ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888568

Integrating Audio Narrations to Strengthen Domain Generalization in Multimodal First-Person Action Recognition

Cagri Gungor and Adriana Kovashka

University of Pittsburgh

https://cagrigungor.github.io/AudioVisualDG/

ABSTRACT

First-person activity recognition is rapidly growing due to the widespread use of wearable cameras but faces challenges from domain shifts across different environments, such as varying objects or background scenes. We propose a multimodal framework that improves domain generalization by integrating motion, audio, and appearance features. Key contributions include analyzing the resilience of audio and motion features to domain shifts, using audio narrations for enhanced audio-text alignment, and applying consistency ratings between audio and visual narrations to optimize the impact of audio in recognition during training. Our approach achieves state-of-the-art performance on the ARGO1M dataset, effectively generalizing across unseen scenarios and locations.

Index Terms— Action recognition, multimodal domain generalization, audio descriptions, multimodal fusion

1. INTRODUCTION

With the growing prevalence of wearable technology and first-person cameras, first-person activity recognition has emerged as a crucial area of research [1, 2, 3, 4, 5]. This field is vital for real-world egocentric vision applications, from human-robot interaction to personalized assistance. However, a significant challenge in developing robust action recognition models is the issue of domain shifts—variations in environmental contexts, objects, and activities that can drastically affect the performance of these models. Some works [6, 7, 8, 9, 10] rely solely on visual features to generalize across different domains. In this paper, we explore the potential of multimodality, specifically the integration of motion and audio with appearance, to enhance domain generalization in first-person action recognition tasks.

Our intuition is that domain shifts arise heavily from variations in the spatial semantics of videos, such as changes in the type of objects or the background, which we refer to as the appearance modality. These variations can lead to a drop in model performance when applied to new, unseen data. However, the motion and audio modalities capture temporal dynamics that remain more consistent across different domains.

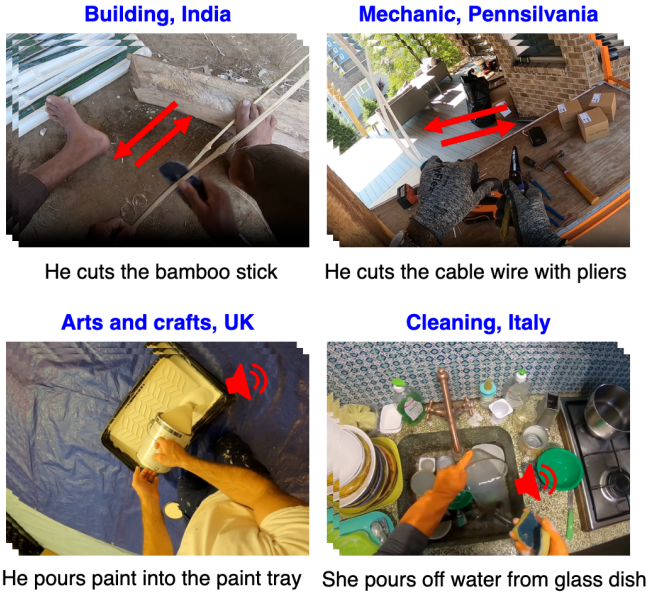

Fig. 1. Illustration of motion and audio resilience to domain shifts compared to appearance. While the motion of 'cutting' (first row) and audio of 'pouring' (second row) remain similar across different scenario-location domains, the appearance varies significantly with different objects and backgrounds.

For example, while objects being 'cut' in different domains, such as bamboo stick with knife (in building in India scenario) or cable wire with pliers (in the mechanic in PA) are visually different, the motion involved in cutting, characterized by the repetitive back-and-forth movement, remains consistent (Fig. 1 top). Similarly, 'mixing' motions are consistent whether stirring ingredients or mixing cement. Audio also provides strong temporal cues less affected by visual changes. The sound of 'pouring', whether paint into a tray or water from glass, remains consistent (Fig. 1 bottom). Despite visual variations, the consistent motion patterns and audio cues reliably indicate the action and are robust to domain shifts. Our analysis validates this intuition, showing lower performance drops of motion (25.8%) and audio (32.7%) compared to appearance (54.8%) in unseen domains.

