
How Much Physics Can Vision Learn from Audio?

Probing and Distilling Damping Structure

Zainab Afolabi
Amazon Web Services
zafolabi@amazon.co.uk

Vadim Selyutin
Amazon Web Services
seliutin@amazon.ae

Zhao Gao
Amazon Web Services
orangez@amazon.co.uk

Babs Khalidson
Amazon Web Services
khalbabs@amazon.co.uk

Mohamed Ghoneim
Amazon Web Services
ghoneimm@amazon.co.uk

Abstract

Physical understanding requires inferring properties, such as damping, that govern material behavior, yet vision models struggle with such latent structure, while audio encodes it directly through measurable acoustic decay. We probe frozen encoders for the quality factor Q , a dimensionless measure of resonance, and find that audio embeddings predict Q at Spearman $\rho = 0.73$, while a single CLIP frame reaches only $\rho = 0.28$. To transfer this structure, we distill a frozen audio teacher into a frozen CLIP encoder through a lightweight projection head. Under a leave-one-material-out protocol in which category recognition alone scores zero, four-frame input raises correlation from 0.23 to 0.29, and distillation raises it further to 0.34 (+0.05 over the matched four-frame baseline, 14 of 17 materials). A single-frame distilled student gains mainly in the damping term rather than pitch. Pooled controls attribute most of the distillation gain to acoustic correspondence rather than to added capacity or category recognition.

1 Introduction

A robot clearing a table observes two cups that look alike, one ceramic and one plastic. A vision model might correctly label their materials, but this categorical distinction does not reveal how they behave, since the ceramic cup shatters when dropped while the plastic cup bounces. Without physical grounding, vision cannot predict these divergent futures. This *appearance–reality gap* pervades modern computer vision. Large vision models achieve strong recognition performance, but what they encode is how things look, not how things behave. Physics benchmarks confirm this gap: PhysBench [2] reports GPT-4o at 47% on physical dynamics (49.5% overall, against 95.9% for humans), and Physion++ [15] shows all models fall short of human performance on latent properties like mass, friction, or elasticity.

Audio captures what appearance cannot. An object’s acoustic response to impact is governed by physical laws: stiffness, density, and geometry set the resonant frequency, and internal friction, radiation, and mounting set the damping. We quantify this gap with the quality factor Q , which counts how many oscillations a struck object completes before its ring dies away, and the Spearman rank correlation ρ : $\rho = 1$ means the model orders impacts by Q perfectly, and $\rho = 0$ means its ordering bears no monotonic relation to Q .

Our question is whether this physical structure can transfer from audio to vision. We extract the quality factor Q from 17,958 impacts as a benchmark; show that, on held-out materials, four-frame input raises Q prediction from 0.23 to 0.29 and distilling from a frozen audio teacher raises it further to 0.34; show that a single-frame distilled student gains mainly in the damping term β rather than

Table 1: Spearman ρ between ridge-probe predictions from frozen embeddings and ground-truth Q . 95% CIs are given in Tables 6 and 8.

Model	Modality	Pooled ρ	LOMO ρ
PE-AV	Audio	0.727	0.600
CAV-MAE	Audio	0.701	0.595
PE-AV	Joint	0.696	0.572
CAV-MAE	Joint	0.608	0.492
PE-AV	Video	0.362	0.197
CLIP	Video	0.281	0.227
DINOv2	Video	0.271	0.239
CAV-MAE	Video	0.151	0.206

pitch; and introduce an evaluation protocol—leave-one-material-out (LOMO) evaluation, complemented by a shuffled-teacher control for incidental structure—that separates physical transfer from category leakage.

