Structured-Noise Masked Modeling for Video, Audio and Beyond

Aritra Bhowmik[1⋆], Carlos Hinojosa[2⋆], Fida Mohammad Thoker[2⋆], Bernard Ghanem[2], and Cees G. M. Snoek[1]
a.bhowmik@uva.nl,{carlos.hinojosa,fida.thoker}@kaust.edu.sa

[1] University of Amsterdam, The Netherlands
[2] King Abdullah University of Science and Technology, Saudi Arabia
https://carloshinojosa.me/project/structured-noise-masking

Abstract. Masked modeling has emerged as a robust self-supervised learning framework. However, most methods rely on random masking, which disregards the structural properties of different data modalities. To align with the spatiotemporal and spectral characteristics of video and audio data, we introduce a structured noise-based masking approach. By filtering white noise into different color noise distributions, we generate structured masks that capture modality-specific patterns without requiring handcrafted heuristics or access to the data. Our approach enhances masked video and audio modeling frameworks without any additional computational cost. Experiments show that structured noise masking consistently outperforms random masking, underscoring the value of modality-aware masking strategies for representation learning.

Keywords: Masked Autoencoders · Structured-Noise Masking · Self-supervised Learning · Data-independent Masking

1 Introduction

Self-supervised masked modeling has emerged as a powerful paradigm for learning representations across modalities, including images [7, 19, 29], videos [39, 48, 49], and audio [4, 10, 24]. The core idea is to mask portions of the input, such as image patches, spectrogram regions, or spatiotemporal tubes, and train a model to reconstruct them. This mask-and-predict objective encourages the network to capture rich contextual information without relying on labels. In most existing methods, masking is performed randomly: tokens are uniformly dropped from the input while the model learns to reconstruct the missing content. Despite its simplicity and strong empirical performance, random masking disregards modality-specific structure. Images exhibit spatial coherence, videos contain strong temporal continuity, and audio spectrograms follow structured time–frequency patterns. As a result, random masking may disrupt informative structures in the signal, leading to suboptimal training supervision.

⋆ Equal contribution.

Natural visual and audio signals exhibit strong structure in the frequency domain. For example, images and videos typically follow approximately $1/f$ spectral statistics, while audio spectrograms display structured frequency–time energy distributions. Masking patterns themselves also possess spectral characteristics, determined by how masked and visible tokens are spatially arranged. Consequently, the interaction between signal spectra and mask spectra can influence reconstruction difficulty and the quality of learned representations. As we show in Appendix A.14 through frequency-domain analysis of masking patterns, structured masks with controlled spectral properties can lead to improved representation learning compared to purely random masking.

Motivated by these limitations, several works have explored alternative masking strategies that better align with the intrinsic structure of the data [6,20,23,25, 34]. In image modeling, for example, adaptive masking methods generate masks guided by semantics, attention maps, or motion cues [6,8,13,22,30]. While effective, these approaches typically rely on predefined heuristics, auxiliary networks, or additional supervision [30] and may incur significant computational overhead (e.g., motion priors [13, 22] or adversarial training [8]). Such dependencies can limit their scalability and make it difficult to apply them consistently across different modalities.

An alternative direction, which we pursue in this paper, is to generate structured masks from filtered noise distributions. By filtering white noise into different color noise distributions, it is possible to produce masks with structured spatial patterns while remaining fully data-independent. This idea was recently explored for static images in ColorMAE [21], where white noise was filtered into distinct 2D color noise types to generate structured masking patterns. However, extending such structured noise masking beyond static images is not straightforward. Video data introduces strong spatiotemporal dependencies, requiring masks that evolve coherently over time, while audio spectrograms exhibit structured time–frequency patterns that call for balanced coverage across both axes. These modality-specific properties raise the question of how structured noise masks can be designed to align with the inductive biases of different data modalities. In this work, we propose modality-specific noise filters that generate masks aligned with each modality's inductive biases, enhancing self-supervised learning via the mask-and-predict objective. Our main contributions are:

- **3D Green masking (Green3D) for video:** a color–noise–based masking strategy that produces spatiotemporally coherent masks, preserving spatial clustering while maintaining temporal smoothness.
- **Regularized Blue noise (R-BN) masking for audio:** a 2D blue-noise-based optimization algorithm that produces masks by enforcing a controlled distribution of visible spectrogram patches across time and frequency, leveraging the structure of audio.
- **Multimodal evaluation:** Through extensive evaluation across video, audio, and audio-visual benchmarks, we demonstrate that our modality-specific masking consistently improves existing self-supervised methods in unimodal and multimodal settings.

To our knowledge, our method is the first data-independent masking strategy to consistently outperform random masking across video, audio, and multimodal masked modeling without introducing additional computational overhead.

2 Related Work

Masked Modeling for Video. Masked modeling is a widely used self-supervised paradigm that originated with masked language modeling in BERT [12], and was later extended to audio with wav2vec [5,41] and vision by MAE [19,53]. The main idea is to randomly mask a portion of the input data, then reconstruct it using an asymmetric encoder-decoder network. VideoMAE [49] extends MAE to videos using spatiotemporal random tube masking. Most follow-up works still rely on random masking but with slightly different architectures [15, 33, 39, 45, 48]. For example, SIGMA [39] refines the learning objective by guiding reconstruction with a Sinkhorn-based matching loss. SMILE [48] infuses both spatial and motion semantics by leveraging image-language pretrained models such as CLIP. Although effective, these methods often increase computational overhead and parameter count due to architectural complexity.

Masked Modeling for Audio. Masked modeling has also been explored in audio. Wav2vec 2.0 [5] masks latent speech units with a contrastive loss, while Data2Vec [4] unifies masked prediction across modalities. SS-AST [16] adapts masked spectrogram prediction to semi-supervised audio tasks, MaskSpec [10] proposes spectrogram-aware masking, and MAE-AST [2] applies MAE to audio spectrogram transformers. Audio-MAE [24] adapts MAE to spectrograms at scale and achieves competitive performance. In addition, multimodal extensions such as MST [31] explore synchronized tube masking across audio and video streams, while MultiMAE [3] leverages shared latent tokens and cross-attention to jointly reconstruct multiple modalities. However, these approaches typically rely on learned or cross-modal masking mechanisms rather than predefined, data-independent patterns. Despite these advances, most works still rely on random masking, which may discard semantically important or structurally coherent regions of the spectrogram.