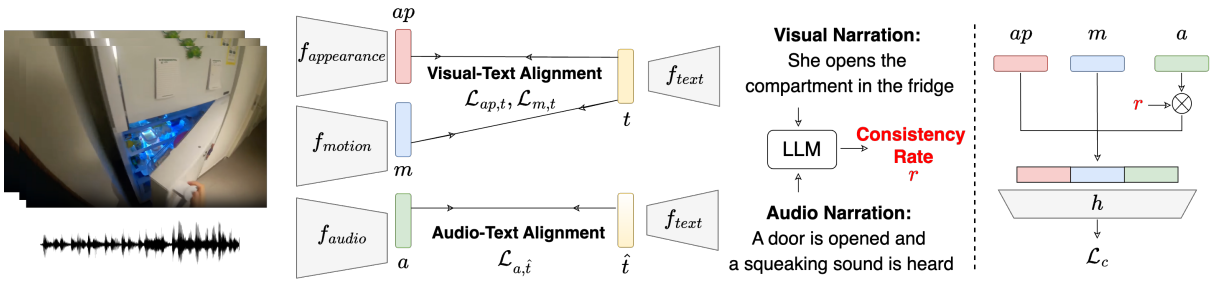

Fig. 2. The proposed framework extracts appearance ap_i, motion m_i, and audio a_i embeddings using trained encoders f. Visual-text and audio-text alignments are performed independently to enhance the robustness of action representations. Consistency rating r_i, calculated offline using a LLM [11], is then multiplied by audio embedding, optimizing the influence of audio in multimodal prediction. Note that narrations and consistency rate are only utilized during training to improve representation learning. During inference, embeddings are directly fused before prediction.

In addition to leveraging motion and audio, our research incorporates the text through narrations. Aligning *visual* features with corresponding textual descriptions (e.g., narrations of actions) has been shown to improve feature representation [12, 10, 13]. However, existing narrations are primarily based on visual content (e.g., in [10]'s dataset, many videos lack audio but all include narrations). Aligning *audio* features with such narrations can introduce noise due to inconsistencies, as actions in videos may not always produce matching audio cues. To address this, we use an audio captioner [14] to generate audio-specific narrations, aligning appearance and motion with visual narrations, and audio with its respective narrations. Additionally, we compute consistency ratings between visual and audio narrations, indicating how well audio and video express the same action. Attention weighing (Fig. 2) is applied, reducing the influence of audio in predictions when consistency is low. These methods reduce noise, enhance generalization, and improve representation learning.

Prior work in multimodal domain generalization for action recognition includes RNA-Net [15], which balances audio and video feature norms, and SimMMDG [16], which employs contrastive learning and cross-modal translation to separate modality-specific and shared features. In contrast, our approach emphasizes the resilience of audio and motion features to domain shifts, showcasing their pivotal role in achieving robust generalization. CIR [10] introduces ARGO1M to test generalization across unseen scenarios and locations, using cross-instance reconstruction with text guidance to align visual narrations with appearance. Differently, we align audio with audio-based narrations, enhancing representation robustness. While [17] uses an LLM to generate audio-centric narrations from visual ones, we instead use an audio captioner [14] to derive narrations directly from audio and calculate a consistency rate to assess how well audio aligns with video.

To summarize, our contributions are: (1) a multimodal framework integrating motion, audio, and appearance features, achieving state-of-the-art domain generalization in first-person activity recognition on the ARGO1M; (2) a detailed analysis demonstrating the resilience of motion and audio features to domain shifts, underscoring their critical role in generalization; (3) the alignment of audio narrations with audio features to enhance action representation robustness; and (4) the use of consistency ratings between audio and visual narrations to optimize audio influence in predictions.

2. METHOD

2.1. Proposed Multimodal Setting

Each training sample consists of a video clip paired with corresponding audio, visual narration and audio narration. The former is provided in the dataset we use [10], while we use Pengi [14] to generate the latter. While we adopt the approach from [10] of using separate frozen encoders for each modality to extract base features, our method differs by incorporating the audio modality and utilizing distinct encoders for each modality, instead of a single encoder for fused appearance and motion features. Specifically, we train (from scratch) separate encoders $f_{appearance}$, f_{motion}, and f_{audio} to derive domain-generalizable features ap, m, and a for appearance, motion, and audio. Additionally, we train a separate encoder f_{text} extracts the features of visual narration t and audio narration $\hat{t}$. See Fig. 2. The action prediction $\hat{y}$ is generated by the classifier h based on the fused multimodal embedding. The cross-entropy loss $\mathcal{L}_c$ is computed between the true and predicted action labels, y and $\hat{y}$.

2.2. Consistency Ratings to Enhance Fusion

Before training, we calculate consistency ratings for each video sample using a LLM [11], assessing the degree to which the audio and visual narrations correspond semantically. This information is used to control fusion. Computing consistency between audio and visual content through text captures abstract concepts that raw features might miss. While features emphasize low-level details, text-based evalu-

ations ensure that the audio and visual content correspond at a deeper, conceptual level.

