Behavioral Geometric Supervision Aligns Video Foundation Models with Human Social Perception

Kathy Garcia
Department of Cognitive Science
Johns Hopkins University
Baltimore, MD 21218
kgarci18@jhu.edu

Leyla Isik
Department of Cognitive Science
Johns Hopkins University
Baltimore, MD 21218
lisik@jhu.edu

Abstract

Current video foundation models, including the strongest self-supervised models such as V-JEPA2, fail to capture how humans organize social information in dynamic scenes. For example, across a range of diverse vision models tested, none were able to predict human similarity judgments to social video clips as well as a sentence embedding model of the caption text (MPNet). We show this gap in vision model performance can be closed by a compact behavioral supervisory signal. We introduce *behavioral geometric supervision (BGS)*: a hybrid objective that constrains local and global pairwise embedding geometry to match the relational similarity structure across videos. We apply this method using a new human similarity dataset, containing 49,484 odd-one-out judgments from 250 naturalistic social video clips, and low-rank adaptation across four ViT backbones (V-JEPA 2/2.1, TimeSformer, VideoMAE, and CLIP). We find that one of the best fine-tuned models, V-JEPA 2.1, more than triples in performance compared to the pre-trained baseline and reaches 75% of the noise ceiling, exceeding the strongest sentence-embedding baseline. In addition, finetuned models (i) capture unique variance in human judgments that caption-based language model embeddings do not, (ii) develop interpretable social-affective attributes (valence, arousal, and dominance) despite never being trained on any of these attributes, (iii) transfer, with frozen features to a separate dataset of out-of-distribution abstract social interactions, and (iv) shift spatial attention toward socially informative regions (faces, gaze, and interacting bodies). A matched language-distillation control fails to reproduce these gains, ruling out caption transfer as the mechanism. Our results show how a modest amount of human behavioral data can steer video models toward abstract, human-aligned social-visual representations that generalize to out-of-distribution social interactions.

1 Introduction

Humans effortlessly perceive the visual social world with remarkable nuance: we can easily tell whether two people are comforting each other, collaborating, or competing, all by watching brief interactions. As AI systems increasingly interpret and interact in human-centered environments, aligning their representations with human social perception is essential. Yet how this capacity arises in artificial systems, and whether the dynamic relational structure of human social perception can emerge in pure vision models remains unknown.

Recent benchmarks have shown that current vision foundation models systematically under-explain human social judgments compared to language models [Garcia et al., 2025], which have been shown to capture multimodal sensory judgments [Marjieh et al., 2023]. While prior alignment work has shown that human similarity judgments can reshape vision model representations [Muttenthaler

et al., 2023, Fu et al., 2023, Sundaram et al., 2024, Mahner et al., 2025], it has not applied these methods to dynamic social videos where similarity depends on temporally integrated information. Recent concurrent work [Policzer et al., 2025, Schad et al., 2025] has "brain-tuned" models on social-cognition using human fMRI data, highlighting the value of cognitive supervision but at infrastructure costs (based on expensive fMRI data) that limit scalability. Two questions therefore remain open: (1) can human social similarity structure be induced in video foundation models through scalable behavioral supervision, and (2) is the resulting social structure visual in nature or is it redundant with the semantic structure language embeddings already encode [Marjieh et al., 2023]?

Here, we introduce *behavioral geometric supervision (BGS)*: a differentiable training signal in which pairwise distances in a video model's embedding space are constrained to match both the local and global relational geometry of human odd-one-out similarity judgments, using a hybrid triplet-RSA objective. We collect a novel dataset of 49,484 similarity judgments from 250 naturalistic two-person social videos, and fine-tune four heterogeneous vision transformer backbones (V-JEPA 2, TimeSformer, VideoMAE, CLIP) with our new hybrid triplet–RSA objective applied through low-rank adaptation (LoRA) [Hu et al., 2022], updating fewer than 2% of parameters. We then test, by direct comparison against contrastive, predictive, and language-distillation alternatives on the same backbones, whether the resulting structure is reducible to other types of supervision.

Contributions. We have three contributions.

(1) A dynamic social video similarity benchmark reveals missing social-visual representations. We release the first large-scale human similarity benchmark for dynamic social video: 49,484 odd-one-out judgments from 245 participants for 250 naturalistic two-person clips. We identify a representation gap in vision models: across 14 pretrained vision encoders (e.g., V-JEPA 2/2.1), none match human social similarity as well as caption-based MPNet embeddings, indicating video pretraining misses key social structure.

(2) Behavioral Geometric Supervision for compact relational alignment. We introduce Behavioral Geometric Supervision (BGS), a LoRA-based hybrid triplet + differentiable RSA objective updating $<2\%$ of parameters. BGS closes the gap in vision models across four heterogeneous ViT backbones, explaining on average 72.6% of the split-half ceiling, and 80.8% from the strongest model. MPNet distillation and RSA-only ablations underperform BGS, and triplet-only ablations trend lower, showing that the gains depend on behavioral targets and benefit from joint local/global geometric supervision.

(3) Transferable, interpretable, capability-preserving social structure. BGS-aligned representations transfer, with frozen features, to out-of-distribution (OOD), abstract, dyadic social interactions, develop interpretable social-affective attributes (valence, arousal, dominance) without re-training, reorganize spatial attention to socially informative regions (faces, gaze, interacting bodies), and preserve action recognition on UCF101 and Kinetics-400.

2 Related Work

We situate our work among three related areas: alignment of models with human similarity judgments in vision, video pretraining beyond category-level recognition, and language-based accounts of perceptual similarity.

Behavioral alignment of Vision models with Human Similarity Judgments. Human similarity judgments have long been used to probe mental representations [Biederman, 1987, Edelman, 1998, Goldstone, 1994, Hebart et al., 2020, Nosofsky, 1986] with large-scale odd-one-out (OOO) behavioral studies mapping latent "similarity spaces" humans use for objects [Hebart et al., 2020], materials [Schmidt et al., 2025], and interaction environments [Josephs et al., 2023].

Recent alignment work uses these judgments to reshape model representations: VICE [Muttenthaler et al., 2022], [Muttenthaler et al., 2023], DreamSim [Fu et al., 2023], and AligNet [Muttenthaler et al., 2025] show that human derived similarity signals yield more perceptually aligned and interpretable features. However, most of this work focuses on static images, low-level perceptual comparisons, or synthetic domains, while [Dima et al., 2022] found that human judgments of dynamic stimuli rely more on social-affective features than surface visual/scene features. Here we collect the first large-scale dataset of social video similarity, and apply new methods in BGS to video models.

A parallel line of work [Kaufmann et al., 2023, Liu et al., 2025, Lu et al., 2025] aligns multimodal models to human preferences via scalar rewards, direct preference optimization, or online negative sampling, but optimizes output quality and leaves internal representational geometry largely unconstrained. These methods also typically require human-in-the-loop annotation at scale [Furuta et al., 2024, Li et al., 2024]. In contrast, we apply BGS directly shaping the pairwise geometry of video embeddings using relational similarity judgments.