Data-driven Masking. These methods adapt the masking process itself, using data semantics, attention, or motion cues. For images, attention-guided masking [43] selects informative patches, SemMAE [30] masks semantic parts, and CL-MAE [34] applies curriculum masking. For video, AdaMAE [6] introduces an adaptive masking strategy that selects informative spatiotemporal patches, enabling effective pretraining even at high masking ratios. Also, MGMAE [22] and MGM [13] guide the masking with optical flow or motion vectors. In audio, spectrogram-aware masking [10] adapts to time-frequency context, while MAVIL [23] balances masking across modalities. Although effective, these methods rely on external supervision, such as flow estimation, pretrained semantics, or reinforcement learning, which increases complexity. This motivates alternatives that achieve competitive performance without incurring additional computational cost.

Data-independent Masking. Unlike adaptive methods, data-independent approaches replace random masking with deterministic rules that capture spatial, temporal, or spectral structure without extra supervision. For images, BEiT [7] explored block- or grid-based masking to enforce uniform coverage. In speech, SpecAugment [35] masks time or frequency bands. Data-independent frequency-based masking variants include MFM [52], FMAE [32] for multimodal biosignals, and iBOT [56] focusing on high-frequency content. More recently, ColorMAE [21] introduced color noise–based masking for images, filtering white noise with low-pass, band-pass, or high-pass filters to generate structured masks that are data-independent, efficient, and better aligned with localized spatial patterns. While effective, it remains restricted to static 2D inputs and does not address temporal coherence in video, spectral structure in audio, or complementary masking needs in multimodal learning. Building on color masking, we propose **Green3D masking** for spatiotemporal video and **Regularized Blue Noise (R-BN) masking** for spectrogram-aware audio, further combining them for multimodal setups. This yields simple, training-free masking strategies designed for different modalities, avoiding motion priors or auxiliary supervision while remaining competitive with adaptive methods.

3 Methodology

3.1 Preliminaries

Masked Modeling. In general, an input X (image, video, or audio spectrogram) is partitioned into patches and embedded into a sequence of token representations via a function ϕ, yielding $X_p{=}\phi(X) \in \mathbb{R}^{N\times d}$, where N is the number of patches and d the embedding dimension. A binary mask $M \in \{0,1\}^N$ is generated by a masking function η with a mask ratio $\gamma \in [0,1]$, using random noise $n_w \sim \mathcal{N}(0,1)$: $M = \eta(X_p, n_w, \gamma)$. Each entry M_i corresponds to patch i, with $M_i{=}1$ if the patch is visible and $M_i{=}0$ if masked. The token sets are

$$X_p^{\text{visible}} = X_p \odot M \quad , \quad X_p^{\text{masked}} = X_p \odot (1 - M),$$

where $\odot$ denotes the Hadamard product. The encoder processes X_p^{visible}, while the decoder reconstructs the full input as $\hat{X}$ by integrating both visible and masked tokens. Training minimizes mean squared error: $\mathcal{L}_{\text{recon}}{=}|X - \hat{X}|_2^2$. Random masking is widely used due to its simplicity [19, 24, 49], but it ignores modality-specific inductive biases such as spatial coherence in images, temporal continuity in videos, and spectral patterns in audio.

Color Noise Masking. Instead of random masking, *structured noise* can be leveraged to introduce modality-specific masks. Unlike white noise, which has a uniform power distribution across all frequencies, filtering it through frequency constraints produces *structured noise patterns* that align with the spatial and temporal structures of the input modality [11, 28]. Given white noise n_w, frequency-constrained noise is constructed by filtering it with a d-dimensional

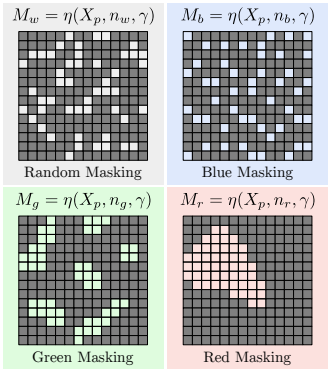

Fig. 1: Generated masks from 2D random, blue, green, and red noise.

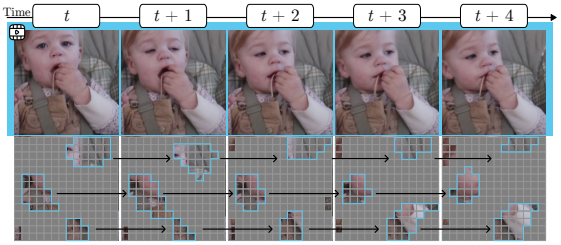

Fig. 2: Unlike traditional random tube masking, which enforces strict temporal consistency, our proposed Green 3D masking generates structured random masks that evolve smoothly across consecutive frames. This smooth evolution prevents abrupt changes in masking, enabling the model to better capture natural temporal dynamics and continuity in video data.

Gaussian kernel $G_\sigma(\mathbf{x}){=}(2\pi)^{-d/2}\sigma^{-d}\exp\big(-\|\mathbf{x}\|^2/(2\sigma^2)\big)$, where $\mathbf{x}\in\mathbb{R}^d$ are spatial coordinates and σ controls the filter scale. Convolving n_w with G_σ of different bandwidths yields distinct noise components:

$$n_r = G_\sigma * n_w, \quad (1) \quad n_b = n_w - G_\sigma * n_w, \quad (2) \quad n_g = G_{\sigma_1} * n_w - G_{\sigma_2} * n_w, \quad (3)$$

where $*$ denotes convolution and $\sigma_1{<}\sigma_2$. These noise patterns define the structured masks: $M_r{=}\eta(X_p, n_r, \gamma)$, $M_b{=}\eta(X_p, n_b, \gamma)$, and $M_g{=}\eta(X_p, n_g, \gamma)$. The precise definition of η is provided in Algorithm 2 in Appendix A.10.