We use the following prompt obtain consistency ratings:

> Rate the consistency between two narrations from the same video out of 100. The first narration describes the visual aspect, and the second describes the audio. Consider how well the audio narration overlaps with and complements the visual narration. Output only the percentage score.

The consistency ratings r are then used as weights to modulate the contribution of the audio embeddings a before the concatenation of appearance, motion, and audio features, as shown in Fig. 2. Specifically, only during training, the audio embeddings are scaled by the consistency rating r, such that $a = a * r$. This consistency-weighted audio approach ensures that audio information with strong semantic alignment to the visual content exerts a greater influence on the final prediction. This approach enhances the quality of audio representations during training, minimizing the impact of noisy or irrelevant audio cues and improving the overall robustness.

2.3. Text Guided Alignment

Aligning visual features with narrations enriches the model with domain-invariant, human-like understanding, enhancing its ability to generalize across domains [13]. As discussed previously, we generate *audio* narrations that are specifically tied to the audio content. The LLM in audio captioner Pengi [14] mimics human-like understanding, providing audio-specific, semantically rich descriptions. Since Pengi uses a separate encoder for audio feature extraction, its generated audio narrations are initially not aligned with our model's audio features a. Thus, in our approach, we align audio features a with these audio narration features $\hat{t}$, while aligning appearance features ap and motion features m with visual narration features t using contrastive learning.

Given a batch of samples $\mathcal{B} = \{(ap_i, m_i, a_i, t_i, \hat{t}_i)\}_{i=1}^{B}$, we frame the alignment as noise contrastive estimation [18]. Specifically, $\mathcal{L}_{ap\to t}$ treats the appearance ap_i as the anchor, with other narrative texts serving as negatives, and minimizes:

$$\mathcal{L}_{ap\to t} = -\frac{1}{|\mathcal{B}|}\sum_{i}^{B} log\frac{exp(s(ap_i, t_i)/\tau)}{\sum_j^B exp(s(ap_j, t_j)/\tau)} \quad (1)$$

where $s(\cdot,\cdot)$ is the cosine similarity and τ is a learnable temperature. In a similar manner, $\mathcal{L}_{t\to ap}$ uses the text t_i as the anchor, with other appearance features as negatives. These losses are then combined to create the appearance-text alignment loss $\mathcal{L}_{ap,t} = \mathcal{L}_{t\to ap} + \mathcal{L}_{ap\to t}$. Likewise, motion-text alignment $\mathcal{L}_{m,t}$ and audio-text alignment $\mathcal{L}_{a,\hat{t}}$ are computed, and all these alignment losses are aggregated into a total alignment loss $\mathcal{L}_{align} = \mathcal{L}_{ap,t} + \mathcal{L}_{m,t} + \mathcal{L}_{a,\hat{t}}$.

To form the overall training objective, we combine the alignment loss with the cross-entropy classification loss:

$$\mathcal{L} = \mathcal{L}_c + \lambda\mathcal{L}_{align} \quad (2)$$

where $\lambda = 0.1$ is used to weight the alignment loss.

Modality Setting	Me SAU	Co JPN	Ar ITA	Sh IND	Cl US-MN	Mean
Audio*	27.6	42.9	30.4	29.3	28.7	31.8
Audio	25.8 (-6.9%)	23.1 (-85.7%)	23.0 (-32.1%)	24.9 (-21.6%)	24.5 (-17.1%)	24.3 (-32.7%)
Motion*	27.0	29.3	32.0	29.6	26.2	28.9
Motion	25.9 (-4.2%)	19.1 (-53.4%)	21.9 (-46.1%)	26.8 (-10.4%)	22.8 (-14.9%)	23.3 (-25.8%)
Appearance*	34.9	56.4	51.2	43.0	39.2	44.9
Appearance	31.6 (-10.4%)	26.6 (-112.0%)	29.3 (-74.7%)	31.1 (-38.2%)	28.7 (-38.2%)	29.5 (-54.8%)
Multimodal*	36.5	59.0	52.5	44.9	40.4	46.7
Multimodal	34.7 (-5.1%)	30.7 (-92.1%)	31.8 (-65.1%)	34.2 (-31.2%)	32.4 (-24.7%)	32.7 (-42.8%)

Table 1. Percentages show performance drop when the training set excludes samples from the test domain (domain noted in column header), vs when it includes them (*). Audio and motion exhibit less drop vs appearance, highlighting their resilience to shifts. The baseline (cross-entropy loss) is used.