Beyond categorical video pretraining. Modern video encoders, including transformer-based architectures trained at scale, [TimeSformer; Bertasius et al., 2021], [ViViT; Arnab et al., 2021], and [VideoMAE; Tong et al., 2022] have achieved strong results on action classification benchmarks. Yet their dominant objectives and evaluations emphasize category-level recognition (e.g., "dancing" vs. "cooking") instead of higher-level aspects of social behavior (e.g., intentions, affect, interactions, etc.). Multimodal video-language models such as VideoCLIP [Xu et al., 2021], All-in-One [Wang et al., 2022], and Qwen3 [Bai et al., 2025] introduce textual supervision, which can support more abstract semantic representations. These video-language approaches still depend on captions and descriptions, which can miss the relational and affective cues people use to compare social scenes.

The V-JEPA family [Assran et al., 2025, Mur-Labadia et al., 2026] provides a particularly strong test case for our question of vision model representations beyond categorization. V-JEPA 2 [Assran et al., 2025] is a large self-supervised predictive video model trained on roughly 1M hours of video with a joint-embedding predictive objective, reporting state-of-the-art results on action understanding, planning, and robotic manipulation. V-JEPA 2.1 [Mur-Labadia et al., 2026] extends this line with a dense prediction loss and deep self-supervision, yielding spatially and temporally consistent features that further improve robotic grasping, navigation, and depth estimation. Yet, neither model has been evaluated on whether it can decode higher-level social features from human social similarity structure. We test both V-JEPA 2 ViT-L ($\sim$300M params) and the distilled V-JEPA 2.1 ViT-B variant ($\sim$ 86M params) to compare across different parameter sizes, and ask whether predictive video pretraining alone is sufficient, or behavioral similarity supervision provides distinct alignment.

Language baselines and caption-derived structure. Marjieh et al. [2023] demonstrated that caption-based language embeddings approximate human similarity across sensory modalities, often outperforming pretrained vision encoders. As such, we test a range of language model embeddings from video captions as our reference, and treat exceeding the highest performing embedding space, MPNet [Song et al., 2020], as a meaningful bar. Exceeding this baseline suggests behavioral fine-tuning is capturing structure beyond what can be recovered from captions alone. It remains an open question whether and to what extent vision models can surpass this baseline. Here, we address this question using not only performance comparison, but also variance partitioning and a matched language-distillation control absent from earlier behavioral similarity approaches [Fu et al., 2023, Muttenthaler et al., 2022, 2025, Sundaram et al., 2024].

3 Methods

Our approach has two stages: (1) Measure the human-perceived similarity: we collect odd-one-out judgments on video triplets to construct a human similarity matrix, (2) Leverage behavior-guided fine-tuning on a video model with the objective of matching its embedding distances to human similarity structure. This is achieved through a hybrid loss function that enforces both local triplet constraints and global alignment of the full similarity matrices (Fig. 1).

3.1 Human Similarity Judgment Dataset

We introduce a dense large-scale dataset of human similarity judgments of short video social clips.

Stimuli. The stimulus set is 250 3-second clips of everyday two-person social interactions from a curated subset [McMahon et al., 2023, Garcia et al., 2025] of the Moments in Time dataset [Monfort et al., 2019], paired with one-sentence captions used to evaluate language models (Appendix §P.1).

Triplet odd-one-out task. Similarity judgments used a triplet odd-one-out paradigm [Hebart et al., 2020]: on each trial, participants saw three videos and were instructed to "focus on what the people are doing and choose the odd-one-out," implicitly nominating the remaining pair as more similar.

To ensure full pairwise coverage with minimal participant load, we generated triplets with a greedy set-cover algorithm (Appendix Alg. 1) that guarantees every video pair appears in at least one triplet. 245 participants (mean age 19.5; 64% female; 75% Asian or White; recruited via a U.S. university research platform) completed the study under IRB approval (see Appendix Fig. 10 for full demographics, and §5.3 for the discussion on sampling limitations).

Splits and similarity matrices. We adopt the pre-determined train-test split [McMahon et al., 2023]: 200 train videos (12,239 triplets) and 50 test videos (368 triplets); the remaining 36,877 triplets, which contained videos from both splits, were excluded. Human similarity is estimated as the frequency with which two videos were judged the more-similar pair, yielding a 200×200 training RSM and a 50×50 test RSM. Crucially, the similarity judgments on the 50 test videos are *stimulus-disjoint* from training (no test stimulus appears in any training triplet) providing a conservative generalization test; the test RSM covers 637 of 1,225 possible pairs. We use Spearman correlation squared (R^2) throughout for robustness to non-uniform sampling density alongside % of the split-half reliability noise ceiling ($R^2 = 0.213$; Spearman-Brown corrected, Appendix §F.5).

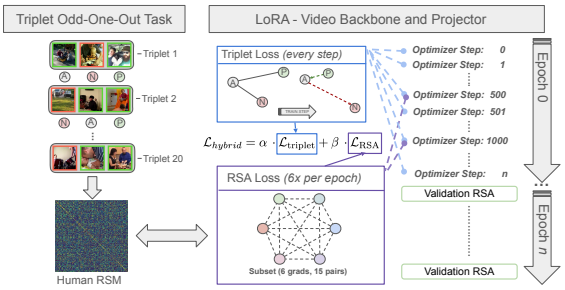

Figure 1: Overview of BGS. **Left:** human odd-one-out judgments on video triplets with anchor (A), positive (P), and odd-one-out (N) examples and the human RSM. **Right:** a LoRA-adapted video backbone is trained with a triplet loss at every step (blue) and an RSA loss applied periodically within each epoch (purple; Eq. 4).

3.2 Pretrained Vision and Language Models

We evaluate 12 pretrained video encoders (CNN- and Transformer-based), 2 image encoders (CLIP ViT-B/32, DINO-ViT-B/16), 2 vision-language models (Qwen3, LFM2.5), and 21 sentence-embedding language model baselines spanning diverse architectures and training objectives (full list: Appendix Tab. 3).[1] To select the evaluation layer, we perform a 5-fold cross-validation on the 200-video training set across all model layers, choose the layer with the highest mean Spearman's ρ across folds, and then fix that layer for evaluation on the held-out 50-video test set.

3.3 Behavior-Guided Fine-Tuning of Vision transformers

Our core approach is to fine-tune vision transformers using the human judgments as supervision. To test whether BGS generalizes across pretraining objectives, we apply lightweight LoRA fine-tuning to four backbones spanning four distinct pretraining regimes, updating less than 2% of parameters while freezing the rest. This approach inserts low-rank matrices into each attention layer (rank = 16). We show that our results generalize across a diverse range of vision transformers, and report fine-tuning results for modern joint-embedding predictive models (V-JEPA-2 ViT-L, Assran et al., 2025 and V-JEPA-2.1 ViT-B, Mur-Labadia et al., 2026), the highest performing image model (CLIP, Radford et al., 2021), the highest performing video transformer (TimeSformer, Bertasius et al., 2021), and a masked auto-encoder (VideoMAE, Tong et al., 2022).

3.3.1 Hybrid Loss Function

Both terms in our loss are instances of BGS: each takes relational judgments as input and produces gradients that adjust pairwise distances in embedding space. We design a loss $\mathcal{L}_{\text{hybrid}}$ that combines a triplet loss term ($\mathcal{L}_{\text{triplet}}$) and an RSA loss term ($\mathcal{L}_{\text{RSA}}$) to address local & global alignment (Fig. 1).