As shown in Fig. 1 for $d{=}2$, red noise (n_r) preserves low frequencies, producing smooth, large-scale masks; blue noise (n_b) enhances high-frequency details, creating fine-grained masks; green noise (n_g) balances both, generating mid-sized, clustered masks. These structured masks encourage the model to learn features aligned with spatial frequency. In the following sections, we extend this principle beyond images to *video (Green3D)* and *audio (Regularized Blue)* data.

3.2 Green 3D Noise for Video Masking

Video masking should respect both spatial coherence and temporal continuity. Standard methods, such as VideoMAE [49] and SIGMA [39], rely on random tube masking, which applies a static mask across frames, preserving temporal consistency but lacking adaptability to motion. To address this, we propose *Green3D Noise Masking*, which introduces structured, evolving masks across frames, enhancing fine-grained temporal representation learning.

Green3D Mask Generation. We generate the proposed noise for video masking by applying equation 3 with a 3D white noise tensor n_w and two 3D Gaussian kernels G_σ ($d{=}3$) defined as:

$$G_\sigma(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{3}{2}}\sigma^3}\exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right), \quad (4)$$

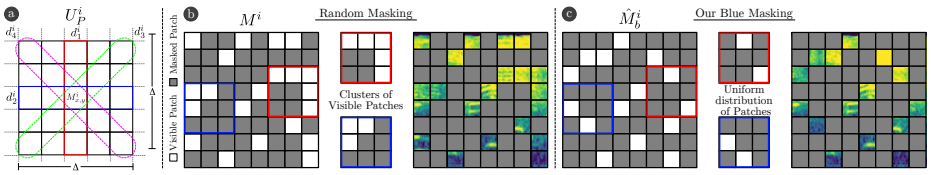

Fig. 3: ⓐ Illustration of the metric used to determine the concentration of visible patches in a window U_P^i of the mask $M_{x,y}^i$. ⓑ Example of the initial mask (M^i), with clusters of visible patches, and ⓒ final mask ($\hat{M}_b^i$) obtained with our regularized blue noise masking algorithm, with uniformly distributed visible patches. Note the improved uniformity in the final mask, which ensures better coverage and reduces undesirable clustering effects.

where $\mathbf{x}=(x,y,z) \in \mathbb{R}^3, \sigma \in \{\sigma_1, \sigma_2\}$, and $\sigma_1<\sigma_2$ control the frequency response. A smaller σ_1 preserves fine details, while a larger σ_2 removes high-frequency components. Instead of fixing (σ_1, σ_2), which produces masks with limited diversity and either overly static or noisy temporal behavior, we randomly sample (σ_1, σ_2) within a bounded range $[0.5, 2]$ for each sequence. This generates an ensemble of 3D green noise tensors, encouraging diverse but still mid-frequency spatiotemporal patterns. Empirically, this stochasticity yields smoother mask evolution and prevents overfitting to a single frequency, making the approach more robust than a fixed parameterization. Our Green3D masks are obtained as $\eta(X_p, n_g^{3D}, \gamma)$, where η follows the masking function in [19,21]. As seen in Fig. 2, our Green3D masks evolve smoothly over time, avoiding abrupt frame-to-frame changes and enabling the model to learn temporal continuity more effectively.

3.3 Regularized Blue Noise for Audio Masking

Self-supervised audio learning relies on spectrogram representations, where structured time–frequency patterns encode meaningful information. However, random masking [24] misaligns with the inherent structure of spectrograms. As seen in Fig. 1, random, green, and red noise masking create clusters of visible patches or large masked regions. While beneficial in vision tasks, these clusters do not correspond to meaningful time-frequency events in audio. Instead, blue noise masking offers a more effective strategy, distributing visible patches more evenly.

Blue noise patterns have been widely studied in computer graphics and image processing [1, 11, 38, 51] for their ability to suppress low-frequency components. A simple way to generate blue noise is to filter white noise with a Gaussian kernel, as in equation 2. However, this does not explicitly control the separation between visible patches, still leading to small clusters. To overcome this, we introduce an optimization-based approach that enforces spatial separation constraints to ensure a better distribution of visible patches. This leads to our proposed *Regularized Blue Noise (R-BN)* Masking, which ensures a more well-distributed masking pattern for spectrogram-based audio representations.

Regularized Blue Noise (R-BN) Mask Generation. We generate an initial set of K candidate masks $\{M^i\}_{i=1}^K$ by thresholding noise values from n_w or n_b at the target ratio γ. Each mask is then iteratively optimized to maintain uniform patch separation. For each spatial position $P{=}(x, y)$, processed in randomized order, we consider a local window $U_P^i \in \mathbb{R}^{\Delta\times\Delta}$ centered at P for each candidate M^i. To measure clustering, we count visible patches aligned with P along four orientations—horizontal (d_1^i), vertical (d_2^i), and the two diagonals (d_3^i, d_4^i). The clustering score is:

$$S_P^i = w_1 d_1^i + w_2 d_2^i + w_3 d_3^i + w_4 d_4^i, \tag{5}$$

where w_1, w_2, w_3, w_4 weight the directional counts (set to 1 by default). The candidate with the lowest score is selected as $\hat{i}{=}\arg\min_i S_P^i$. The patch update is then applied:

$$\hat{M}_{x,y}^i = \begin{cases} 1, & \text{if } i = \hat{i} \text{ (visible)}, \\ 0, & \text{otherwise (masked)}. \end{cases} \tag{6}$$

This process continues until the number of visible patches in each mask reaches $(1-\gamma)N$, ensuring the target masking ratio. The optimized masks $\hat{M}_b$ are retained for training, yielding more uniform and well-separated visible patches than the standard blue-noise masks M_b. As shown in Fig. 3, this procedure reduces local clustering effects observed in prior work [21]. Empirically, we find that R-BN improves representation learning on audio tasks, particularly in settings where time–frequency coverage is crucial. Full pseudocode is provided in Appendix A.9.

