Rethinking Masking Strategies for Masked Prediction-based Audio Self-supervised Learning

Daisuke Niizumi[†‡¶], Daiki Takeuchi[‡], Masahiro Yasuda[‡], Binh Thien Nguyen[‡],
Noboru Harada[‡], and Nobutaka Ono[†]
[†]*Tokyo Metropolitan University*, Tokyo, Japan
[‡]*NTT Inc.*, Atsugi, Japan
[¶]daisukelab.cs@gmail.com

***Abstract*—Since the introduction of Masked Autoencoders, various improvements to masking techniques have been explored. In this paper, we rethink masking strategies for audio representation learning using masked prediction-based self-supervised learning (SSL) on general audio spectrograms. While recent informed masking techniques have attracted attention, we observe that they incur substantial computational overhead. Motivated by this observation, we propose dispersion-weighted masking (DWM), a lightweight masking strategy that leverages the spectral sparsity inherent in the frequency structure of audio content. Our experiments show that inverse block masking, commonly used in recent SSL frameworks, improves audio event understanding performance while introducing a trade-off in generalization. The proposed DWM alleviates these limitations and computational complexity, leading to consistent performance improvements. This work provides practical guidance on masking strategy design for masked prediction-based audio representation learning.**

***Index Terms*—masking strategy, masked autoencoders, spectrogram, audio representation learning**

I. Introduction

Masked Autoencoders (MAE) [1] have had a significant impact on representation learning through masked prediction-based self-supervised learning (SSL), initially in the image domain, and have subsequently influenced various studies on general-purpose audio representations [2]–[4]. While the original MAE demonstrated the effectiveness of fully random masking, subsequent work in the image domain showed that more structured strategies can lead to improved representations, such as block masking in BEiT [5] and inverse block masking (IBM) in data2vec 2.0 [6]. In addition to masking strategies that rely purely on randomness regardless of data content, several recent studies have explored informed masking approaches [7], [8], which exploit attention patterns or patch-level representations derived from pretrained models or from the model itself during training.

In the audio domain, many existing methods largely follow the random masking strategy introduced in MAE, whereas recent SSL approaches have reported performance gains by adopting IBM [9], [10]. However, in these studies, the contributions of masking strategies are often entangled with other training components, and evaluations are typically conducted on a limited set of downstream tasks. As a result, the role of masking strategy design in masked prediction-based audio representation learning remains insufficiently understood, especially in terms of generalization performance, leaving room for further investigation into alternative masking strategies.

This work was partially supported by JST Strategic International Collaborative Research Program (SICORP), Grant Number JPMJSC2306, Japan.

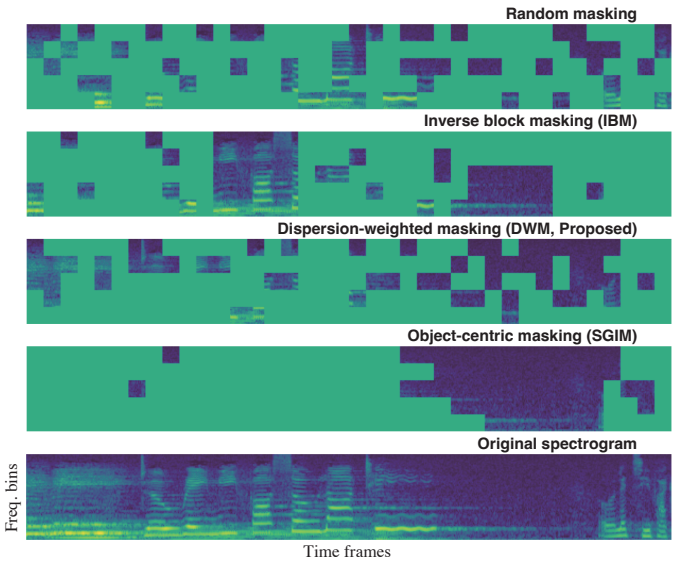

Fig. 1. Masking examples on an audio spectrogram (hint ratio $r_h = 0.0$). Object-centric masking (SGIM) almost entirely covers audio events (objects), whereas the proposed DWM applies dispersion-weighted random masking based on patch-wise spectral variability, yielding masking patterns that loosely align with object-centric behavior.

In this study, we investigate how various masking strategies, ranging from random masking to informed masking, contribute to masked prediction-based representation learning for general audio (e.g., environmental sound, speech, and music) using spectrogram inputs. Using three masked prediction-based SSL frameworks, we conduct a comparative analysis of random masking, IBM, informed masking, and the proposed dispersion-weighted masking (DWM), as illustrated in Fig. 1.