2 Related Work

That impact sounds carry material information is established: Owens et al. [12] built the Greatest Hits dataset on it, RealImpact [3] adds calibrated measurements, and in robotics, audio often beats vision for physical property estimation [8, 14]. Audio-visual representation learning has meanwhile moved from contrastive objectives [1, 11] to masked autoencoders [4, 7] and large-scale models such as PE-AV [17], which we use as our teacher. These models are evaluated on retrieval and classification; we add a physical evaluation axis by probing for a continuous, physically defined quantity and testing generalization to unseen materials. Cross-modal distillation dates back to Gupta et al. [5], who transferred supervision from labeled RGB images to depth and optical-flow representations, and knowledge distillation more generally to Hinton et al. [6]. We differ in targeting specific physically relevant structure rather than general-purpose representations, and we contribute an evaluation protocol (LOMO, complemented by a shuffled-teacher control for incidental structure) that separates physical transfer from category leakage.

3 The Quality Factor as a Physical Target

The quality factor $Q = \pi f_0 / \beta$ combines resonant frequency (f_0) and damping rate (β) into a dimensionless scalar counting oscillation cycles before decay. Two properties make it a useful target. Being dimensionless, it is comparable across objects regardless of size or pitch. More importantly, it varies continuously *within* material categories: a thin metal sheet and a solid metal block are both “metal” yet differ substantially in Q . Predicting it therefore demands more than category recognition.

We extract Q from Greatest Hits [12], which contains 977 videos of drumstick impacts across 17 materials (28,444 impacts in total). For each impact, we compute the Hilbert envelope of the audio waveform and fit an exponential decay $A(t) = A_0 e^{-\beta t} + C$ to obtain β , take the dominant spectral peak above 50 Hz as f_0 , and retain fits with $R^2 \geq 0.8$. This yields 17,958 usable impacts (63%), with rejections concentrated on granular surfaces, where no single resonant mode exists. Extracted values match physical expectation: median Q rises from cloth (15) through wood (55) and plastic (89) to metal (134), ceramic (327), and glass (412).

4 Physical Structure in Frozen Embeddings

We probe four frozen encoders with ridge regression ($\lambda = 1.0$) on $\log(Q + 1)$. We use Spearman ρ because Q spans two orders of magnitude with a heavy tail (ceramic and glass at roughly 330–410, cloth at 15), and because ranking physical properties matters more than predicting exact values. PE-AV [17] and CAV-MAE [4] are audio-visual models with separate audio, video, and joint embeddings. CLIP ViT-L/14 [13] and DINOv2 ViT-L/14 [10] are vision-only baselines. Video inputs are the frame at impact; audio inputs are 2 s clips centered on the impact.

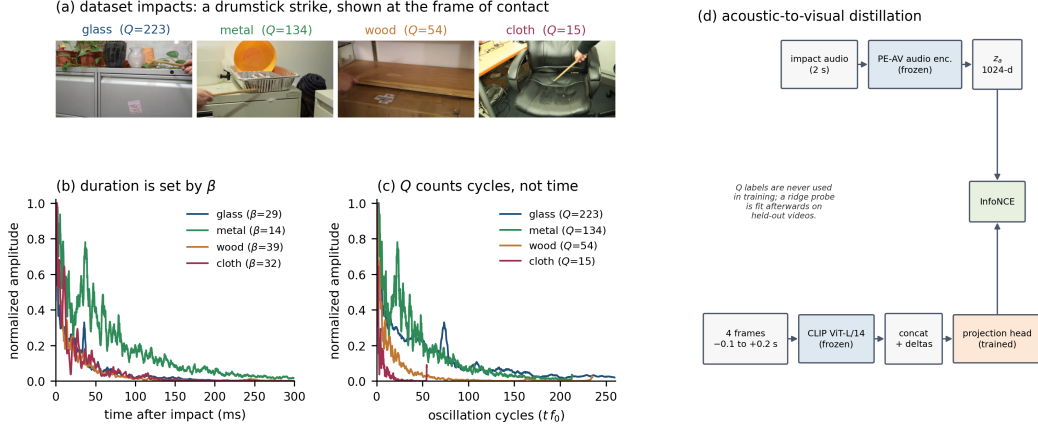


Figure 1: Method. (a) Example impacts (individual values, not medians), from glass ($Q = 223$) to cloth (15). (b–c) Hilbert envelopes: Q measures cycles before decay, not absolute duration. (d) Both encoders stay frozen, and only the projection head is trained, with an InfoNCE loss.