Our design choices for the masks are motivated by modality-specific structural properties: videos exhibit mid-frequency spatiotemporal statistics, whereas audio spectrograms contain fine-grained time–frequency harmonic patterns with energy concentrated in low-to-mid frequencies. We provide a frequency-domain analysis in Appendix A.14 that supports these structured mask choices.

3.4 Blue & Green Noise for Audio-Visual Masking

Our color masking extends naturally to joint audio-video setups. Specifically, we apply Green masking M_g to video frames for structured spatial masking with frame-wise consistency, while Distributed Blue noise masks $\hat{M}_b$ enforce a uniform distribution of visible patches in audio. This modality-specific masking better aligns with the structural properties of each domain, enhancing joint pretraining within a unified masked modeling framework.

Notably, all our proposed masks, including Green3D and R-BN, are precomputed as mask tensors (e.g., Green3D masks of shape $N{\times}64{\times}64{\times}64$). During training, they undergo standard augmentations such as random flipping, normalization, and resizing to match the input volume (e.g., $14{\times}14{\times}8$), preserving their noise properties while providing diverse and efficient structured masking without any added computational overhead.

4 Experiments

4.1 Results for Video

Experimental Setup. We evaluate two masked video modeling frameworks: VideoMAE [49], which reconstructs raw video pixels from masked inputs, and SIGMA [39], which predicts DINO features of masked frames using a Sinkhorn-cluster matching loss. For each framework, we replace the default random tube masking with our proposed Green3D masking while keeping the architecture and pretraining hyperparameters unchanged. Following the original setups [39, 49], we use an encoder-decoder framework with a ViT-B backbone. After pretraining, the decoder is discarded, and the encoder is fine-tuned for downstream tasks. Additional implementation details are provided in Appendix A.11.

Standard Action Recognition

Datasets. Following prior works [13, 22, 45, 49], we evaluate on two standard action recognition benchmarks: **Kinetics-400 (K400)** [26] and **Something-Something V2 (SSv2)** [18]. K400 comprises 240k training and 20k validation videos across 400 human action categories, emphasizing spatial and object-centric cues. In contrast, SSv2 contains 169k training and 25k validation clips across 174 categories, focusing on fine-grained temporal interactions, making it a more challenging benchmark for video SSL learning. We follow the fine-tuning and evaluation protocol of [49] and report top-1 accuracy for both.

Something-Something Results. We evaluate SSv2 under two settings: (i) *in-domain*, where models are pretrained and finetuned on SSv2, and (ii) *cross-domain*, where pretraining is on K400 and finetuning on SSv2. Results are reported in Table 1, alongside state-of-the-art masked video modeling methods.

Our Green3D masking improves VideoMAE by +1.2% in both in-domain (SSv2→SSv2) and cross-domain (K400→SSv2) transfer, validating its effectiveness over random tube masking for learning richer spatiotemporal representations. We attribute these gains to the harder mask-reconstruction patterns generated by Green3D, which encourage stronger spatiotemporal modeling.

Notably, VideoMAE+Green3D matches or surpasses motion-guided approaches such as MGMAE [22] and MGM [13]. Specifically, it outperforms MGM in the in-domain setting and MGMAE in the cross-domain setting. Unlike these data-adaptive methods, which require access to motion priors (e.g., optical flow or motion vectors) and add substantial computational overhead (MGMAE is $\sim$1.5$\times$ slower than VideoMAE), Green3D is data-independent and incurs no additional cost because masks are precomputed.

Finally, Green3D benefits recent frameworks such as SIGMA [39], yielding +0.8% in-domain and +0.7% cross-domain gains, highlighting its versatility as a lightweight, plug-and-play masking strategy for current and future video models.

Table 1: Comparison of masked video modeling methods on Something-Something V2 and Kinetics-400 for standard action recognition. All results use a ViT-B backbone pretrained on K400 or SSv2 for 800 epochs. Our Green3D masking consistently improves both in-domain and cross-domain performance for VideoMAE as well as advanced architectures such as SIGMA.

	Masking Type		SSv2 Pretraining	K400 Pretraining	
Method	**Data-independant**	**Data-adaptive**	**SSv2 Top-1**	**SSv2 Top-1**	**K400 Top-1**
VideoMAE	Random	-	69.6	68.5	80.0
VideoMAE + Our masking	Green3D	-	70.8 (+1.2%)	69.7 (+1.2%)	80.5 (+0.5%)
CMAE-V	Random	-	69.7	-	80.2
OmniMAE	Random	-	69.5	69.0	80.8
MME	Random	-	70.0	70.5	81.5
MGM	-	Motion	70.6	71.1	80.8
MGMAE	-	Motion	71.0	68.9	81.2
SIGMA	Random	-	71.2	71.1	81.5
SIGMA + Our masking	Green3D	-	72.0 (+0.8%)	71.8 (+0.7%)	82.1 (+0.6%)

Kinetics Results. For K400, we evaluate in-domain pretraining as in Table 1, following prior work. Similar to SSv2, our Green3D masking yields consistent gains over random masking (+0.5% for VideoMAE and +0.6% for SIGMA), showing that our method also enhances spatial semantics, which is crucial for datasets like K400, where many actions are distinguished by spatial cues.

Unsupervised Video Object Segmentation

Setup. Following [39], we evaluate the spatial, temporal, and semantic representations learned by our method on the unsupervised video object segmentation benchmark from [40]. Unlike action recognition, which pools space–time features into a global clip representation, this benchmark assesses the encoder's ability to produce temporally consistent segmentation maps. Space–time features are clustered using k-means with a predefined cluster count K and aligned with ground-truth masks via the Hungarian algorithm [27]. Segmentation quality is measured by mean Intersection over Union (mIoU), with two evaluation modes: *clustering* when K matches the ground-truth object count, and *overclustering* when K exceeds it. We report results on **DAVIS** [37] and **YTVOS** [54]. Dataset and evaluation details are provided in Appendix A.11.