DWM is motivated by the observation that typical audio events do not occupy all frequency components uniformly, resulting in blank or low-energy regions in spectrogram frequency bins. By leveraging this property, DWM assigns higher masking probabilities to regions with larger dispersion, thereby incorporating spectral structure into the masking

process. Compared to previous informed masking approaches that require non-negligible computational overhead to identify candidate regions, DWM provides a simple and lightweight algorithm, enabling a spectrogram-friendly and efficient form of informed masking.

Our experimental results provide insights into how different masking strategies affect representation learning. While IBM facilitates the learning of representations that are well suited for audio event semantics, it degrades performance on other downstream tasks, resulting in an overall trade-off. Informed masking approaches, on the other hand, incur non-negligible computational costs due to the need to compute masking candidates using the model during training. In contrast, the proposed DWM improves performance on audio event understanding without increasing computational complexity, while suppressing performance degradation on other tasks and preserving generalization performance.

The findings of this study, together with the proposed DWM, are applicable to a broad range of masked prediction-based audio representation learning methods and contribute to masking strategy design in the following ways:

- We clarify the impact of masking strategies on audio spectrogram-based representation learning, highlighting their effects across different downstream tasks.
- We propose DWM, a lightweight masking strategy tailored to audio spectrograms, which leverages dispersion patterns with minimal additional computational cost.

We publicly release our code[1] to support reproducibility and to facilitate further advances in masked prediction-based audio representation learning.

II. Preliminary

We introduce the background and prior work necessary to frame our study on masking strategies for masked prediction-based representation learning in this section.

A. Masked prediction-based representation learning for audio spectrograms

Masked prediction-based representation learning operates by dividing an input into visible and masked patches, denoted as x_v and x_m, respectively. An encoder $f(\cdot)$ processes only the visible patches to produce latent representations $z_v = f(x_v)$. A predictor $g(\cdot)$ then infers the masked part of the input as $\hat{z}_m = g(\mathrm{concat}(z_v, m))$, where m denotes mask tokens concatenated at the masked positions.

The form of the prediction target $\hat{z}_m$ and the role of the predictor $g(\cdot)$ depend on the specific method. For example, in MAE [1], $\hat{z}_m$ corresponds to the reconstructed input signal in the masked regions, and $g(\cdot)$ acts as a reconstruction decoder. In Masked Modeling Duo (M2D) [11], also used in our experiments, $\hat{z}_m$ represents latent representations of the masked regions, and $g(\cdot)$ serves as a predictor in the representation space.

[1]https://github.com/onolab-tmu/audio_ssl_masking

For audio spectrograms, MAE-AST [2], MSM-MAE [3], Audio-MAE [4], and CED [12] are MAE-based methods. Prior to MAE, SSAST [13] also learns representations using a reconstruction-based objective. In contrast, M2D [11], ATST [14], EAT [9], and AaPE [10] learn by predicting masked patch representations, while BEATs [15] predicts tokenized labels. M2D-CLAP [16] further extends M2D by combining masked prediction with contrastive language-audio pre-training (CLAP), enabling the model to directly learn semantic information from natural language descriptions of audio. Among these methods, most adopt a random masking strategy. More recently, EAT and AaPE have demonstrated performance improvements by leveraging IBM.

B. Masking strategies

Among existing masking strategies, we consider random masking, inverse block masking (IBM) [6], and self-guided informed masking (SGIM) [8] in our evaluation.

a) Random Masking: In MAE, several masking variants, such as block and grid masking, were investigated, and random masking was found to be the most effective, enabling higher masking ratios while maintaining strong performance.

b) Inverse Block Masking: IBM starts with a complete mask over all patches and iteratively restores original content in block-shaped regions, which may overlap, until the number of masked embeddings matches a target masking ratio. As a result, visible patches remain in contiguous block regions; however, the selection of these regions is performed randomly and independently of the input content, making IBM a content-agnostic masking strategy. EAT [9] extends IBM to audio spectrograms and provides empirical evidence of its utility for audio representation learning.

c) Self-Guided Informed Masking: SGIM, proposed in SG-MAE [8], is an object-centric masking strategy that applies masks to estimated object-centric regions.

SGIM assigns an object relevance score S_i to each of the N input patches, where i denotes the patch index, and determines masked and visible patches based on their ranking. A hint ratio is used to randomly replace a subset of masked patches with visible ones to control task difficulty, and is gradually annealed toward zero through scheduling, as in [7].