[1]We also include Qwen3-VL-2B-Instruct as a vision-language reference point, selected for parameter-count parity with V-JEPA 2 (Appendix §F.3). For each model layer, we extract embeddings for all videos, or sentence captions where applicable, compute a 50×50 cosine-similarity matrix, and compare it to the human test-set RSM using RSA [Kriegeskorte et al., 2008]. Embedding dimensions are standardized via Johnson–Lindenstrauss random projection with $\varepsilon = 0.1$ (Appendix F).

Shared notation and distance. Let $f(v)$ be the embedding of video v. We use ℓ_2-normalized embeddings $\mathbf{z}_i = f(v_i)/\|f(v_i)\|_2$ and define a single cosine-distance operator shared by both losses:

$$d(i,j) \;=\; 1 - \langle \mathbf{z}_i, \mathbf{z}_j \rangle. \tag{1}$$

Triplet Loss (local constraints) For each human odd-one-out judgment we seek to ensure the distance between anchor video i and its positive pair j is less than the distance to its negative pair k (odd-one-out) by a margin of γ. Specifically, we penalize violations of a margin $\gamma = 0.2$:

$$\mathcal{L}_{\text{triplet}}(i,j,k) \;=\; \max\big\{0,\; d(i,j) - d(i,k) + \gamma\big\}. \tag{2}$$

RSA Loss (global geometry) To shape the broader geometry toward human similarity structure, we inject a step based on representational similarity analysis (RSA; Kriegeskorte et al. [2008]) step six times per epoch. At each RSA step, we sample a batch of $K{=}24$ videos $\mathcal{K}$ and designate a subset of $M{=}6$ indices $\mathcal{G} \subset \mathcal{K}$ whose embeddings carry gradients. We limit gradients to $M{=}6$ to keep memory manageable: each RSA step considers all $\binom{24}{2} = 276$ pairwise distances but backpropagates only through the 123 pairs involving at least one gradient carrying video, which empirically provided sufficient supervision without the overhead of updating all 24 embeddings.

We calculate model RSM entries with $d(\cdot,\cdot)$ for all unordered pairs $\{i,j\} \subset \mathcal{K}$ with $i \neq j$ and $i \in \mathcal{G}$ or $j \in \mathcal{G}$. Corresponding human distances $d^{\text{H}}(i,j)$ are taken from the split-specific behavior RSM, masking out pairs without judgments to create a masked index set $\mathcal{M}$.

The RSA loss is the negative RSA score between the z-scored model and human distances of the masked index set $\mathcal{M}$:

$$\mathcal{L}_{\text{RSA}} \;=\; -\operatorname{corr}\Big(z\big(\operatorname{vec}(d)\big)[\mathcal{M}],\; z\big(\operatorname{vec}(d^{\text{H}})\big)[\mathcal{M}]\Big), \tag{3}$$

where $\operatorname{vec}(\cdot)$ denotes vectorization of the upper triangle, and $z(\cdot)$ denotes per-step standardization to zero mean and unit variance.

Pearson correlation is used for the RSA loss during training to ensure the loss is differentiable.

Hybrid Loss. We combine triplet (local) and RSA (global) supervisions with a weighted objective:

$$\mathcal{L}_{\text{hybrid}}^{(t)} \;=\; \alpha\, \mathcal{L}_{\text{triplet}}{}^{(t)} \;+\; \mathbb{1}_{\text{RSA}}(t)\, \beta\, \mathcal{L}_{\text{RSA}}{}^{(t)}, \tag{4}$$

where $\mathcal{L}_{\text{triplet}}$ captures fine-grained constraints from odd-one-out judgments and $\mathcal{L}_{\text{RSA}}$ encourages broader geometric alignment on sampled subsets. The indicator $\mathbb{1}_{\text{RSA}}(t)$ equals 1 if step t is one of the scheduled RSA steps and 0 otherwise. Specifically, we compute the total number of optimizer steps in an epoch, divide by 6, and activate the RSA loss at these evenly spaced intervals. We fix $\alpha = 0.7$ and linearly ramp β from 0.3 to 0.7 over training epochs. We do this to emphasize the local triplet loss early and ensure the model starts by getting the odd-one-out relationships correct, then gradually increase the weight of the global RSA loss (β) as training progresses to fine-tune the overall similarity structure (see Appendix §F.2).

Training Procedure. We fine-tune for 50 epochs with AdamW [see Loshchilov and Hutter, 2017] with learning rate = 1×10^{-4}, mixed precision, and gradient-checkpointing, using a batch size of 4. At each optimizer step, we apply the triplet loss; the RSA term is injected periodically as described above. We select the best checkpoint based on RSA validation performance on a held-out 20% split of the training judgments (monitoring explained variance R^2). For ablations, we also train models with triplet-only and RSA-only objectives under the same optimizer and schedule. Each V-JEPA 2.1 objective was trained with 5 random seeds; we report the mean. Downstream analyses use the seed-0 hybrid checkpoint.

Language-distillation control We also train a matched language-distillation control that replaces human triplet judgments with similarity of caption embeddings from MPNet (the highest-performing language model) as the supervisory target. Per-video captions are concatenated and encoded (L2-normalized outputs z_i), yielding $D_{ij}^{\text{MPNet}} = 1 - z_i^{\top} z_j$ over the 200 training videos. We optimize the same hybrid loss as the behavioral run (Eqs. 2-3) with same weights as hybrid run, substituting

D^{MPNet} for $d^{\mathrm{H}}(i, j)$ in the RSA term and synthesizing odd-one-out triplets directly from MPNet space, mirroring the human odd-one-out task with MPNet as the rater. All other hyperparameters match the behavioral hybrid run, isolating the supervision target as the only difference.

3.3.2 Linear-Probe Evaluations: Social-Affective Generalization & Action Recognition

To assess whether human similarity alignment changes model representations beyond the primary similarity task, we conducted two auxiliary linear-probe evaluations. First, we tested generalization to five social-affective attributes labeled in the original video dataset [McMahon et al., 2023]: *Intimacy* (how intimate or personal the interaction is), *Valence* (overall emotional positivity vs negativity), *Arousal* (energy or intensity of the action), *Dominance* (power dynamics between people), and *Communicating* (whether people in the video are communicating with one another). They were independently rated by and averaged across annotators, then z-scored across the 250 videos. We trained ridge-regression linear probes on layer-wise model embeddings using the same train-test split as in the main experiments. Second, to test whether human-aligned fine-tuning preserved the models' original action-recognition capacity, we evaluated frozen pretrained and fine-tuned video backbones on UCF101 [Soomro et al., 2012] and Kinetics-400 validation set [Kay et al., 2017]. We extracted TimeSformer and V-JEPA2.1 embeddings and trained linear probes on UCF101 split1 across three random seeds (Appendix I) and on kinetics, reused TimeSformer's original K400 classification head (Appendix §J).

