contribute to dataloading itself being the primary bottleneck [15]. BLIP and CLIP both are frozen.

Limitations. One limitation is the reliance upon captions. However, the captions used in our work are far from perfect. For instance, BLIP is an image captioning model and frame-wise captioning may not capture temporal details of video. Despite this limitation, we showed the efficacy of these noisy captions within our framework. Video captioning is a difficult problem and we expect that as state-of-the-art improves, our performance would also improve. Another limitation is that strong vision-text pretraining is needed to leverage captions for mask generation. However, the growing availability of CLIP-like models makes research into how to leverage these resources all the more timely, especially for designing real-world systems. We have proposed both a new direction of research and a simple yet effective baseline that offers room for further research.

6 Conclusion

We motivated a new research direction into leveraging natural language to improve masked video representation learning. We first presented TGM, a novel masking algorithm that masks the regions of video with highest alignment to text captions. Next, we introduced masked video-to-text contrastive learning as an optional module that can be combined with video MAE to enrich semantic learning across a host of masking algorithms. When combined with specifically our TGM, we observe improved performance in finetuning, linear probe, and even zero-shot retrieval across six different downstream datasets. Our approach is simple yet effective and sets a baseline for future research in this new direction of language-guided MAE.

Acknowledgements. We thank Linda Liu and Pichao Wang for their valuable feedback on this work.

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: International Conference on Learning Representations (2021)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Chen, H., Zhang, W., Wang, Y., Yang, X.: Improving masked autoencoders by learning where to mask. arXiv preprint arXiv:2303.06583 (2023)
4. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International conference on machine learning. pp. 1691–1703. PMLR (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
7. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
10. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9588–9597 (2021)
11. Fan, D., Wang, J., Liao, S., Zhu, Y., Bhat, V., Santos-Villalobos, H., MV, R., Li, X.: Motion-guided masking for spatiotemporal representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5619–5629 (2023)
12. Fan, D., Yang, D., Li, X., Bhat, V., Rohith, M.: Look globally and locally: Inter-intra contrastive learning from unlabeled videos. In: ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models (2023)
13. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
14. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3299–3309 (2021)

15. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems **35**, 35946–35958 (2022)
16. Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., Abbeel, P.: M3ae: Multimodal masked autoencoders learn transferable representations. Tech. rep., Technical Report
17. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. pp. 5842–5850 (2017)
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
20. Huang, B., Zhao, Z., Zhang, G., Qiao, Y., Wang, L.: Mgmae: Motion guided masking for video masked autoencoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13493–13504 (2023)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
22. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
23. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
24. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. Advances in Neural Information Processing Systems **35**, 14290–14302 (2022)
25. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
26. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
27. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23390–23400 (2023)
28. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 513–528 (2018)
29. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)

31. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
32. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=jlAjNL8z5cs
33. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
34. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15691–15701 (2023)
35. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Recasens, A., Luc, P., Alayrac, J.B., Wang, L., Strub, F., Tallec, C., Malinowski, M., Pătrăucean, V., Altché, F., Valko, M., et al.: Broaden your views for self-supervised video learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1255–1265 (2021)
38. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
39. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. ArXiv **abs/2203.12602** (2022)
40. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
41. Wang, J., Bertasius, G., Tran, D., Torresani, L.: Long-short temporal contrastive learning of video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14010–14020 (2022)
42. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
43. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence **41**(11), 2740–2755 (2018)
44. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: The Twelfth International Conference on Learning Representations (2023)
45. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
46. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)

47. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)