To compute these relevance scores, SGIM performs a Normalized Cut [17]-based partitioning using pairwise similarities between patch features [18]. Specifically, it applies an eigenvalue decomposition to the similarity matrix constructed from patch representations, and the resulting partition is used to evaluate the relevance score of each patch.

However, this procedure involves eigenvalue decomposition, whose computational complexity generally scales as $O(N^3)$. This high computational cost poses a challenge to training efficiency in self-supervised pre-training, where a large number of patches and batch-wise processing are required.

In our preliminary implementation based on M2D using four A100 GPUs, random masking required approximately 7 minutes per epoch, whereas SGIM incurred a substantially higher cost of approximately 35 minutes, resulting in roughly

a fivefold slowdown. Due to this computational overhead, applying SGIM broadly across masked prediction-based learning frameworks can be impractical.

These limitations motivate the need for a lightweight alternative that retains the benefits of informed masking while remaining computationally efficient.

III. Proposed Masking Method: DWM

DWM aims to approximate informed masking tailored to audio spectrograms while maintaining low computational cost. As illustrated by the object-centric masking example produced by SGIM in Fig. 1 (Object-centric masking), patches corresponding to spectrally sparse regions are often assigned as visible in SGIM. Unlike images, audio spectrograms naturally contain blank regions when certain frequency components are absent. Based on this observation, DWM is motivated by the idea that the dispersion of a spectrogram patch can serve as a proxy for the amount of information related to audio events contained in that patch.

As shown in Algorithm 1, DWM operates on an input patch sequence X and first estimates patch-wise importance using mean absolute deviation (MAD) as a dispersion metric. The MAD of a patch x_i is defined as $\mathrm{MAD}(x_i) = \frac{1}{n}\sum_j |x_{i,j} - \bar{x}_i|$, where n is the number of elements in the patch and $\bar{x}_i$ is the patch mean, and ϵ is a small constant to prevent division by zero. The resulting dispersion values are converted into sampling probabilities, which are used to sample an initial set of masked patches M_0 in a weighted manner. From M_0, a subset of patches is randomly selected according to the hint ratio and returned to the visible set. To preserve the target mask ratio, the same number of patches is then randomly selected from the visible set and reassigned as masked.

As a result, patches with higher dispersion are probabilistically more likely to be masked, while the hint-based exchange process prevents the masking task from becoming overly difficult. Although DWM does not explicitly model object-centric regions, it assigns lower masking probability to spectrally sparse patches, which partially aligns its masking behavior with the effects observed in SGIM for audio spectrograms, without expensive computations.

Similar to SGIM, the hint ratio r_h is determined by a scheduling strategy. Following SemMAE [7], r_h is updated based on training progress as

$$r_h = 1.0 - \left(\frac{\text{epoch}}{\text{total\_epochs}}\right)^{\gamma}, \tag{1}$$

where γ is a scheduling parameter.

Algorithm 1 Dispersion-weighted Masking (DWM)

Input: Input sequence X, mask ratio r_m, hint ratio r_h.
Output: Visible indices $\mathcal{V}_{\text{final}}$, Masked indices $\mathcal{M}_{\text{final}}$.

 # Step 1: Setup parameters
$L \leftarrow$ Total number of patches
$N_{\text{keep}} \leftarrow \lfloor L \cdot (1 - r_m) \rfloor$, $N_{\text{mask}} \leftarrow L - N_{\text{keep}}$
$N_{\text{hint}} \leftarrow \lfloor N_{\text{mask}} \cdot r_h \rfloor$
 # Step 2: Importance estimation
for each patch x_i in X **do**
  $\omega_i \leftarrow \mathrm{MAD}(x_i)$ {Calculate mean absolute deviation}
end for
$P(i) \leftarrow (\omega_i + \epsilon) / \sum(\omega_j + \epsilon)$ {Sampling probability}
 # Step 3: Initial weighted sampling
$\mathcal{M}_0 \leftarrow \mathrm{Sample}(\{1, \ldots, L\}, N_{\text{mask}}, \text{prob} = P)$
$\mathcal{V}_0 \leftarrow \{1, \ldots, L\} \setminus \mathcal{M}_0$
 # Step 4: Hint-based exchange process
$\mathcal{H} \leftarrow \mathrm{RandomSelect}(\mathcal{M}_0, N_{\text{hint}})$ {Select hints}
$\mathcal{V}_{\text{temp}} \leftarrow \mathcal{V}_0 \cup \mathcal{H}$
$\mathcal{C} \leftarrow \mathrm{RandomSelect}(\mathcal{V}_{\text{temp}}, N_{\text{hint}})$ {Maintain N_{keep}}
 # Step 5: Final index sets
$\mathcal{V}_{\text{final}} \leftarrow \mathcal{V}_{\text{temp}} \setminus \mathcal{C}$
$\mathcal{M}_{\text{final}} \leftarrow \{1, \ldots, L\} \setminus \mathcal{V}_{\text{final}}$