Results. As shown in Table 2, equipping VideoMAE with Green3D masking yields substantial improvements across all settings. On DAVIS clustering, it improves VideoMAE by +8.7% and surpasses MGMAE by +7.2%, demonstrating stronger object awareness and spatiotemporal representations. Consistent gains on YTVOS further validate its effectiveness. Since our setup matches MGMAE and MGM except for the masking strategy, these results confirm that Green3D better preserves spatiotemporal object continuity. Again, applying Green3D to SIGMA improves performance, highlighting its ability to generalize across pretraining frameworks and downstream tasks.

Table 2: Comparison of masked video methods for unsupervised video object segmentation. Following the evaluation protocol from [40], we report mIoU for clustering and overclustering. We evaluate the ViT-B backbone pretrained on K400 and use the officially released checkpoints for all prior works. Equipping VideoMAE with our masking significantly improves its performance, surpassing even motion-guided masking methods. Our masking also improves SIGMA when used as a plugin.

	Clustering		Overclustering	
Method	**YTVOS**	**DAVIS**	**YTVOS**	**DAVIS**
VideoMAE	34.1	29.5	61.3	56.2
VideoMAE + Our masking	35.6 (+1.5%)	**38.2** (+8.7%)	62.5 (+1.5%)	58.2 (+2.0%)
MGM	36.6	36.5	61.2	56.6
MGMAE	34.5	31.0	60.1	57.5
SIGMA	41.1	33.1	67.1	59.0
SIGMA + Our masking	**42.1** (+1.3%)	34.2 (+1.2%)	**68.4** (+1.3%)	**60.0** (+1.0%)

SEVERE Generalization Benchmark

Setup. Following recent works in video self-supervised learning [46] [47] [39], we further assess the robustness and generalization of our learned video representations on the SEVERE benchmark [46], a comprehensive suite comprising eight experiments across four axes of generalization: *domain shift*, *sample efficiency*, *action granularity*, and *task shift*. Domain shift is assessed on **SomethingSomething v2** and **FineGym (Gym99)** [42], which diverge from the Kinetics-400 pretraining domain. Sample efficiency is evaluated through low-shot action recognition with only 1,000 fine-tuning samples, denoted as **UCF**(10^3) and **Gym**(10^3). Action granularity examines fine distinctions between semantically related categories using the **FX-S1** and **UB-S1** splits of FineGym, where differences hinge on subtle gymnastic elements such as variations in "jump" or "balance". Finally, task shift evaluates out-of-distribution tasks, namely temporal repetition counting on **UCFRep** [55] and multi-label recognition on **Charades** [44]. We adhere to the original SEVERE training and evaluation protocols, with further details provided in Appendix A.11.

Results. Table 3 summarizes the comparison with recent self-supervised methods on the SEVERE benchmark. For *Domain Shift*, we observe higher accuracy on SSv2 and Gym99 relative to the respective baselines, suggesting stronger robustness to unseen domains. Under *Sample Efficiency*, VideoMAE+Ours outperforms motion-guided approaches MGM and MGMAE on **Gym**(10^3), while SIGMA+Ours achieves a 7% gain over the SIGMA baseline, indicating effective adaptation in low-shot settings. In the *Action Granularity* evaluation, our method yields better results on FX-S1 and UB-S1 for both frameworks, reflecting improved handling of fine-grained motion categories. For *Task Shift*, our approach maintains competitive performance on repetition counting and improves the performance on Charades multi-label recognition. Note that for UCF-RC,

Table 3: Comparison on the SEVERE benchmark [46] evaluating domain shift, sample efficiency, action granularity, and task shift of learned video representations. All methods are pretrained on K400 with a ViT-B backbone. Our method improves the downstream generalization of both VideoMAE and SIGMA architectures.

Method	Domain shift		Sample efficiency		Action granularity		Task shift		Mean
	SSv2	Gym99	UCF(10^3)	GYM(10^3)	FX-S1	UB-S1	UCF-RC↓	Charades	
VideoMAE	68.6	86.6	74.6	25.9	36.6	74.3	0.172	17.2	58.3
VideoMAE + Our masking	69.7	88.1	75.0	29.9	38.6	74.8	0.170	18.2	59.6
MVD	70.0	82.5	67.1	17.5	31.3	50.5	0.184	16.1	52.1
MGMAE	68.9	87.2	77.2	24.1	33.7	79.5	0.181	17.9	58.8
MGM	71.1	89.1	78.4	26.4	38.6	86.9	**0.152**	22.5	62.2
MME	70.1	89.7	79.2	29.8	55.5	**87.2**	0.155	23.6	65.0
SIGMA	70.9	89.7	84.1	28.0	55.1	79.9	0.169	23.1	64.2
SIGMA + Our masking	**71.8**	**90.7**	**85.0**	**35.0**	**56.5**	85.9	0.163	**27.5**	**67.3**

lower values indicate better repetition-counting accuracy, as the metric reports mean counting error. *Overall.* On average, Green3D masking increases the SEVERE mean score by 1.3% with VideoMAE and by 3% with SIGMA. These improvements across all SEVERE factors demonstrate that our modality-aware masks generalize well beyond the training distribution, strengthening cross-domain and long-horizon robustness.

4.2 Results for Audio

Setup. We evaluate our proposed **Regularized Blue Noise (R-BN)** masking within both AudioMAE [24] and MaskSpec [10], replacing their default random masking while leaving the rest of the framework unchanged. Following the AudioMAE protocol, we use a ViT-B backbone, apply an 80% masking ratio during pretraining and 30% during fine-tuning, and discard the decoder after pretraining. Pretraining is conducted on the large-scale **AudioSet-2M** [14], which contains over 2M audio clips spanning a diverse range of sound events, including non-speech audio. For downstream evaluation, the pretrained encoder is fine-tuned separately on **AudioSet-20K (AS-20k)**, a balanced 20k subset of AudioSet for efficient benchmarking; **ESC-50** [36], consisting of 2,000 environmental sound recordings across 50 classes; and **SPC-2** [50], which provides additional coverage of sound classification tasks. Note that the publicly available AudioSet test split has changed over time due to YouTube removals; hence, results from our evaluation are marked with $*$. We evaluate both the baselines and our method on the same current split to ensure a fair comparison. More training and evaluation details are provided in Appendix A.12. This setup enables us to assess the effect of R-BN across both large-scale pretraining and fine-grained audio recognition tasks, ensuring fair comparison with prior baselines.