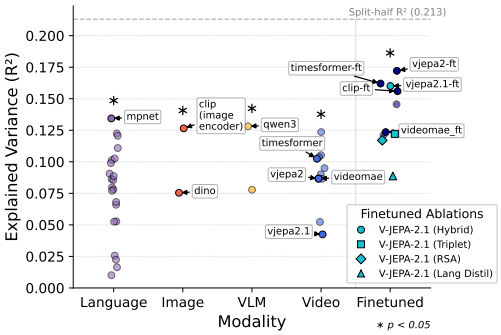

Figure 2: Model–human alignment (R^2, Spearman RSA) by modality. Shapes: V-JEPA 2.1 fine-tuning ablations (mean of 5 seeds; CIs in Tab. 5). Dashed line: split-half noise ceiling ($R^2 = 0.213$; §F.5). $*$: $p < 0.05$ within modality (1,000 permutations).

4 Results

4.1 Pretrained vision models are worse at matching human video similarity than language model embeddings of video captions

Across twelve pretrained video encoders, two image encoders, two vision-language models, and 21 caption-based language encoders, pretrained vision models all provided a worse match to human social similarity than the best sentence encoder (*paraphrase-multilingual-mpnet-base-v2*, $R^2 = 0.134$; OOO accuracy 70.38%; Figure 2). Due to the high performance, we consider these sentence embeddings as a baseline going forward. Among pretrained baselines, the best video model (X3D-M) reaches only $R^2 = 0.124$ (OOO accuracy 68.48%), V-JEPA-2.1 ViT-B (distilled) and V-JEPA-2 ViT-L (large) achieve only $R^2 = 0.042$ (OOO 63.85%) and $R^2 = 0.086$ (OOO 60.05%), respectively (Appendix Tab. 3). This pattern accords with prior work [Marjieh et al., 2023] showing that language embeddings approximate human similarity across perceptual domains. Yet participants performed a purely visual task without captions, exposing critical gaps in current pretrained video representations.

4.2 Behavioral fine-tuning induces human-like social similarity structure in video models

We next ask whether we can imbue video models with more human-like similarity structure via fine-tuning. After fine-tuning, the strongest fine-tuned model, V-JEPA 2 ViT-L reached $R^2 = 0.172$ (81% of ceiling), exceeding MPNet ($R^2 = 0.134$; 63% of ceiling). The distilled V-JEPA 2.1 ViT-B variant, also exceeds the MPNet baseline, reaching $R^2 = 0.160$, 75.1% of ceiling (Fig. 2). Fine-tuning with the hybrid loss function outperforms both the local triplet ($R^2 = 0.122$; 57% of ceiling) and global RSA loss ($R^2 = 0.117$; 55% of ceiling) alone, and all BGS fine-tuned models outperform the language distillation control ($R^2 = 0.089$; 42% of ceiling). The same hybrid procedure produces consistent gains across all four backbones (TimeSformer, VideoMAE, CLIP, V-JEPA 2/2.1), spanning supervised classification, masked reconstruction, contrastive vision-language, and self-supervised predictive pretraining, while preserving action recognition performance with no

catastrophic forgetting (Appendix I). BGS hence recovers human-aligned social structure regardless of pretraining objective.

Statistical robustness. We performed a paired bootstrap over the 637 observed pairs (of 1,225 possible; 10,000 resamples; 5 training seeds averaged per resample). BGS raises V-JEPA 2.1 alignment from $R^2 = 0.042$ [95% CI : 0.016, 0.080]; 20% of ceiling, to 0.160 [0.109, 0.216]; 75% of ceiling (paired $\Delta R^2 = +0.118$ [+0.065, +0.173], $p < 10^{-4}$). The hybrid objective significantly outperforms the language-distillation control ($\Delta R^2 = +0.073$ [0.026, 0.120], $p = 0.002$) and RSA-only (+0.044 [0.007, 0.083], $p = 0.020$); advantage over triplet-only is similar in magnitude, but not significant (+0.039 [−0.005, 0.083], $p = 0.077$). RSA and triplet ablations are similar ($\Delta R^2 = 0.005$), so neither term alone accounts for the hybrid's advantage (Appendix Tab. 5).

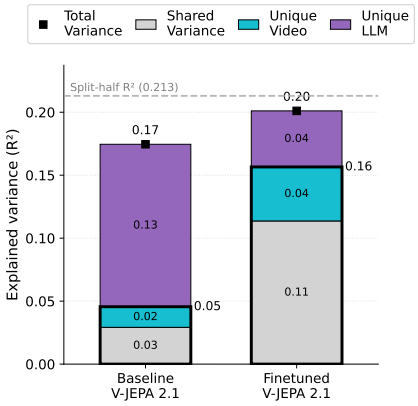

Figure 3: Variance partitioning between V-JEPA 2.1 and MPNet before and after BGS. BGS shifts explained variance from language-specific to shared and video-specific structure, bringing total variance near the noise ceiling. Black outlines: total variance explained by the video model.

Hyperparameter robustness and sample efficiency. We ran a 32 configuration sweep on V-JEPA 2.1 (Appendix §N). Held-out R^2 ranged from 0.101 to 0.176; every run beats the pretrained baseline ($R^2 = 0.042$) and 29/32 surpass MPNet ($R^2 = 0.134$). Alignment is insensitive to RSA batch composition, LoRA rank, and the β schedule shape; it benefits modestly from larger RSA weight and triplet margin, and depends most on RSA cadence, with fewer RSA updates per epoch degrading alignment. We report the pre-specified configuration rather than the best sweep result. A sample efficiency curve of the training triplets (3 seeds per size) shows monotonic gains: BGS exceeds the pretrained baseline by 1,000 triplets and recovers roughly half of the full-data gain with 2,500–5,000 (Appendix Tab. 8).

4.3 Finetuning captures social structure seen in language models and beyond

Variance partitioning using the best language embedding as a reference further supports the strength of our fine-tuning approach (Fig. 3). We fit multiple regressions predicting human similarity distances from model distances and decomposed the R^2 contributions of the pretrained V-JEPA 2.1, the fine-tuned model, and the language embedding. In the pretrained baseline, the video model contributed little unique variance ($R^2 = 0.02$) beyond what overlapped with the language model ($R^2 = 0.03$). After hybrid fine-tuning, the partitioning shifts in two distinct ways. First, structure that was uniquely accessible to language pre-tuning ($R^2 = 0.13$) collapses into the shared component ($R^2 = 0.11$), indicating the fine-tuned video model now expresses most of what caption embeddings uniquely encoded. Second, video-unique variance doubles from $R^2 = 0.02$ to 0.04; small in absolute terms, but persistent: the fine-tuned model captures social-similarity structure that even an idealized caption embedding cannot reach. Total explained variance reaches 94% of the noise ceiling. We also observe this trend with other fine-tuned video models (Appendix Fig. 8).

4.4 OOD generalization to abstract social interactions with no catastrophic forgetting

A central question is whether BGS learns a representation that is merely better at the training benchmark or recovers broader social structure. We test this in three complementary ways: social-affective linear probes, cross-domain transfer to abstract social interactions, and retention of standard action-recognition performance.