IV. Experiments

We investigated the effects of different masking strategies on masked prediction-based representation learning for general audio, focusing on both task-specific performance and generalization across diverse downstream tasks. Learned representations were evaluated through linear evaluation and fine-tuning on benchmarks used in M2D [11], and the proposed method was further analyzed via ablation studies on hint ratio scheduling.

Linear evaluation focuses on the linear separability of frozen representations across diverse downstream tasks and also serves as an indicator of generalization. Fine-tuning evaluates end-to-end performance under task-specific optimization, reflecting practical effectiveness on standard benchmarks. Our ablation studies examined the sensitivity of DWM to hint ratio scheduling, providing insight into its design behavior.

We evaluated three masking strategies, random masking, IBM, and the proposed DWM, by integrating each into masked prediction-based SSL frameworks. SGIM was excluded from evaluation due to its significant computational overhead, which renders large-scale pre-training impractical in our experimental setup.

Experiments were conducted using a series of three SSL frameworks: MSM-MAE [3], which adapts MAE to audio spectrograms; M2D [11], which learns by predicting masked patch representations; and M2D-CLAP [16], which extends M2D by incorporating semantic supervision from audio-text alignment. These frameworks share most of the encoder architecture and training pipeline, differing primarily in their learning objectives, which makes them well suited for a controlled comparison in our experiments.

A. Experimental Setup

All pre-training settings followed those in the base SSL methods, including a masking ratio of 0.75 for MSM-MAE and 0.7 for M2D and M2D-CLAP. We pre-trained two models per masking strategy for MSM-MAE and M2D, while only one model per strategy was trained for M2D-CLAP due to resource constraints. For DWM, the hint scheduling parameter γ was

TABLE I
LINEAR EVALUATION RESULTS (%) WITH 95% CI COMPARING DIFFERENT MASKING STRATEGIES.