3. EXPERIMENTS

Implementation details. We leverage the frozen pretrained SlowFast model [19], utilizing its slow pathway for capturing appearance features and its fast pathway for extracting motion features. For audio features, we employ BEATs [20], while CLIP-ViT-B-32 [12] is used to extract text features. The trained encoders f each consist of two fully connected layers, each followed by ReLU activation and Batch Normalization [21]. We use batch size 128 for 50 epochs with the Adam optimizer [22]. The learning rate is initially set to 2e-4, with decay factor 10 applied at epochs 30 and 40.

Dataset. ARGO1M [10] is an egocentric video dataset curated from Ego4D [23], specifically designed to analyze scenario and location-based domain shifts. The dataset includes 10 distinct train-test splits, ensuring that the domains of the samples—both scenario and location—do not overlap between the training and test sets. The test set scenarios and geographic locations include Gardening in Pennsylvania (Ga, US-PNA), Cleaning in Minnesota (Cl, US-MN), Knitting in India (Kn, IND), Shopping in India (Sh, IND), Building in Pennsylvania (Bu, US-PNA), Mechanic in Saudi Arabia (Me, SAU), Sport in Colombia (Sp, COL), Cooking in Japan (Co, JPN), Arts and Crafts in Italy (Ar, ITA), and Playing in Indiana (Pl, US-IN). ARGO1M contains 1,050,371 video clips. However, due to the absence of audio in some clips, we utilized approximately 60% of the total dataset where audio is available. Test sets are fairly large (more than 10,000 samples for each domain) indicating results are reliable. We report top-1 accuracy for each test split, along with the mean accuracy across the splits.

3.1. Audio and Motion Resilience to Domain Shifts

In Table 1, the mean percentage drops in parentheses indicate the extent of domain shifts across various domains,

Method	Me SAU	Co JPN	Ar ITA	Sh IND	Cl US-MN	Pl US-IN	Sp COL	Ga US-PNA	Bu US-PNA	Kn IND	Mean
Baseline	34.7	30.7	31.8	34.2	32.4	36.1	33.8	34.6	33.9	28.9	33.1
CIR [10]	35.6	**32.4**	32.4	35.4	33.1	38.2	34.7	36.1	35.4	30.4	34.3
Ours	**36.0**	32.2	**32.7**	**36.0**	**33.5**	**38.3**	**35.1**	**36.5**	**35.8**	**30.6**	**34.7**

Table 2. A comparison with the state of the art on ARGO1M using appearance, motion, and audio.

Method	Modality Setting	Me SAU	Co JPN	Ar ITA	Sh IND	Cl US-MN	Mean
Baseline	Ap	31.6	26.6	29.3	31.1	28.7	29.5
Baseline	Ap-Mo	32.7	27.5	30.0	31.3	29.5	30.2
Baseline	Ap, Mo	33.2	27.8	30.8	31.7	29.7	30.6
Baseline	Ap, Mo, Au	**34.7**	**30.7**	**31.8**	**34.2**	**32.4**	**32.7**

Table 3. Impact of combining different modalities—**Ap**pearance, **Mo**tion, and **Au**dio—across domains.

Method	Me SAU	Co JPN	Ar ITA	Sh IND	Cl US-MN	Mean
Baseline (B)	34.7	30.7	31.8	34.2	32.4	32.7
$B+$ weighted a by r	34.9	30.9	31.9	34.4	32.4	32.9
$B + \mathcal{L}_{vt} + \mathcal{L}_{mt} + \mathcal{L}_{at}$	35.4	31.6	32.3	35.2	32.9	33.5
$B + \mathcal{L}_{vt} + \mathcal{L}_{mt} + \mathcal{L}_{a\hat{t}}$	35.8	31.9	32.6	35.6	33.3	33.8
Ours	**36.0**	**32.2**	**32.7**	**36.0**	**33.5**	**34.1**

Table 4. Ablation showing the impact of alignment losses and consistency-weighted audio. All methods use (Ap, Mo, Au).

comparing modality settings where the training set either contains samples that share location and/or scenario with the test domain, or share neither. The mean performance drop for motion features is 25.8%, while audio exhibits a drop of 32.7%. In contrast, appearance shows a significantly higher shift (54.8%). This aligns with our hypothesis that temporal dynamics—such as consistent patterns of movement and continuity of sound—remain more stable across different environments and scenarios. In contrast, spatial semantics represented by appearance features, are more variable due to differences in objects, backgrounds, and other visual elements that can vary significantly from one domain to another. Moreover, the multimodal approach, which integrates audio, motion, and appearance, demonstrates a reduced shift with a mean drop of 42.8% compared to appearance alone (54.8%). Our analysis underscores the critical role of incorporating audio and motion features for robust domain generalization.