Table 1 shows the gap: audio ($\rho = 0.73$) far exceeds vision (CLIP $\rho = 0.28$; best visual branch 0.36). Co-training on audio does not by itself place that structure in the visual pathway: PE-AV’s video branch reaches 0.36 and CAV-MAE’s only 0.15, both far below their own audio branches.

Under LOMO, where the probe is trained on 16 materials and evaluated on the held-out 17th, audio encoders reach $\rho \approx 0.60$, against 0.20–0.24 for vision. LOMO ρ is the within-material Spearman correlation, averaged over the 17 held-out materials (constant predictions within a material score 0). A probe that never saw the held-out material still ranks its impacts by Q from audio embeddings.

Within a single material category, where category identity supplies nothing, the separation sharpens: PE-AV audio reaches $\rho = 0.67$ on metal and 0.57 on wood, against CLIP’s 0.06 and 0.03 (Appendix A).

5 Distilling Physical Structure to Vision

We distill from frozen PE-AV audio (teacher) into frozen CLIP ViT-L/14 (student), training only a three-layer MLP projection head (GELU, LayerNorm, 1024-dimensional output); Figure 1 illustrates both the quantity and the setup. The visual input is four frames at $-0.1, 0, +0.1$, and $+0.2$ s around the impact, each encoded independently and concatenated with their successive differences (“four frames + deltas”). Training uses a symmetric InfoNCE loss [16] ($\tau = 0.07$), batch size 256, and AdamW [9] with learning rate 10^{-4} . No Q labels enter training. We select the feature variant and epoch count on a validation split (20% of training videos), then evaluate on the test split once, reporting three seeds.

A data-handling detail matters: we train the head only on the 780 training-split videos (22,793 of 28,444 impacts). The teacher predicts Q at $\rho = 0.73$, so any impact seen during distillation has effectively had its target leaked, even without explicit Q labels.

Before adding audio, we check whether single-frame input is the bottleneck. Damping concerns how quickly motion subsides after impact, which is inherently temporal. Widening from one to four frames raises LOMO ρ from 0.234 to 0.285 (+0.051) with no acoustic supervision, roughly half of our eventual improvement.

To test physical transfer rather than category recognition, the held-out material must be excluded from *both* the projection head and the probe; we call this protocol *true LOMO*. A model outputting category medians scores zero by construction under this protocol, which makes it the discriminating test.

Table 2 shows the result. Distillation improves LOMO ρ from 0.234 to 0.338 (+0.104; 15/17 materials improve; Wilcoxon $p = 0.002$) and by +0.053 over the four-frame baseline (14/17; $p = 0.001$). Gains are uneven, and the five largest (single-frame $\rightarrow$ distilled) are on granular and natural surfaces: gravel 0.36 $\rightarrow$ 0.71, grass 0.38 $\rightarrow$ 0.61, dirt 0.29 $\rightarrow$ 0.61, leaf 0.29 $\rightarrow$ 0.54, and rock 0.25 $\rightarrow$ 0.47.

Table 2: True LOMO: the held-out material is excluded from both the projection head and the probe. Mean over 17 materials and three seeds, with 95% CIs bootstrapped over materials.

Model	LOMO ρ	95% CI	vs. baseline
Baseline, single frame	0.234	[0.175, 0.289]	—
Baseline, four frames + deltas	0.285	[0.200, 0.372]	+0.051
Distilled, four frames + deltas	0.338	[0.249, 0.430]	+0.104
Teacher (PE-AV audio)	0.600	[0.528, 0.658]	—

Baselines are recomputed in the pipeline used for distillation; the single-frame baseline’s frozen-probe value in Table 1 is 0.227.