Interpretable social-affective attributes recovered. We tested whether BGS reorganizes representations along human-interpretable social-affective dimensions despite never receiving attribute labels. Following [McMahon et al., 2023], we trained ridge-regression probes on five densely annotated attributes: communication, valence, arousal, intimacy, and dominance. BGS substantially improved

over pretrained baselines on arousal, dominance, and valence, with modest gains on intimacy and a small decline on communication (Appendix H). Because they never appear in the supervision signal, such gains indicate that human similarity judgments encode an interpretable social-affective geometry that BGS makes linearly accessible. A matched language-distillation control failed to reproduce these gains on 4/5 attributes; the sole exception is communication, where MPNet's advantage is unsurprising given that words like "talking" and "discussing" are often directly in caption text.

Frozen-feature transfer to abstract social interactions. We next tested whether BGS-induced structure transfers beyond the naturalistic training distribution. We evaluated the frozen pretrained and BGS-aligned encoders for four backbones on PHASE [Netanyahu et al., 2021], a dataset of two-agent interactions, rendered as geometric shapes (no faces, bodies, and minimal scene context) moving in a 2D physical world (Appendix Fig. 6), which presents a strong case of OOD generalization relative to the naturalistic videos in model pre-training or fine-tuning. Following Malik and Isik [2023], we trained a logistic-regression probe on 400 videos and evaluated on the 100-video held-out generalization set. No backbone degraded after BGS. V-JEPA 2.1 showed the largest gain, reaching **81% accuracy** (Wilson 95% CI: 0.72–0.87), a 26-point gain over the pretrained baseline (55%), numerically exceeding SocialGNN (76%), which uses privileged ground-truth scene-graph inputs, and matching average human agreement (∼83%) within sampling error (Fig. 4). Gains were smaller for TimeSformer (60%→68%) and CLIP (65%→68%), and VideoMAE, the strongest pretrained encoder on this task, was unchanged (69%). The V-JEPA 2.1 gain is not only recovery from a weak baseline: although its pretrained accuracy was the lowest of the four backbones, its fine-tuned accuracy is the highest by 12 points. To our knowledge, this is the first demonstration that pixel-level video representations can reach prior baselines on this benchmark.

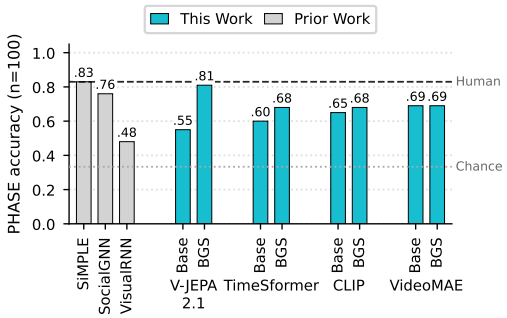

Figure 4: PHASE frozen-feature transfer: accuracy on the 100-video held-out generalization set for pretrained (Base) and BGS-aligned encoders vs. prior baselines (SiMPLE, SocialGNN, VisualRNN; Appendix §K). Dashed line: average human agreement; dotted line: chance (0.33).

Action recognition is preserved. Finally, we tested whether BGS sacrifices generic action-recognition capacity on UCF101 split1 [Soomro et al., 2012] and the full Kinetics-400 validation set [Kay et al., 2017]. With frozen linear probes (three seeds), V-JEPA 2.1 improved after BGS on both benchmarks (UCF101: $78.17 \pm 0.03\% \rightarrow 80.31 \pm 0.06\%$, $\Delta = +2.14$ pp; K400: $50.05 \pm 0.12\% \rightarrow 53.30 \pm 0.04\%$, $\Delta = +3.25$ pp Top-1), while TimeSformer was unchanged on UCF101 ($95.75 \pm 0.18\% \rightarrow 95.70 \pm 0.14\%$; $\Delta = -0.05$ pp). For TimeSformer we additionally ran a stricter, training-free test: reusing its original K400 classification head on the BGS-adapted backbone retains 98.8% of Top-1 accuracy ($77.28\% \rightarrow 76.32\%$; $\Delta = -0.96$ pp; Appendix §J). BGS therefore preserves, and for V-JEPA 2.1 marginally improves, generic action-recognition performance. Together, these results show that BGS does not merely improve fit to the trained similarity task. It produces representations that generalize to interpretable social-affective judgments and to abstract OOD social interactions, while preserving the discriminative capacities acquired during pretraining.

4.5 Spatial attention reorganization toward socially informative regions

To examine how fine-tuning changes video model representations, we performed attention rollout on V-JEPA 2.1 ViT-B (Appendix Fig. 13); the pretrained model concentrates spatial attention on a single focal region per frame (e.g., one agent's torso), while fine-tuning redistributes attention across both agents, gaze directions, interacting hands, and shared activity regions. Across all 250 videos, we observe a mean attention divergence of $1 - \bar{r} = 0.18 \pm 0.03$ (mean $\pm$ sd), indicating systematic spatial reorganization rather than isolated shifts.

Causal occlusion. To test whether these regions are causally used rather than merely attended, we re-scored the test set with individual social regions masked (Appendix §O, Tab. 9). Occluding faces (3.0% of frame area) reduces BGS-model alignment by 30% (R^2: $0.165 \rightarrow 0.115$), whereas a control

mask matched in size, shape, and motion but placed in a person-free location reduces it by only 1.5% (0.162). The face-specific drop is $4.2\times$ larger for BGS than the pretrained model ($\Delta R^2 = -0.050$ vs. -0.014). Background occlusion degrades both models equally (70% vs. 69% of alignment lost), consistent with masking two-thirds of the frame, whereas person occlusion disproportionately degrades the fine-tuned model (72% vs. 56%), despite covering half the area of the background mask. Per proportion of masked area, BGS sensitivity is ordered face $\gg$ person $>$ background, indicating that BGS specifically increases reliance on socially informative regions. Qualitative triplet examples further show that fine-tuning agrees on odd-one-out judgments with the human choice while the pretrained model disagreed, prioritizing human-relevant cues (Appendix Figs. 11–12). Together with the variance-partitioning result (§4.3) and the cross-domain transfer (§4.4), this spatial reorganization provides a third, mechanistically grounded line of evidence that BGS recovers genuinely social-visual structure rather than statistical regularities of the training corpus.

5 Discussion

Here we show that human social perception, measured via similarity judgments on dynamic social clips, can be induced in video representations through lightweight behavioral fine-tuning, BGS. Our fine-tuned model develops interpretable social features (e.g., valence, arousal, intimacy, dominance) without any direct supervision, captures variance in human judgments that caption-based language embeddings cannot, and reorganizes its spatial attention toward socially informative regions. Our novel objective generalizes across four backbones, and preserves standard action recognition performance.

Our results reposition social perception in video models as a latent property rather than a learned task. Pretrained backbones, optimized for action classification, masked reconstruction, or latent prediction, on their own fail to match human social similarity judgments, but carry enough information about the social world that a compact behavioral supervision signal suffices to make that information usable. How readily that information is unlocked may depend on the pretraining objective: the largest PHASE gain occurred for V-JEPA 2.1, whose BGS accuracy exceeded all other's by 12 points despite the weakest base score. These results suggest that predictive latent-space pretraining yields representations that behavioral supervision can most readily reorganize toward socially-relevant structure. Future work comparing predictive and non-predictive backbones at matched scale should test this prediction.