	Env. sound tasks		Speech tasks				Music tasks			
Masking strategy	ESC-50	US8K	SPCV2	VC1	VF	CRM-D	GTZAN	NSynth	Surge	Avg.
(MSM-MAE variants)										
Random	89.2 ±0.4	87.2 ±0.2	**96.1 ±0.1**	**73.4 ±0.4**	**97.9 ±0.1**	**71.3 ±0.2**	80.3 ±1.0	74.6 ±0.4	43.2 ±0.2	79.2
IBM [6]	**90.1 ±0.3**	**88.2 ±0.4**	95.4 ±0.1	65.1 ±0.5	97.7 ±0.1	70.9 ±0.6	80.5 ±2.3	74.2 ±0.6	42.8 ±0.1	78.3
vs. random [pp]	+0.9	+1.0	-0.7	-8.3	-0.2	-0.4	+0.2	-0.4	-0.4	-0.9
DWM (proposed)	89.9 ±0.7	87.4 ±0.1	95.9 ±0.1	72.5 ±0.7	97.8 ±0.1	70.9 ±0.3	**81.8 ±1.1**	**75.5 ±0.1**	**43.3 ±0.3**	**79.5**
vs. random [pp]	+0.7	+0.2	-0.2	-0.9	-0.1	-0.4	+1.5	+0.9	+0.1	+0.3
(M2D variants)										
Random	91.3 ±0.6	87.6 ±0.2	96.0 ±0.1	**73.4 ±0.2**	98.3 ±0.0	73.0 ±0.7	84.1 ±2.7	**75.7 ±0.1**	42.1 ±0.2	80.2
IBM [6]	**92.8 ±0.3**	**88.2 ±0.2**	95.8 ±0.1	69.1 ±0.3	98.3 ±0.0	**73.2 ±0.7**	**85.3 ±2.0**	75.5 ±0.6	**42.5 ±0.2**	80.1
vs. random [pp]	+1.5	+0.6	-0.2	-4.3	0.0	+0.2	+1.2	-0.2	+0.4	-0.1
DWM (proposed)	91.9 ±0.3	87.9 ±0.2	**96.1 ±0.1**	72.8 ±0.4	**98.4 ±0.0**	73.0 ±0.8	85.1 ±1.2	75.4 ±0.5	42.0 ±0.3	**80.3**
vs. random [pp]	+0.6	+0.3	+0.1	-0.6	+0.1	0.0	+1.0	-0.3	-0.1	+0.1
(M2D-CLAP variants)										
Random	95.2 ±0.3	89.5 ±0.2	**95.6 ±0.1**	**68.6 ±0.5**	**98.1 ±0.1**	73.4 ±0.7	86.6 ±1.6	**76.5 ±0.4**	42.2 ±0.3	80.6
IBM [6]	**95.8 ±0.2**	89.5 ±0.2	95.2 ±0.1	65.7 ±0.2	97.9 ±0.1	**73.6 ±0.6**	86.4 ±0.9	76.5 ±0.6	42.5 ±0.2	80.3
vs. random [pp]	+0.6	0.0	-0.4	-2.9	-0.2	+0.2	-0.2	0.0	+0.3	-0.3
DWM (proposed)	95.6 ±1.0	**90.2 ±0.4**	95.5 ±0.1	67.9 ±0.5	98.0 ±0.0	73.2 ±1.8	**87.7 ±4.3**	75.2 ±0.4	**43.0 ±0.2**	**80.7**
vs. random [pp]	+0.4	+0.7	-0.1	-0.7	-0.1	-0.2	+1.1	-1.3	+0.8	+0.1
(Reference: Results of prior methods evaluated in this work)										
BEATs [15] (Random)	86.9 ±1.4	84.8 ±0.1	89.4 ±0.1	41.4 ±0.7	94.1 ±0.3	64.7 ±0.8	72.6 ±4.3	75.9 ±0.2	39.3 ±0.4	72.1
EAT [9] (IBM)	85.6 ±0.4	81.7 ±0.3	81.5 ±0.4	39.6 ±0.5	92.6 ±0.0	64.9 ±2.8	73.7 ±0.5	71.9 ±0.1	39.0 ±0.3	70.0

set to 2, coinciding with the value reported in SemMAE [7]. The hint ratio was initialized at 1.0 and gradually annealed to zero, as illustrated in Fig. 2. For IBM, we followed EAT [9] for the block size (i.e., 5×5).

All evaluation settings followed those in [3], [11], [16], [19], including the evaluation platform (EVAR[2]) and the downstream classification tasks spanning environmental sound, speech, and music. For each model and downstream task, we conducted three runs and report the resulting statistics as the final results. When two models are pre-trained, this results in six evaluations per pre-training setting.

Environmental sound tasks include AudioSet [20], ESC-50 [21], UrbanSound8K [22] (US8K); speech tasks include Speech Commands V2 [23] (SPCV2), VoxCeleb1 [24] (VC1), VoxForge [25] (VF), and CREMA-D [26] (CRM-D); and music tasks include GTZAN [27], NSynth [28], and the Pitch Audio Dataset [29] (Surge). AudioSet is evaluated under two settings: AS2M using the full 2M training samples and AS20K using 21K samples from the balanced training set. Surge is a pitch classification task over 88 MIDI notes.

All tasks are formulated as classification problems, and performance is reported in terms of accuracy or mean average precision (mAP).

B. Comparison of Masking Strategies in Linear Evaluation

With the linear evaluation setup, we compared masking strategies by assessing the linear separability of frozen representations across downstream tasks, which also serves as an indicator of representation generalization.

[2]https://github.com/nttcslab/eval-audio-repr

TABLE II
FINE-TUNING RESULTS WITH 95% CI COMPARING MASKING STRATEGIES.