Table 3: Q decomposed into its spectral (f_0) and damping (β) terms. Audio embeddings predict β better than f_0 , and the margin widens on unseen materials. The distilled row is a single-frame student (the frame at impact), directly comparable with the CLIP row; its LOMO columns use true LOMO (held-out material excluded from both the projection head and the probe). Bold marks the better of f_0 and β within each row and setting.

Model	Pooled ρ		LOMO ρ	
	f_0	β	f_0	β
PE-AV (audio)	0.700	0.729	0.485	0.686
CAV-MAE (audio)	0.667	0.732	0.423	0.719
CLIP	0.347	0.221	0.137	0.257
DINOv2	0.353	0.210	0.152	0.244
CLIP distilled	0.345	0.311	0.177	0.375

Three pooled controls bound what this improvement can mean (Table 7, Appendix C). First, category recognition: mapping the true material label to its per-category median Q reaches pooled $\rho = 0.430$, and CLIP’s own material classifier (45.6% accuracy), mapped the same way, reaches 0.279, so pooled correlations near 0.28 are indistinguishable from category recognition. Second, capacity: on single-frame features, an MLP head of the same shape given direct supervision on $\log(Q + 1)$, with no teacher, reaches only 0.275, below the 0.300 of a linear probe, so head capacity alone does not explain the gain. Third, correspondence: holding the seed fixed (the first, and highest, of the three seeds), distilling from correctly matched audio reaches 0.466, against 0.298 with no distillation, while a fully shuffled teacher (10.6% material agreement) still reaches 0.359. Roughly a third of the raw gain therefore survives destroying the correspondence, so reported gains from cross-modal distillation should be read against a shuffled-teacher control.

5.1 Damping, Not Pitch, Transfers

A high ρ on Q alone would not establish that anything physical had transferred. Across the 17,958 impacts, Q is strongly rank-correlated with f_0 ($\rho = 0.87$) and moderately anti-correlated with β ($\rho = -0.54$), while the two components are close to independent ($\rho = -0.11$), so an encoder that merely recovered pitch would score well on Q . Table 3 probes each target separately. For both audio encoders, the damping term is the better predicted, and its advantage over f_0 grows under LOMO, where category identity stops helping. Vision-only encoders instead favor f_0 when pooled, consistent with pitch being partly inferable from visible size, though that ordering reverses on unseen materials. The single-frame distilled student improves most on β ($0.26 \rightarrow 0.38$ under true LOMO), so it is the damping term, not pitch, that crosses the modality gap.

6 Discussion

Our pooled distillation result (0.44) barely exceeds the 0.43 of a category-median predictor, so we do not claim transfer on that basis (Appendix B); under true LOMO, where category-median strategies score zero by construction, the improvement ($0.234 \rightarrow 0.338$, $p = 0.002$) cannot be explained by category leakage.

Table 3 shows that the single-frame student’s gain is concentrated in β ($0.26 \rightarrow 0.38$ under true LOMO) rather than pitch ($0.14 \rightarrow 0.18$), which is what licenses reading the transferred structure as damping-related rather than pitch-related. Separately, for Q under true LOMO, half of the four-

frame model’s improvement (+0.051 of +0.104) comes from widening from one to four frames with no acoustic supervision, suggesting that video encoders trained on longer clips may narrow the gap even without distillation.

The teacher reaches $\rho = 0.60$ under LOMO; the student reaches 0.34. Characterizing what fraction of damping is visually recoverable in principle is an open problem that our benchmark makes it possible to study. Two caveats bound the present result. First, evaluation is confined to Greatest Hits (977 videos, 17 materials) and one visual domain, so whether the projection head learns a general appearance-to-damping mapping or associations tied to controlled drumstick footage is unresolved. Second, the $R^2 \geq 0.8$ filter retains only 63% of annotated impacts, at strongly material-dependent rates (water 24%, gravel 33%, glass 91%), so per-material correlations cover a selected subset of each category, and the discarded impacts cannot be assessed with Q at all, because the target is undefined there.