5.1 Behavioral relational similarity as latent-structure supervision

BGS occupies a niche in the space of human aligned training signal: unlike preference learning, it shapes representations rather than outputs and does not require a human in-the-loop; unlike attribute labeling, it captures the integrated structure of human judgment without enumerating individual axes; and unlike contrastive learning, which gets positive and negative pairs from augmentations/category labels, it draws its structure directly from human behavior. This approach is complementary to recent brain-tuning approaches for social video [Policzer et al., 2025, Schad et al., 2025], while triplet judgments provide a scalable behavioral target for the relational structure people can explicitly compare.

This framework extends naturally beyond social perception: any domain with human similarity data (e.g., objects, faces, scenes, materials, sounds) admits the same supervision, and large-scale triplet datasets already exist in several of these domains [Muttenthaler et al., 2023, Josephs et al., 2023, Fu et al., 2023, Schmidt et al., 2025].

5.2 What our behavioral fine-tuning accesses that language models do not

Fine-tuned video models explain unique human similarity judgments variance beyond the strongest caption based language encoder, and simultaneously increase the variance shared with language (Fig. 3). More unique variance in video embeddings and increased shared variance combined show the signature of a model that has acquired visual social structure that humans rely on [McMahon and Isik, 2023] rather than distilled language: pure distillation would increase shared variance while leaving unique video variance flat/shrinking it. One question naturally arises: why does the language baseline outperform all video models in the first place? Understanding social interactions often requires

abstract inferences (goals, roles, affect) that go beyond form and motion. Video models, trained mainly for action classification, may emphasize kinematics and object cues, while caption-based language embeddings encode high-level semantics (e.g., "friends boxing for fun" vs. "strangers fighting angrily"). This makes the increased unique-variance for fine-tuned video models particularly compelling, suggesting humans encode aspects of social structure visually beyond verbal descriptions.

Three convergent lines of evidence support this: (1) On triplets where the pretrained video model disagrees with humans but the language model agrees, the fine-tuned video model flips its judgments toward the human choice (e.g., prioritizing social relational over surface visual cues; Appendix §Q). (2) Spatial attention reorganizes toward faces, gaze, and interacting agents after BGS, without any spatial supervision during training (Appendix Fig. 13), and causal occlusion confirms these regions are used: masking faces or people disproportionately degrades the fine-tuned model, whereas background masking affects both models equally (§4.5, Appendix §O). (3) Linear probes recover interpretable axes of valence, arousal, and dominance from the fine-tuned representations despite none of them appearing in the supervision signal (Appendix Fig. 5). Together, they show the BGS model's similarity structure reflects high-level social distinctions, not just richer visual statistics.

5.3 Limitations

Our dataset, though diverse in action categories, draws from a single source corpus (Moments in Time, Monfort et al. [2019]) and reflects predominantly Western social norms: all 245 participants were recruited through a U.S. university platform, and captions were provided by native English speakers from Prolific. Because social similarity judgments vary with cultural background [Pang et al., 2024], the learned structure should not be assumed to generalize across all populations without cross-cultural validation. Our evaluation focuses on similarity alignment; we test generalization along multiple axes including cross-domain transfer (§4.4) and cross-task transfer (Appendix §I, §J). After BGS finetuning, V-JEPA 2 and 2.1 reach 75–81% of the explainable noise ceiling. One thing to note is the level of OOO human agreement (R^2 = 0.213), which is relatively low compared to other, more concrete rating or labeling tasks. The OOO judgment provides a relatively unconstrained measure of video similarity. Subjects may make judgments on a range of factors that vary by individual and cultural context. The aggregate-consensus nature of our similarity data also smooths over individual differences; future work on personalized, inter-subject variability or subgroup-specific alignment would test the limits of the current approach. Finally, the 250-video stimulus set, while behaviorally densely sampled, is small relative to image-domain benchmarks; cross-dataset replication is the natural next step. The existing video captions are relatively short and centered on the human actions and interactions. An interesting open question is whether LLM embeddings of richer captions could better account for the remaining unexplained variance in human similarity judgments.

5.4 Broader Impact

Video models that are aligned with human social similarity judgments provide a path to more trustworthy and intuitive AI systems. Embeddings that align with human behavior may improve interpretability, video retrieval, and recommendation by way of organizing content that is reflective of human categorization. Our findings suggest that such alignment also promotes representations of social-affective features, with potential applications in affective computing and safety-sensitive domains. However, models that reflect human perception may also inherit human biases. While our dataset is diverse, culturally specific notions of similarity may also be encoded as a result of the aforementioned factors. This motivates future broader studies that include analysis for bias with more diverse annotation sources, ensuring fairness and robustness across populations.

6 Conclusion

Together, these results reposition social perception as a latent property of video representations that becomes accessible only with the right supervision signal, and suggest that relational behavioral data offer a form of alignment different from the preference learning and cross-modal distillation that currently dominate multimodal training. We release our benchmark, trained adapters, and training code as resources for the community.

Acknowledgments and Disclosure of Funding

This work was funded in part by NSF GRFP DGE-2139757 awarded to K.G. and NIMH R01MH132826 awarded to L.I. The authors declare no competing interests.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, Montreal, Canada, 2021. IEEE. doi: 10.1109/ICCV48922.2021.00676.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding?, June 2021. URL http://arxiv.org/abs/2102.05095. arXiv:2102.05095 [cs].

Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, April 1987. ISSN 0033-295X. doi: 10.1037/0033-295x.94.2.115. URL http://dx.doi.org/10.1037/0033-295x.94.2.115.

Colin Conwell, Jacob S. Prince, Kendrick N. Kay, George A. Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15(1), October 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53147-y. URL http://dx.doi.org/10.1038/s41467-024-53147-y.

D. C. Dima, T. M Tomita, C. J. Honey, and L. Isik. Social-affective features drive human representations of observed actions. *eLife*, 11, 2022. doi: 10.7554/eLife.75027.

Shimon Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4):449–467, August 1998. ISSN 1469-1825. doi: 10.1017/s0140525x98001253. URL http://dx.doi.org/10.1017/s0140525x98001253.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020. URL https://arxiv.org/abs/2004.04730.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. *arXiv: 1812.03982*, 2018.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. *arXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2306.09344.

Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback, 2024. URL https://arxiv.org/abs/2412.02617.

Kathy Garcia, Emalie McMahon, Colin Conwell, Michael F. Bonner, and Leyla. Isik. Modeling dynamic social vision reveals gaps between deep learning and the humans. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b1bdb0f22c9748203c62f29aa297ac57-Paper-Conference.pdf.

Robert L. Goldstone. The role of similarity in categorization: providing a groundwork. *Cognition*, 52(2):125–157, August 1994. ISSN 0010-0277. doi: 10.1016/0010-0277(94)90065-5. URL http://dx.doi.org/10.1016/0010-0277(94)90065-5.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11):1173–1185, October 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-00951-3. URL http://dx.doi.org/10.1038/s41562-020-00951-3.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2106.09685. URL https://arxiv.org/abs/2106.09685.