Masking strategy	AS2M mAP	AS20K mAP	ESC-50 acc(%)	SPCV2 acc(%)	VC1 acc(%)
(MAE variants)					
Random	47.4 ±0.1	37.9 ±0.0	95.4 ±0.1	**98.4 ±0.0**	**96.6 ±0.2**
IBM [6]	47.4 ±0.1	**38.6 ±0.1**	95.4 ±0.3	98.3 ±0.0	95.4 ±0.5
diff.[pp]	0.0	+0.7	0.0	-0.1	-1.2
DWM	**47.5 ±0.1**	38.3 ±0.1	**95.5 ±0.2**	**98.4 ±0.1**	96.4 ±0.2
diff.[pp]	+0.1	+0.4	+0.1	0.0	-0.2
(M2D variants)					
Random	47.8 ±0.1	38.6 ±0.1	96.0 ±0.2	**98.4 ±0.1**	**96.3 ±0.2**
IBM [6]	**47.9 ±0.2**	**39.3 ±0.1**	**96.2 ±0.2**	**98.4 ±0.0**	95.4 ±0.7
diff.[pp]	+0.1	+0.7	+0.2	0.0	-0.9
DWM	47.8 ±0.0	38.9 ±0.0	96.0 ±0.2	**98.4 ±0.1**	96.1 ±0.2
diff.[pp]	0.0	+0.3	0.0	0.0	-0.2
(M2D-CLAP variants)					
Random	49.0 ±0.1	42.1 ±0.0	97.8 ±0.1	98.4 ±0.0	**95.5 ±0.3**
IBM [6]	**49.1 ±0.1**	**42.3 ±0.1**	**98.0 ±0.1**	98.4 ±0.1	94.2 ±0.1
diff.[pp]	+0.1	+0.2	+0.2	0.0	-1.3
DWM	49.0 ±0.2	41.8 ±0.1	97.8 ±0.1	**98.5 ±0.0**	95.1 ±0.4
diff.[pp]	0.0	-0.3	0.0	+0.1	-0.4
(Reference: Performance reported in prior studies)					
BEATs [15]	48.0	38.3	95.6	98.3	-
EAT [9]	48.6	40.2	95.9	98.3	-

The results in Table I indicate that both IBM and DWM influence downstream task performance consistently across different base SSL frameworks. Their effects vary by task domain: performance improvements are generally observed for environmental sound tasks, whereas degradations tend to appear in speech tasks. For music tasks, the effect differs across tasks: performance improves for genre classification (GTZAN) but degrades for instrument classification (NSynth).

TABLE III
ABLATION OF HINT RATIO SCHEDULING IN DWM: LINEAR EVALUATION RESULTS (%) WITH 95% CI.

	Env. sound tasks		Speech tasks				Music tasks			
Scheduling	ESC-50	US8K	SPCV2	VC1	VF	CRM-D	GTZAN	NSynth	Surge	Avg.
Baseline MSM-MAE	89.2 ±0.4	87.2 ±0.2	96.1 ±0.1	**73.4 ±0.4**	97.9 ±0.1	71.3 ±0.2	80.3 ±1.0	74.6 ±0.4	43.2 ±0.2	79.2
Hint ratio r_h=0.0	**89.9 ±1.1**	87.3 ±0.7	95.5 ±0.1	69.1 ±0.1	97.7 ±0.0	70.4 ±0.6	**83.4 ±4.5**	75.4 ±0.2	42.5 ±0.9	79.0
Hint ratio r_h=0.5	87.4 ±0.9	86.3 ±0.6	96.0 ±0.2	71.8 ±0.5	97.6 ±0.3	70.5 ±0.9	82.1 ±3.1	74.3 ±0.4	42.5 ±0.5	78.7
Schedule γ=1	88.1 ±0.2	87.3 ±0.3	**96.2 ±0.2**	70.0 ±0.5	97.7 ±0.2	**71.4 ±0.9**	82.0 ±6.4	**76.0 ±0.2**	42.6 ±0.7	79.0
Schedule γ=2	**89.9 ±0.7**	**87.4 ±0.1**	95.9 ±0.1	72.5 ±0.7	97.8 ±0.1	70.9 ±0.3	81.8 ±1.1	75.5 ±0.1	**43.3 ±0.3**	**79.5**
Schedule γ=4	89.5 ±0.8	87.3 ±0.3	95.9 ±0.2	**73.4 ±0.5**	**98.0 ±0.1**	69.9 ±1.9	81.7 ±4.5	74.7 ±0.2	43.1 ±0.4	79.3
Schedule γ=8	89.5 ±2.7	87.2 ±0.6	**96.2 ±0.0**	**73.4 ±0.3**	**98.0 ±0.2**	70.1 ±2.0	82.5 ±5.8	72.5 ±0.7	42.9 ±0.4	79.2

Although the variance across tasks is non-negligible, IBM overall exhibits a larger impact on performance compared to DWM.

IBM particularly improves performance on environmental sound tasks (ESC-50 and US8K) by approximately 1 pp, which is consistent with the performance gains on AudioSet reported in EAT that employs IBM. In contrast, notable performance degradation is observed for speaker identification on VC1, with drops of up to 8.3 pp in MSM-MAE and 4.3 pp in M2D. This suggests that IBM may suppress fine-grained speaker-specific cues that are crucial for tasks requiring sensitivity to subtle vocal differences.