References

- [1] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017.
- [2] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations (ICLR)*, 2025.
- [3] Samuel Clarke, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L James, and Jiajun Wu. ReallImpact: A dataset of impact sound fields for real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1516–1525, 2023.
- [4] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [5] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2827–2836, 2016.
- [6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [7] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. MAViL: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36:20371–20393, 2023.
- [8] Jiaxun Liu and Boyuan Chen. SonicSense: Object perception from in-hand acoustic vibration. In *Conference on Robot Learning (CoRL)*, 2024.
- [9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- [10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khilodov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.
- [11] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multi-sensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.
- [12] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413, 2016.

- [13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.
- [14] Abitha Thankaraj and Lerrel Pinto. That sounds right: Auditory self-supervision for dynamic robot manipulation. In *Conference on Robot Learning*, pages 1036–1049. PMLR, 2023.
- [15] Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.
- [16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [17] Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. *arXiv preprint arXiv:2512.19687*, 2025.

A Within-Material Q Prediction

Table 4: Within-material Q prediction (ρ). Category identity carries no information here.

Model	Metal	Wood	Ceramic	Plastic
PE-AV (audio)	0.67	0.57	0.31	0.45
CAV-MAE (audio)	0.65	0.45	0.12	0.32
PE-AV (video)	0.36	0.07	−0.08	0.00
CLIP	0.06	0.03	−0.14	−0.19
DINOv2	0.30	0.09	0.09	−0.03

Table 4 shows that audio still works within a material (metal 0.67, wood 0.57), while visual encoders are near chance within wood, ceramic, and plastic and reach at most 0.36 within metal. CLIP’s correlations on ceramic (−0.14) and plastic (−0.19) are negative, suggesting that its features are, if anything, inversely related to the physical variation that remains once the material is fixed.

B Why Pooled Correlation Is a Weak Measure

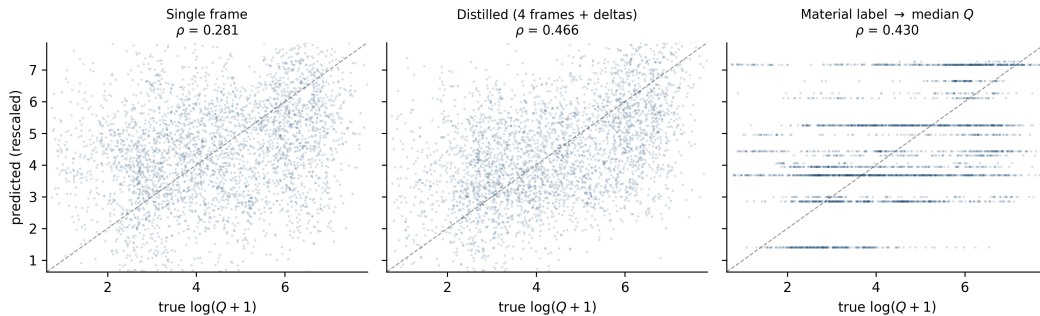


Figure 2: Predicted against true Q on the held-out split ($n = 3,510$), with predictions standardized to the target scale and the diagonal shown for reference. Left: a single frame at impact supports only a weak trend. Middle: after distillation, the cloud tightens (first of three seeds, $\rho = 0.466$, against the three-seed mean of 0.440). Right: mapping the *true* material label to its per-category median Q produces at most 17 distinct predictions, visible as horizontal bands, and still reaches $\rho = 0.430$.

Figure 2 shows why we do not rest the claim on pooled correlation. The category-median predictor is a step function over 17 levels and cannot separate two metal objects at all, yet its pooled ρ of 0.430 sits within 0.01 of the distilled model’s 0.440. Pooled correlation on this task rewards material recognition.