Emilie L. Josephs, Martin N. Hebart, and Talia Konkle. Dimensions underlying human understanding of the reachable world. *Cognition*, 234:105368, May 2023. ISSN 0010-0277. doi: 10.1016/j.cognition.2023.105368. URL http://dx.doi.org/10.1016/j.cognition.2023.105368.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2312.14925.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design, 2024. URL https://arxiv.org/abs/2410.05677.

Shujun Liu, Siyuan Wang, Zejun Li, Jianxiang Wang, Cheng Zeng, and Zhongyu Wei. Ovip: Online vision-language preference learning for vlm hallucination, 2025. URL https://arxiv.org/abs/2505.15963.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

Jinda Lu, Jinghan Li, Yuan Gao, Junkang Wu, Jiancan Wu, Xiang Wang, and Xiangnan He. Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization, 2025. URL https://arxiv.org/abs/2504.15619.

Florian P. Mahner, Lukas Muttenthaler, Umut Güçlü, and Martin N. Hebart. Dimensions underlying the representational alignment of deep neural networks with humans. *Nature Machine Intelligence*, 7(6):848–859, 2025. doi: 10.1038/s42256-025-01041-7.

Manasi Malik and Leyla Isik. Relational visual representations underlie human social interaction recognition. *Nature Communications*, 14(1):7317, November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43156-8. URL https://www.nature.com/articles/s41467-023-43156-8. Number: 1 Publisher: Nature Publishing Group.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. Large language models predict human sensory judgments across six modalities, 2023. URL https://arxiv.org/abs/2302.01308.

Emalie McMahon and Leyla Isik. Seeing social interactions. *Trends in Cognitive Sciences*, 27 (12):1165–1179, December 2023. ISSN 13646613. doi: 10.1016/j.tics.2023.09.001. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661323002486.

Emalie McMahon, Michael F. Bonner, and Leyla Isik. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047.e8, December 2023. ISSN 0960-9822. doi: 10.1016/j.cub.2023.10.015. URL http://dx.doi.org/10.1016/j.cub.2023.10.015.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–8, 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2019.2901464.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Lukas Muttenthaler, Charles Y. Zheng, Patrick McClure, Robert A. Vandermeulen, Martin N. Hebart, and Francisco Pereira. Vice: Variational interpretable concept embeddings, 2022. URL https://arxiv.org/abs/2205.00756.

Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A. Vandermeulen, Katherine Hermann, Andrew K. Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments, 2023. URL https://arxiv.org/abs/2306.04507.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C. Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K. Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, November 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09631-6. URL http://dx.doi.org/10.1038/s41586-025-09631-6.

Aviv Netanyahu, Tianmin Shu, Boris Katz, Andrei Barbu, and Joshua B. Tenenbaum. PHASE: PHysically-grounded Abstract Social Events for Machine Social Perception, March 2021. URL http://arxiv.org/abs/2103.01933. arXiv:2103.01933 [cs, stat].

Robert M. Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57, 1986. ISSN 0096-3445. doi: 10.1037/0096-3445.115.1.39. URL http://dx.doi.org/10.1037/0096-3445.115.1.39.

Hio Tong Pang, Xiaolin Zhou, and Mingyuan Chu. Cross-cultural differences in using nonverbal behaviors to identify indirect replies. *Journal of Nonverbal Behavior*, 48(2):323–344, February 2024. ISSN 1573-3653. doi: 10.1007/s10919-024-00454-z. URL http://dx.doi.org/10.1007/s10919-024-00454-z.

Nico Policzer, Cameron Braunstein, and Mariya Toneva. The one where they brain-tune for social cognition: Multi-modal brain-tuning on friends. *arXiv preprint arXiv:2511.07988*, 2025. doi: 10.48550/arXiv.2511.07988.

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, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Daniel Carlström Schad, Shrey Dixit, Janis Keck, Viktor Studenyak, Aleksandr Shpilevoi, and Andrej Bicanski. Vibe: Video-input brain encoder for fmri response modeling. *arXiv preprint arXiv:2507.17958*, 2025. doi: 10.48550/arXiv.2507.17958.

Filipp Schmidt, Martin N. Hebart, Alexandra C. Schmid, and Roland W. Fleming. Core dimensions of human material perception. *Proceedings of the National Academy of Sciences*, 122(10), March 2025. ISSN 1091-6490. doi: 10.1073/pnas.2417202122. URL http://dx.doi.org/10.1073/pnas.2417202122.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 16857–16867, 2020. URL https://arxiv.org/abs/2004.09297.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Y. Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations? *arXiv*, 2024. doi: 10.48550/ARXIV.2410.10817. URL https://arxiv.org/abs/2410.10817.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 3487–3501, 2022. doi: 10.5555/3600270.3601002.

Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. doi: 10.48550/arXiv.2203.07303.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6787–6800, 2021. doi: 10.18653/v1/2021.emnlp-main.544. URL https://aclanthology.org/2021.emnlp-main.544/.

A Reproducibility Statement

We release the data we have collected on the odd-one-out similarity judgments (with the canonical 200/50 train-test split), along with human RSMs and video annotations on all 250 videos. The pertinent metrics for all of the models we have evaluated (both pretrained/baseline and fine-tuned) are provided in detail in §3, with additional details on RSA and variance partitioning in the Appendix (§F.4). Details on the availability for coding material concerning embedding extraction, similarity computation, and model evaluation (both for R^2 scores and odd-one-out accuracy) are provided in the Appendix (§G). This section also outlines the way by which others could obtain the Moments in Time dataset we used for our analyses. After gaining access, the mapping we used between raw video files and their representative IDs (0-249) is also available in the code base.

Furthermore, we release the configuration details for our vision models (V-JEPA 2/2.1, TimeSformer, CLIP, and VideoMAE) with LoRA adapters, including the scripts for hybrid, triplet-only, RSA-only, and triplet-budget-matched versions. See §3 for a more high-level description of training details. Next, the validation and reporting metrics follow the procedure outlined in §3.3.2. To facilitate this, we also release the scripts necessary for full pre-processing and training. You can find more information about the UCF101 linear-probe action recognition and social-affective probing experiments in Appendix §I). Finally, we wish to support both full retraining and more lightweight reproduction for accessibility. Therefore, we make available all training and evaluation code, as well as pretrained adapters and precomputed RSMs to reproduce these analyses. All parts listed above are available at `https://github.com/garciakathy/bgs-finetuning` (Full details in Appendix §G)

B Ethics Statement

All procedures conducted throughout this paper adhere to ethical standards. All participants gave informed consent, and data was collected under the JHU Homewood Institutional Review Board (IRB) approval. The experiment itself was straightforward: they were instructed to make quick "odd-one-out" choices on 3-second video clips that showed everyday social interactions without any identifying details. We did not collect any personal data beyond demographic information, all of which were optional. Participants were recruited from an online university crowd-sourcing platform and were awarded course credits for participation. The data we obtained was not used to infer private characteristics on the participants, only for model-human representational alignment. We release the dataset (except the actual videos due to licensing, see Appendix §G) and the code to encourage transparency and replication.

C LLM Usage

LLMs were used for minor edits, such as grammar, phrasing, and shortening, throughout this paper. LLMs were not used during the data collection or the analysis stages. No major task, such as ideation, full content generation, or substantive interpretation of results, was delegated to LLMs. All conceptual framing, experimental design, and analytical decisions were carried out by the authors.