DWM exhibits relatively smaller gains on ESC-50 and US8K compared to IBM, while being characterized by notably smaller performance degradation on speech tasks. Although the improvements are modest, DWM consistently improves performance on ESC-50, US8K, and GTZAN, contributing to overall performance gains across tasks. Performance degradation on speech tasks is largely mitigated, with the drop on VC1 remaining within 1 pp, which is substantially smaller than that observed with IBM. Overall, these results suggest that DWM modestly improves performance while better preserving generalization across diverse downstream tasks.

C. Comparison of Masking Strategies in Fine-tuning

In fine-tuning comparisons shown in Table II, the impact of different masking strategies is generally smaller than that observed in linear evaluation; nevertheless, similar trends can still be identified, with performance improvements on environmental sound tasks and notable degradation on VC1.

IBM yields little gain on AS2M but more consistent gains on AS20K, where training data are limited, suggesting that it facilitates learning representations that generalize across diverse patterns of audio events. Conversely, the substantial performance drop on speaker identification (VC1) indicates a trade-off with the fine-grained discriminative information required for this task.

Compared to IBM, DWM exhibits more moderate performance changes across tasks, while consistently reducing degradation on VC1, in line with the observations from linear evaluation. On AS20K, DWM generally maintains favorable performance for MSM-MAE and M2D, indicating improved performance with preserved generalization. In contrast, AS20K performance decreases for M2D-CLAP; although

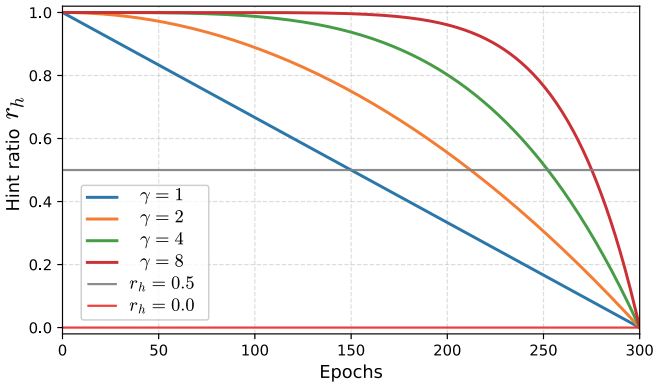

Fig. 2. Hint ratio r_h schedules over training epochs for ablation settings.

based on limited runs due to computational constraints, it suggests that the effectiveness of DWM may be sensitive to training conditions when competing for high-end performance.

D. Ablations on DWM Hint Ratio Scheduling

We conducted an ablation study on the scheduling of the hint ratio r_h in DWM, which is designed to prevent the prediction task from becoming overly difficult in the early stages of training. As illustrated in Fig. 2, we either fixed r_h to a constant value or varied the scheduling parameter γ, and evaluated the resulting MSM-MAE models using linear evaluation. The results are summarized in Table III.

When r_h was fixed to 0.0, some tasks showed improved performance; however, a substantial degradation was observed on VC1. When r_h was fixed to 0.5, performance improvements were observed only on GTZAN, while degradation was evident on the remaining tasks.

In contrast, scheduling r_h using different values of γ resulted in more stable behavior. From $\gamma = 2$, degradation on VC1 was effectively mitigated, and with $\gamma = 8$, performance became nearly comparable to that of random masking. Consistent with observations from SG-MAE [8] and SemMAE [7], these results highlight the importance of appropriately scheduling task difficulty during training, which also holds for DWM.

E. Summary of Experiments

The experiments reveal several insights into the role of masking strategies. First, masking strategies substantially influence downstream performance across tasks, even when the

underlying SSL framework is fixed, indicating that masking design is a non-trivial factor in masked prediction-based representation learning. Second, while IBM can yield notable gains on environmental sound tasks, it tends to incur a trade-off with fine-grained discriminative tasks such as speaker identification, leading to degraded generalization. In contrast, DWM achieves more moderate yet consistent improvements across tasks, while showing less performance degradation on speech-related tasks. Finally, the ablation results highlight the importance of appropriately scheduling task difficulty via the hint ratio, demonstrating that controlled masking progression is critical for stable and generalizable representation learning.

V. Conclusion

In this work, we investigated the role of masking strategies in masked prediction-based self-supervised learning for general audio using spectrogram inputs. Through comparisons across multiple SSL frameworks and downstream tasks, we showed that masking strategies substantially influence both task performance and generalization.