Results. Table 4 compares AudioMAE [24] and MaskSpec [10] with their R-BN variants, alongside other self-supervised audio baselines. Our masking method consistently improves performance across all benchmarks: for AudioMAE, R-BN

Table 4: Comparison of self-supervised audio pretraining methods. Our R-BN masking improves over AudioMAE and MaskSpec across all benchmarks. ∗ denotes results from our evaluation.

Method	AS-20k	AS-2M	ESC-50	SPC-2
SS-AST	31.0	-	88.8	98.0
MaskSpec	32.3	47.1	89.6	97.7
MaskSpec + Ours	33.4 (+1.1%)	**47.6** (+0.5%)	90.4 (+0.8%)	98.2 (+0.5%)
MAE-AST	30.6	-	90.0	97.9
Audio-MAE*	36.1	46.3	94.1	98.3
Audio-MAE + Ours	**36.8** (+0.7%)	47.2 (+0.9%)	**94.6** (+0.5%)	**98.7** (+0.4%)

Table 5: Comparison on VGG-Sound with audio-only, video-only, and audio-visual inputs. Our Green3D and R-BN masking improve performance across all settings. ∗ denotes results from our evaluation.

Method	Audio	Video	Audio-Video
MBT	52.3	**51.2**	64.1
CAV-MAE*	58.5	45.6	64.3
CAV-MAE + Ours	**59.1** (+0.6%)	46.4 (+0.8%)	**64.9** (+0.6%)

yields +0.7% on AS-20k, +0.9% on AS-2M, +0.5% on ESC-50, and +0.4% on SPC-2; for MaskSpec, the gains are even larger, reaching +1.1%, +0.5%, +0.8%, and +0.5%, respectively. These results confirm that R-BN enhances both frameworks, outperforming prior baselines without requiring additional supervision or heuristics. Unlike MaskSpec, which relies on predefined time–frequency rules, or MAE-AST [2], which benefits from extra speech data, R-BN provides a simple, data-independent masking strategy that naturally aligns with the spectral structure of audio signals, offering a robust and generalizable alternative to rigid masking schemes.

4.3 Results for Audio-Visual

Setup. We evaluate our proposed modality-specific masking within the CAV-MAE framework [17]. Random masking in the original model is replaced by **Green3D** masking for video frames and **Regularized Blue Noise (R-BN)** masking for audio spectrograms, while all other components remain unchanged. Following prior work, we adopt a ViT-B backbone, apply a masking ratio of 75% to both modalities, and pretrain on the large-scale **VGGSound** dataset [9], which consists of ∼200K audio-visual clips across 309 classes, including non-speech audio, with strong natural correspondence between sound and vision. Since the original AudioSet setup cannot be fully reproduced, we follow the authors' publicly released VGGSound pipeline and evaluate both models under identical conditions to ensure a fair comparison. Pretraining runs for 25 epochs, after which the encoder is evaluated under unimodal (audio-only, video-only) and multimodal (audio–visual) settings. Full details are in Appendix A.13.

Results. Table 5 reports results on VGGSound. Incorporating Green3D for video and R-BN for audio consistently improves performance across all evaluation modes: audio-only (+0.6%), video-only (+0.8%), and audio–visual (+0.6%). The unimodal gains indicate that structured noise masking enhances modality-specific representation learning, while the multimodal gains highlight improved cross-modal alignment. As CAV-MAE [17] employs random masking for both streams, these results demonstrate that simple, data-independent structured

Table 6: Impact of color masks. Left: 3D masks for video, where Green3D performs best. Right: 2D masks for audio, where Blue is strongest among other color variants.

	Video			Audio		
Color Mask	$\mathcal{L}_{\text{recon}}$	mini Kinetics	mini SSv2	$\mathcal{L}_{\text{recon}}$	AS-20k	ESC-50
Random	0.67	51.6	52.8	0.52	36.1	94.1
Blue	0.41	50.9	52.1	0.45	36.5	94.2
Red	0.85	51.0	52.3	0.61	35.5	92.6
Green	0.60	52.7	54.5	0.57	36.4	94.1

Table 7: Blue vs R-BN and Green3D vs Green2D. For audio (top), R-BN beats vanilla Blue masking. For video (bottom), Green3D outperforms tube and Green 2D masking.

Masking type	$\mathcal{L}_{\text{recon}}$	AS-20k	ESC-50
Blue	0.45	36.5	94.2
R-BN	0.49	36.8	94.6
Masking type	**$\mathcal{L}_{\text{recon}}$**	**mini-Kinetics**	**mini-SSv2**
Tube	0.67	51.6	52.8
Green2D	0.73	51.9	52.9
Green3D	0.60	52.7	54.5

noise can serve as a plug-and-play replacement, strengthening both unimodal and joint representations without adding objectives or computational cost.

4.4 Ablations

We ablate mask colors and types using smaller subsets of standard datasets: **mini-Kinetics** (25% of K400) and **mini-SSv2** (50% of SSv2) for video, and **AudioSet-20K (AS-20k)** and **ESC-50** for audio. Additional ablations and qualitative results are provided in Appendices A.1-A.8.

Impact of color noise on video masking. Table 6 reports performance and reconstruction loss for different color noise masks. **Green** masking achieves the best accuracy on both mini-Kinetics and mini-SSv2, striking a balance between reconstruction difficulty and solvability (loss = 0.60). In contrast, **Blue** lowers reconstruction loss too much (0.41), making the task overly easy and limiting representation quality, while **Red** noise imposes excessive difficulty (0.85), also reducing accuracy. This shows that Green provides a more effective inductive bias for video masking by respecting spatiotemporal coherence.

Impact of color noise on audio masking. For audio, the trend is reversed: among the color variants in Table 6, **Blue** yields the best accuracy on AS-20k and ESC-50 with a low reconstruction loss (0.45). **Green** provides no benefit and underperforms **Blue**, while **Red** makes the reconstruction task overly difficult (loss = 0.61), leading to degraded accuracy. This shows that for audio, blue noise provides the most effective inductive bias for masking.

