Table 4: **Action Recognition on UCF101 and HMDB51 and Audio Classification on ESC50.** We report action recognition accuracy after full fine-tuning and linear probe evaluation. We indicate the pre-training dataset, resolution, the number of frames, iterations (or epochs in brackets), and pre-training data modalities (V=RGB, A=audio).

							UCF101		**HMDB51**		**ESC50**
Method	**Dataset**	**Res.**	**Frames**	**It. [Ep.]**	**Network**	**Mod.**	FT	Lin.	FT	Lin.	Lin.
TE-CVRL (**?**)	K400	112	16	[200]	R(2+1)D-18	V	88.2		62.2		
CVRL (**?**)	K600	224	32	[800]	R3D-50	V	93.4	90.6	68.0	59.7	
MMV (**?**)	AS	224	32	500K	R(2+1)D-18	V+A	91.5	83.9	70.1	60.0	
BraVe (**?**)	AS	224	32	620K	R(2+1)D-18	V+A	**93.6**	90.0	70.8	63.6	
AVTS (**?**)	K400	224	25	[90]	MC3	V+A	85.8		56.9		76.7
XDC (**?**)	K400	224	32	900K	R(2+1)D-18	V+A	84.2		47.1		78.5
GDT (**?**)	K400	112	32	[200]	R(2+1)D-18	V+A	88.7		57.8		78.6
AVID (**?**)	K400	224	32	[400]	R(2+1)D-18	V+A	87.5		60.8		79.1
Ours	VGG-S	112	16	160K [240]	R(2+1)D-18	V+A	90.9	86.8	70.2	55.9	**87.9**
Ours	K400	112	16	200K [240]	R(2+1)D-18	V+A	91.8	88.0	71.2	58.2	84.8
Ours	K600	112	16	200K [150]	R(2+1)D-18	V+A	92.2	90.3	72.2	62.6	86.4
Ours	K600	224	16	400K [300]	R3D-34	V+A	**93.6**	**91.8**	**74.6**	**65.8**	85.5

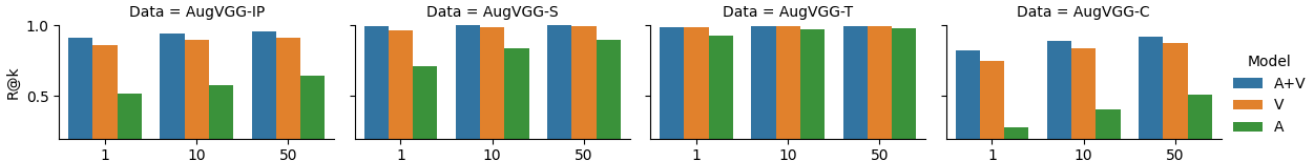

Figure 3: **Video Fingerprinting Performance.** We report instance retrieval performance under video content manipulation on the different AugVGG variants. We show results using a video only (V), audio only (A), and a joint audio-visual model (A+V).

Table 5: **Video Retrieval on UCF101 and HMDB51.** We report recall at k (R@k) for k-NN video retrieval. All methods use a R(2+1)D-18 network.

	UCF101			**HMDB51**		
Method	**R@1**	**R@5**	**R@20**	**R@1**	**R@5**	**R@20**
TCLR	56.9	72.2	84.6	24.1	45.8	75.3
GDT	57.4	73.4	88.1	25.4	51.4	75.0
Robust-xID	60.9	79.4	90.8	30.8	55.8	79.7
TE-CVRL	64.2	81.1	92.6	33.1	60.8	84.1
Ours (R(2+1)D-18)	**80.6**	**90.4**	**96.4**	**44.9**	**70.4**	**87.6**
Ours (R3D-34)	**85.2**	**93.0**	**97.3**	**51.3**	**74.3**	**91.4**

novel contrastive loss design and a model with both intra- and cross-modal contrastive objectives to learn from the audio-visual correspondence in videos. Experiments demonstrate that representations that integrate both temporal and aural features achieve state-of-the-art video classification and retrieval performance.

Dolorem ea laudantium libero, impedit excepturi voluptatum maxime cupiditate illum non harum maiores atque, veniam corrupti fuga fugit excepturi, odit amet iusto fuga neque eaque autem ex rem veritatis blanditiis expedita, repellat aut quisquam?Numquam officia vero dignissimos, perspiciatis eos perferendis molestiae libero eaque fugit impedit architecto quas magni, veniam cum enim nulla dolor reiciendis eius asperiores doloremque, odit accusamus ex nostrum aspernatur earum alias sequi, amet eos blanditiis magnam animi libero quidem sunt molestiae vero?Voluptatem voluptatibus ratione a soluta vitae harum voluptatum quod, nesciunt nihil id nobis incidunt temporibus saepe facere nemo quo esse voluptate?Saepe iure eum architecto libero error aspernatur exercitationem eligendi animi quod, quia dolorum mollitia error porro autem enim, ratione omnis voluptatibus, ipsa error doloribus ipsum assumenda distinctio numquam libero.Doloribus nihil ducimus sint sapiente repellendus, consequatur iste facere voluptates, magni neque est molestiae aut alias porro praesentium, ea suscipit sequi, cumque provident vero veritatis deserunt repellat reprehenderit numquam possimus consequatur quod officiis.Soluta odit a deleniti unde cumque ab

Table 6: **Modality Fusion.** We explore the fusion of our audio-visual features for downstream video classification.

Modalities	**VGG-Sound**	**K600**
Audio	39.1	15.7
Video	39.7	56.8
Audio+Video	**53.9**	**58.4**

enim inventore quibusdam, animi dolore ratione facere odio dolorum enim eligendi blanditiis accusamus libero, eveniet vel laudantium quod culpa odio, non