3.2. Comparison with State of the Art

Table 2 compares our proposed method, the state-of-the-art CIR [10] approach, and a baseline model on the ARGO1M dataset. The baseline is trained solely with cross-entropy loss, whereas CIR enhances its performance through cross-instance reconstruction guided by text. Notably, although the original CIR results did not include audio, we have extended their approach by incorporating audio features and multiple trained encoders f for each modality to ensure a fair comparison with our method. Our method consistently outperforms CIR across all domain splits, except for Co-JPN, achieving up to a 1.7% improvement and a 1.2% higher average performance overall. Additionally, our method outperforms the baseline by an average of 4.8%.

3.3. Ablations

Table 3 illustrates the impact of different modality settings on performance. In the 'Ap-Mo' setting used by CIR [10], appearance and motion features are fused early and processed through a single encoder. We propose an alternative approach in the 'Ap, Mo' setting, where separate encoders ($f_{appearance}$, f_{motion}) are employed to learn distinct domain-generalizable features, particularly because motion exhibits greater resistance to domain shifts. As a result, our 'Ap, Mo' outperforms 'Ap-Mo'. Furthermore, the integration of audio with separate encoders in 'Ap, Mo, Au' yields the best overall performance.

Table 4 evaluates the effectiveness of each component of our approach. Starting with the baseline (trained using only cross-entropy), we observe that weighting the audio embeddings a by the consistency ratings r enhances the model's performance, owing to the more reliable audio representations. Aligning all modalities with narration ('$B+\mathcal{L}_{vt}+\mathcal{L}_{mt}+\mathcal{L}_{at}$') further improves performance. Notably, when we align audio features with their corresponding audio narrations $\hat{t}$ using $\mathcal{L}_{a\hat{t}}$ the model outperforms the version with $\mathcal{L}_{at}$ where audio features are aligned with the original visual-based narrations t. Finally, the combination of alignment losses and the consistency-weighted audio approach in our final method 'Ours' achieves the best performance across all domains.

Conclusion. We proposed a novel multimodal framework where motion, audio, and appearance improve domain generalization in first-person action recognition. We demonstrated that audio and motion features exhibit greater resilience to domain shifts than appearance features. By aligning audio features with audio-specific narrations and applying consistency-weighted audio during training, our method enhanced the robustness of action representations.

Acknowledgement: This work was supported by a National Science Foundation Award No. 2046853.

4. REFERENCES

[1] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *ICCV*, 2019, pp. 5492–5501. 1

[2] Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan, "Mmg-ego4d: Multimodal generalization in egocentric action recognition," in *CVPR*, 2023, pp. 6481–6491. 1

[3] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar, "Learning video representations from large language models," in *CVPR*, 2023, pp. 6586–6597. 1

[4] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang, "Interactive prototype learning for egocentric action recognition," in *ICCV*, 2021, pp. 8168–8177. 1

[5] Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo, "E2 (go) motion: Motion augmented event stream for egocentric action recognition," in *CVPR*, 2022, pp. 19935–19947. 1

[6] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales, "Deeper, broader and artier domain generalization," in *ICCV*, 2017, pp. 5542–5550. 1

[7] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf, "Domain generalization via invariant feature representation," in *ICML*, 2013, pp. 10–18. 1

[8] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in *ICML*, 2020, pp. 7728–7738. 1

[9] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018. 1

[10] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen, "What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations," in *ICCV*, 2023, pp. 13656–13666. 1, 2, 3, 4

[11] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763. 2, 3

[13] Seonwoo Min, Nokyung Park, Siwon Kim, Seunghyun Park, and Jinkyu Kim, "Grounding visual representations with texts for domain generalization," in *ECCV*, 2022, pp. 37–53. 2, 3

[14] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang, "Pengi: An audio language model for audio tasks," *NeurIPS*, vol. 36, pp. 18090–18108, 2023. 2, 3

[15] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo, "Domain generalization through audio-visual relative norm alignment in first person action recognition," in *WACV*, 2022, pp. 1807–1818. 2

[16] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink, "Simmmdg: A simple and effective framework for multi-modal domain generalization," *NeurIPS*, vol. 36, 2024. 2

[17] Andreea-Maria Oncescu, João F Henriques, Andrew Zisserman, Samuel Albanie, and A Sophia Koepke, "A sound approach: Using large language models to generate audio descriptions for egocentric text-audio retrieval," in *ICASSP*, 2024, pp. 7300–7304. 2

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 3

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *ICCV*, 2019, pp. 6202–6211. 3

[20] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei, "Beats: Audio pre-training with acoustic tokenizers," in *ICML*, 2023. 3

[21] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015. 3

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015. 3

[23] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al., "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022, pp. 18995–19012. 3