Blue vs R-BN. Table 7 compares the vanilla Blue noise with our regularized blue noise **R-BN** for audio. Our **R-BN** improves the performance on both downstream datasets compared to vanilla 2D blue noise masking. This validates our assumption that regularizing the vanilla blue noise to introduce greater separability among visible patches for audio masking yields better representations.

3D vs. 2D video masking. To examine whether structured noise masking for images can be directly applied to video, we compare our proposed **Green3D**

Table 8: SEVERE fine-grained and segmentation results for 3D masking. Compared with the random baseline and other 3D masks, Green3D achieves the strongest overall performance across selected SEVERE tasks and unsupervised video object segmentation.

Method	SEVERE				Clustering		Overclustering	
	Gym99	FX-S1	UB-S1	Charades	YTVOS	DAVIS	YTVOS	DAVIS
Random	72.0	35.7	71.3	14.2	29.7	25.3	56.2	43.9
Red3D	72.5	37.5	**73.0**	15.4	31.2	25.9	54.7	48.6
Blue3D	73.4	34.7	69.3	15.8	31.3	25.5	55.6	44.1
Green3D	**75.4**	**38.5**	72.6	**16.0**	**32.9**	**27.3**	**60.0**	**50.9**

masking with a naive extension of the 2D Green noise masks introduced in ColorMAE [21]. In this baseline, the same 2D Green noise mask is applied independently to each video frame, ignoring temporal relationships between frames. As shown in Table 7, this 2D masking strategy performs only marginally better than random tube masking on both mini-Kinetics and mini-SSv2, indicating that simply transferring the image-based masking scheme to video provides limited benefit. In contrast, our proposed **Green3D** masking generates masks with explicit spatiotemporal coherence, leading to consistently lower reconstruction loss and clear performance gains across both datasets. These results highlight that extending structured noise masking from images to video is not a trivial adaptation. Modeling temporal dependencies through 3D masking is crucial for producing effective masking patterns and learning robust video representations.

Masking ratio ablation We examine the effect of the masking ratio for both video and audio. VideoMAE typically peaks around 90% video masking, and AudioMAE around 80% audio masking; our Green3D and R-BN follow the same trends (gains from 80% to 90% for video, with performance peaking near 80% for audio). Results are provided in Appendix A.3 (Tables A.3-A.4).

Fine-grained evaluation We evaluated masking strategies on fine-grained tasks in SEVERE and video object segmentation in Table 8. While Red3D and Blue3D can improve certain metrics, our Green3D consistently achieves strong performance, supporting the importance of modality-aligned masking design.

Comparison with per-frame 2D Green masking. We additionally evaluated independently sampled 2D Green masks per frame, achieving 51.7/53.1 on mini-Kinetics/mini-SSv2, compared to 51.9/52.9 for tube-style Green2D masking and 52.7/54.5 for our Green3D. This suggests that simply increasing frame-wise mask diversity is insufficient; coherent 3D spatiotemporal structure is important for video masking.

Qualitative Results on Video Object Segmentation. Figure 4 shows qualitative examples on the DAVIS dataset, comparing segmentation masks from VideoMAE and VideoMAE+Green3D. Consistent with our quantitative results (Sec. 4.1), Green3D masking produces sharper, more object-centric segmentations and preserves temporal consistency across frames. These visualizations

confirm that structured masking not only improves mIoU but also enhances the interpretability of learned spatiotemporal representations.

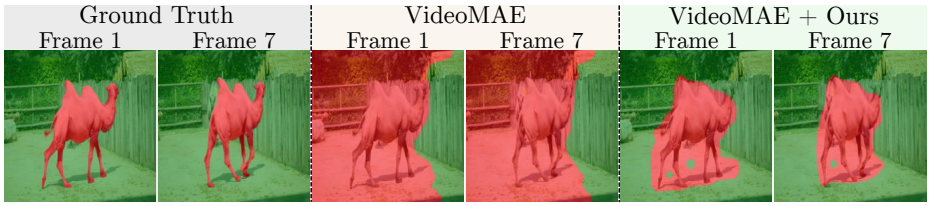

Fig. 4: Qualitative results on DAVIS. Green3D masking produces sharper and temporally consistent segmentations compared to VideoMAE.

5 Conclusion

Self-supervised learning via masked modeling overlooks inherent structures within different data modalities by relying on random masking. In this work, we show that structured noise-based masking offers a simple yet effective alternative that naturally aligns with the spatial, temporal, and spectral characteristics of video and audio data. By leveraging color noise distributions, our approach introduces structured masking without requiring handcrafted heuristics or additional data. Consistent improvements in multiple benchmarks demonstrate that modality-aware masking enhances representation learning without increasing computational costs. These findings reinforce a broader perspective: self-supervised masked modeling benefits not just from masking large portions of data, but from doing so in a way that respects the structure of the modality.

Acknowledgement. This work has been financially supported by TomTom, the University of Amsterdam, and the allowance of Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy. Cees G. M. Snoek is also (partially) funded by the Horizon Europe project ELLIOT (GA No. 101214398). The research was also supported by funding from King Abdullah University of Science and Technology (KAUST) Center of Excellence for Generative AI, under award number 5940. For computing time, this research used Ibex, managed by the Supercomputing Core Laboratory at King Abdullah University of Science and Technology (KAUST) in Thuwal, Saudi Arabia. We also extend our gratitude to the anonymous reviewers for their valuable feedback and insightful suggestions during the rebuttal stage, which considerably improved this work.