C Pooled Results and Controls

Table 5 gives the pooled figures behind Section 5. Widening the visual window helps only modestly on the pooled measure, with ρ rising from 0.281 to 0.298 (four frames + deltas) and overlapping intervals; the effect is clearer under LOMO, where the same change is worth +0.051. Distillation raises pooled prediction to 0.440, marginally above the 0.430 obtainable from perfect material-category knowledge and still far below the audio teacher.

Table 5: Q prediction from vision on the held-out video split, with and without distillation from the audio teacher.

Visual input / model	ρ	Interval	Type
<i>No audio supervision</i>			
Single frame at impact	0.281	[0.247, 0.313]	95% CI
Four frames, concatenated	0.299	[0.269, 0.332]	95% CI
Four frames + deltas	0.298	[0.267, 0.329]	95% CI
<i>Distilled from the audio teacher</i>			
Four frames + deltas	0.440	[0.426, 0.466]	three-seed range
Material-label reference	0.430	[0.400, 0.457]	95% CI
Teacher (PE-AV audio)	0.727	[0.707, 0.746]	95% CI

Table 7 collects the controls summarized in Section 5; two details are not given there. First, the capacity result has a natural explanation: with roughly 14,000 noisy scalar targets, a regression head has little to learn from, whereas the teacher supplies a dense 1024-dimensional target whose geometry already encodes damping. Second, the positional indexing used by an earlier version of our pipeline (13.3% material agreement) reaches 0.360, close to the fully shuffled teacher’s 0.359. Against the matched teacher’s 0.466, the part of the gain that requires correspondence is therefore +0.107.

Table 6: Audio embeddings capture the quality factor Q , while single-frame vision largely does not. Spearman ρ between ridge-probe predictions from frozen embeddings and ground-truth Q , on the held-out video split.

Model	Modality	ρ	95% CI
PE-AV	Audio	0.727	[0.707, 0.746]
CAV-MAE	Audio	0.701	[0.681, 0.721]
PE-AV	Joint	0.696	[0.676, 0.717]
CAV-MAE	Joint	0.608	[0.583, 0.633]
PE-AV	Video	0.362	[0.333, 0.393]
CLIP	Video	0.281	[0.247, 0.313]
DINOv2	Video	0.271	[0.242, 0.303]
CAV-MAE	Video	0.151	[0.121, 0.181]

Table 7: Controls, as pooled ρ on the held-out video split. Correspondence arms hold the seed fixed, so the matched-teacher row is the first (and highest) of the three seeds (0.466) rather than their mean (0.440).

Control	Condition	ρ
<i>How much is category recognition alone worth?</i>		
	CLIP material classifier (45.6% acc.) $\rightarrow$ median Q	0.279
	true material label $\rightarrow$ median Q	0.430
<i>Is the trainable head doing the work? (cached single-frame features[†])</i>		
	linear probe on $\log(Q + 1)$	0.300
	three-layer MLP on $\log(Q + 1)$, no teacher	0.275
<i>Does the teacher need to correspond to the frame? (four frames + deltas)</i>		
	no distillation	0.298
	shuffled teacher (10.6% material agreement)	0.359
	positional indexing (13.3% material agreement)	0.360
	matched teacher	0.466

[†]The capacity pair is measured on cached single-frame embeddings rather than the impact-window pipeline; the two rows are comparable with each other, and their values bracket the 0.281 single-frame figure.

Table 8: Physical structure generalizes to unseen materials. Mean over the 17 held-out materials, with intervals bootstrapped over materials.

Model	Modality	LOMO ρ	95% CI
PE-AV	Audio	0.600	[0.528, 0.658]
CAV-MAE	Audio	0.595	[0.538, 0.645]
PE-AV	Joint	0.572	[0.505, 0.630]
CAV-MAE	Joint	0.492	[0.441, 0.541]
DINOv2	Video	0.239	[0.177, 0.299]
CLIP	Video	0.227	[0.167, 0.289]
CAV-MAE	Video	0.206	[0.148, 0.264]
PE-AV	Video	0.197	[0.144, 0.245]