Our analysis revealed that IBM substantially improves environmental sound performance under a limited-data condition, while reducing generalization, particularly to speech tasks. In our experiments, informed masking approaches such as SGIM exhibited non-negligible computational overhead. Motivated by these observations, we proposed DWM, a lightweight alternative that exploits dispersion characteristics inherent to audio spectrograms.

Experimental results demonstrated that DWM yields modest but consistent performance improvements while largely maintaining performance on speech tasks, thereby better preserving generalization across diverse downstream tasks. In addition, ablation studies highlighted the importance of scheduling the hint ratio to control task difficulty during training for stable and effective representation learning.

Overall, our findings provide empirical insights into masking strategy design and suggest that lightweight, spectrogram-aware masking can serve as a practical alternative to computationally intensive informed masking approaches. Future work includes exploring dispersion metrics beyond MAD that better capture semantically relevant content, evaluating DWM with varying patch sizes and larger-scale datasets, and extending comparisons to computationally intensive methods such as SGIM under matched budgets.

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Proc. CVPR*, Jun. 2022, pp. 16 000–16 009.

[2] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked Autoencoding Audio Spectrogram Transformer," in *Proc. Interspeech*, 2022, pp. 2438–2442.

[3] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Spectrogram Modeling using Masked Autoencoders for Learning General-purpose Audio Representation," in *Proc. HEAR: Holistic Evaluation of Audio Representations (NeurIPS 2021 Competition)*, vol. 166, 2022, pp. 1–24.

[4] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in *Proc. NeurIPS*, vol. 35, 2022, pp. 28 708–28 720.

[5] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. ICLR*, 2022.

[6] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *Proc. ICML*, 23–29 Jul 2023, pp. 1416–1429.

[7] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng, "SemMAE: Semantic-guided masking for learning masked autoencoders," in *Proc. NeurIPS*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 14 290–14 302.

[8] J. Shin, I. Lee, J. Lee, and J. Lee, "Self-guided masked autoencoder," in *Proc. NeurIPS*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 58 929–58 954.

[9] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *Proc. IJCAI*, 8 2024, pp. 3807–3815.

[10] K. Yamamoto and K. Okusa, "AaPE: Aliasing-aware patch embedding for self-supervised audio representation learning," *arXiv preprint arXiv:2512.03637*, 2025.

[11] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Modeling Duo: Towards a Universal Audio Pre-Training Framework," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2391–2406, 2024.

[12] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "CED: Consistent ensemble distillation for audio tagging," in *Proc. ICASSP*, 2024.

[13] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," in *Proc. AAAI*, vol. 36, No. 10, 2022, pp. 10 699–10 709.

[14] X. Li, N. Shao, and X. Li, "Self-Supervised Audio Teacher-Student Transformer for Both Clip-Level and Frame-Level Tasks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 1336–1351, 2024.

[15] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio Pre-Training with Acoustic Tokenizers," in *Proc. ICML*, 2023, pp. 5178–5193.

[16] D. Niizumi, D. Takeuchi, M. Yasuda, B. Thien Nguyen, Y. Ohishi, and N. Harada, "M2D-CLAP: Exploring general-purpose audio-language representations beyond clap," *IEEE Access*, vol. 13, pp. 163 313–163 330, 2025.

[17] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[18] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz, "Self-supervised transformers for unsupervised object discovery using normalized cut," in *Proc. CVPR*, Jun. 2022, pp. 14 523–14 533.

[19] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Exploring Pre-trained General-purpose Audio Representations," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, p. 137–151, 2023.

[20] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017, pp. 776–780.

[21] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. ACM-MM*, Oct. 2015, pp. 1015–1018.

[22] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM-MM*, Nov. 2014, pp. 1041–1044.

[23] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv preprint arXiv::1804.03209*, Apr. 2018.

[24] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[25] K. MacLean, "Voxforge," 2018, available: http://www.voxforge.org/home.

[26] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affective Comput.*, vol. 5, no. 4, 2014.

[27] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Speech Audio Process.*, vol. 10, no. 5, 2002.

[28] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proc. ICML*, 2017, pp. 1068–1077.

[29] J. Turian, J. Shier, G. Tzanetakis, K. McNally, and M. Henry, "One billion audio sounds from GPU-enabled modular synthesis," in *Proc. DAFx2020*, Sep. 2021, pp. 222–229.