References

1. Ahmed, A.G., Wonka, P.: Screen-space blue-noise diffusion of monte carlo sampling error via hierarchical ordering of pixels. ACM Transactions on Graphics (TOG) **39**(6), 1–15 (2020)
2. Baade, A., Peng, P., Harwath, D.: Mae-ast: Masked autoencoding audio spectrogram transformer. arXiv preprint arXiv:2203.16691 (2022)
3. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: European Conference on Computer Vision. pp. 348–367. Springer (2022)
4. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International conference on machine learning. pp. 1298–1312. PMLR (2022)
5. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
6. Bandara, W.G.C., Patel, N., Gholami, A., Nikkhah, M., Agrawal, M., Patel, V.M.: Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14507–14517 (2023)
7. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
8. Chen, H., Zhang, W., Wang, Y., Yang, X.: Improving masked autoencoders by learning where to mask. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 377–390. Springer (2023)
9. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020)
10. Chong, D., Wang, H., Zhou, P., Zeng, Q.: Masked spectrogram prediction for self-supervised audio pre-training. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
11. Correa, C.V., Arguello, H., Arce, G.R.: Spatiotemporal blue noise coded aperture design for multi-shot compressive spectral imaging. Journal of the Optical Society of America A **33**(12), 2312–2322 (2016)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
13. 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)
14. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
15. Girdhar, R., El-Nouby, A., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Omnimae: Single model masked pretraining on images and videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10406–10417 (2023)

16. Gong, Y., Lai, C.I., Chung, Y.A., Glass, J.: Ssast: Self-supervised audio spectrogram transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10699–10709 (2022)
17. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.: Contrastive audio-visual masked autoencoder. arXiv preprint arXiv:2210.07839 (2022)
18. 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)
19. 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)
20. Hernandez, J., Villegas, R., Ordonez, V.: Vic-mae: Self-supervised representation learning from images and video with contrastive masked autoencoders. In: European Conference on Computer Vision (ECCV). pp. 444–463. Springer (2024)
21. Hinojosa, C., Liu, S., Ghanem, B.: ColorMAE: Exploring data-independent masking strategies in Masked AutoEncoders. In: European Conference on Computer Vision. pp. 432–449. Springer (2024)
22. 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)
23. Huang, P.Y., Sharma, V., Xu, H., Ryali, C., Li, Y., Li, S.W., Ghosh, G., Malik, J., Feichtenhofer, C., et al.: Mavil: Masked audio-video learners. Advances in Neural Information Processing Systems **36**, 20371–20393 (2023)
24. Huang, P.Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., Feichtenhofer, C.: Masked autoencoders that listen. Advances in Neural Information Processing Systems **35**, 28708–28720 (2022)
25. Kara, D., Kimura, T., Chen, Y., Li, J., Wang, R., Chen, Y., Wang, T., Liu, S., Abdelzaher, T.: Phymask: An adaptive masking paradigm for efficient self-supervised learning in iot. In: Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems. pp. 97–111 (2024)
26. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
27. Kuhn, H.W.: The hungarian method for the assignment problem. In Naval research logistics quarterly **2**(1-2), 83–97 (1955)
28. Lau, D.L., Ulichney, R., Arce, G.R.: Blue and green noise halftoning models. IEEE Signal Processing Magazine **20**(4), 28–38 (2003)
29. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3041–3050 (2023)
30. 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)
31. Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., et al.: Mst: Masked self-supervised transformer for visual representation. Advances in Neural Information Processing Systems **34**, 13165–13176 (2021)

32. Liu, R., Zippi, E.L., Pouransari, H., Sandino, C., Nie, J., Goh, H., Azemi, E., Moin, A.: Frequency-aware masked autoencoders for multimodal pretraining on biosignals. In: ICLR 2024 Workshop on Learning from Time Series For Health (2024)
33. Lu, C., Jin, X., Huang, Z., Hou, Q., Cheng, M., Feng, J.: CMAE-V: contrastive masked autoencoders for video action recognition. CoRR **abs/2301.06018** (2023)
34. Madan, N., Ristea, N.C., Nasrollahi, K., Moeslund, T.B., Ionescu, R.T.: Cl-mae: Curriculum-learned masked autoencoders. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2492–2502 (2024)
35. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019)
36. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1015–1018 (2015)
37. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
38. Rauhut, H.: Compressive sensing and structured random matrices. Theoretical foundations and numerical methods for sparse recovery **9**(1), 92 (2010)
39. Salehi, M., Dorkenwald, M., Thoker, F.M., Gavves, E., Snoek, C.G., Asano, Y.M.: Sigma: Sinkhorn-guided masked video modeling. In: European Conference on Computer Vision. pp. 293–312. Springer (2024)
40. Salehi, M., Gavves, E., Snoek, C.G.M., Asano, Y.M.: Time does tell: Self-supervised time-tuning of dense image representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16536–16547 (2023)
41. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862 (2019)
42. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
43. Sick, L., Engel, D., Hermosilla, P., Ropinski, T.: Attention-guided masked autoencoders for learning image representations. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 836–846. IEEE (2025)
44. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In: European Conference on Computer Vision (ECCV) (2016)
45. Sun, X., Chen, P., Chen, L., Li, C., Li, T.H., Tan, M., Gan, C.: Masked motion encoding for self-supervised video representation learning. In: CVPR. pp. 2235–2245. IEEE (2023)
46. Thoker, F.M., Doughty, H., Bagad, P., Snoek, C.G.: How severe is benchmark-sensitivity in video self-supervised learning? In: European Conference on Computer Vision (ECCV) (2022)
47. Thoker, F.M., Doughty, H., Snoek, C.G.M.: Tubelet-contrastive self-supervision for video-efficient generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
48. Thoker, F.M., Jiang, L., Zhao, C., Ghanem, B.: Smile: Infusing spatial and motion semantics in masked video learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8438–8449 (2025)
49. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)

50. Warden, P.: Speech commands: A dataset for limited-vocabulary speech recognition. arXiv preprint arXiv:1804.03209 (2018)
51. Wolfe, A., Morrical, N., Akenine-Möller, T., Ramamoorthi, R., Ghosh, A., Wei, L.: Spatiotemporal blue noise masks. In: EGSR (ST). pp. 117–126 (2022)
52. Xie, J., Li, W., Zhan, X., Liu, Z., Ong, Y.S., Loy, C.C.: Masked frequency modeling for self-supervised visual pre-training. arXiv preprint arXiv:2206.07706 (2022)
53. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
54. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
55. Zhang, H., Xu, X., Han, G., He, S.: Context-aware and scale-insensitive temporal repetition counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
56. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)