D Compute & Resource Use

We report a rough estimate of a little over 150 GPU-hours of compute across all 48 models (21 language encoders, 2 image encoders, 2 VLMs, 12 pretrained video backbones, and 11 fine-tuned video models). Across this compute, the predominant cost is associated with our finetuning process, with a single V-JEPA 2.1 ViT-B + LoRA hybrid training taking $\sim$2.5 GPU-hours alone (12 epochs on a single A100, `num-epochs=50`, `early-stopping=5`). When we include all V-JEPA 2.1 and TimeSformer ablations, and the hybrid finetuning process for the baselines of CLIP and VideoMAE, finetuning as a whole consumed about 110 GPU-hours. Additional seeds and data sampling runs cost around 17.5 GPU-h. The remaining compute can be attributed to the checkpoint evaluation across different tasks, including behavioral RSA ($\sim$17 GPU-h), OOO triplet accuracy ($\sim$2 GPU-h), the PHASE OOD probe ($\sim$3 GPU-h), and attention-rollout visualizations ($\sim$0.5 GPU-h). All experiments ran on HPC cluster on a node of 4x NVIDIA A100 (80 GB) GPUs.

E Triplet Selection Algorithm

$\mathbf{S}^{(human)}$ is an estimate (aggregate), rather than a fully observed matrix. This is because the triplet sample is sparse relative to all pairs $\binom{250}{2} = 31,125$. We came up with a specialized algorithm to create the triplets to ensure adequate coverage with the least amount of participants possible. So, we designed this procedure so that *every possible pair* $(i,\ j)$ *appears in at least one triplet* $(i,\ j,\ k)$. Since the similarity (probability) matrix is constructed through how many times a pair of videos was rated similar based on how many times they appeared, this guarantees that each "pair" would have at least one rating.

This problem is conceptually equivalent to a *set cover*. Namely, the universe of elements consists of all video pairs, and each triplet corresponds to a subset that covers three of those video pairs. Finding the truly minimal set of triplets that covers all possible pairs is NP-hard. So, we implemented a greedy approximation strategy to iteratively choose the most informative triplet at each step, given all the triplets selected before and remaining.

- First, we randomly sample a candidate pool for the triplets at each iteration.
- Then, from this pool, we select the triplet that covers the largest number of pairs (max: 3) *not yet included*.
- Next, we mark those pairs as "covered" and continue the iteration until every pair has been assigned at least one triplet.

We prioritized efficiency with this "greedy" search, since we effectively minimized the number of triplets (and thus the number of participants) we needed to guarantee full pairwise coverage. After coverage was achieved, we adjusted the total number of triplets by adding more so that it was divisible by 220 (11 sets of 20 trials for each participant).

$$\text{Minimum possible triplets} = \frac{\binom{250}{2}}{3} = \frac{31{,}125}{3} = 10{,}375.$$

$$\text{Our greedy algorithm produced } 10{,}780 \text{ triplets} \rightarrow \frac{10{,}780}{220} = 49 \text{ participants required.}$$

Algorithm 1 Triplet Selection Covering All Pairs (Greedy Set Cover Approximation)

Require: Number of items N (e.g., $N = 250$ for 250 video stimuli)
Ensure: Set of triplets T covering all pairs, with $|T|$ divisible by 220

$P \leftarrow \{(i, j) \mid 0 \le i < j < N\}$ ▷ All pairs
$S \leftarrow \{(i, j, k) \mid 0 \le i < j < k < N\}$ ▷ All triplets
$T \leftarrow \emptyset$ ▷ Selected triplets
while $P \neq \emptyset$ **do**
    $C \leftarrow$ random sample of $\min(|S|, 10,000)$ triplets from S
    $best_triplet \leftarrow$ triplet in C maximizing coverage w.r.t. P
    $T \leftarrow T \cup \{best_triplet\}$
    Remove all pairs in $best_triplet$ from P
end while
$r \leftarrow |T| \bmod 220$
if $r \neq 0$ **then**
    Sample $220 - r$ triplets randomly from S and add to T
end if
return T

F Supplementary Evaluation and Analysis Procedures

F.1 RSA objective

During training we use Pearson-correlation RSA on z-scored pairwise distances. Pearson is smooth, so gradients propagate from the correlation through distances back to the embeddings. (For evaluation we report Spearman ρ^2 (reported as R^2), which is rank-based and non-differentiable.)

F.2 On the Hybrid Loss.

Our fine-tuning objective combines a triplet loss with an RSA loss, balancing local and global alignment. The triplet component ensures that fine-grained distinctions from the original model are preserved while pulling together pairs judged similar by humans. With the addition of the RSA component, it is complemented by aligning the model's overall pairwise structure with human RSMs. This distills the relational knowledge at a global level, reflecting the findings by Muttenthaler et al. [2023], who showed that constraints on global geometry that match human similarity yields better, and more interpretable task-effective features when preserving local structure. Where RSA is typically used as a tool for analysis [Kriegeskorte et al., 2008], our contribution takes it a step further and re-purposes it as a differentiable objective. With this, the hybrid loss leverages both local and global supervision to push the representation towards richer semantic space that is reflected by human judgments.

F.3 Pretrained Model Details

Video encoders. We evaluate nine video encoders spanning CNN and Transformer architectures: X3D-M, X3D-S, and X3D-XS from the X3D family optimized for efficient video classification [Feichtenhofer, 2020]; SlowFast-R50, a two-pathway CNN capturing both slow and fast temporal dynamics [Feichtenhofer et al., 2018]; Slow-R50 and C2D-R50 as single-pathway CNN baselines; I3D-R50; TimeSformer, a video Transformer factorizing spatial and temporal attention trained on Kinetics-400 [Kay et al., 2017, Bertasius et al., 2021]; and VideoMAE-Base, a masked autoencoder for self-supervised video pretraining [Tong et al., 2022]. We feed each 3s clip into these models after resizing frames to the required model resolution and extract embeddings at every layer using the DeepJuice software package [Conwell et al., 2024].

Image encoders. We benchmark two image models: CLIP ViT-B/32 (vision encoder only) [Radford et al., 2021] and DINO-ViT-B/16. For each video, we extract seven equally spaced frames, compute per-frame embeddings, and average across frames before continuing with the standard evaluation pipeline.

Vision-language models. We include Qwen3-VL-2B-Instruct as a modern multimodal reference point, selected for approximate parameter-count parity with V-JEPA 2. We input both the video caption and video frames (using the frame-extraction procedure above) and extract embeddings for RSA evaluation.

Dimensionality standardization. For fairness across architectures with varying hidden dimensionalities, we down-sample embeddings using sparse random projection (SRP) based on the Johnson–Lindenstrauss lemma with $\varepsilon = 0.1$. This automatically sets the projection size according to the number of samples, yielding 4,732 dimensions for the training split ($N = 200$) and 3,353 dimensions for the test split ($N = 50$), preserving pairwise distances within $\pm 10\%$ with high probability.

F.4 Variance partitioning analysis

We model human distances $d_{\text{human}}(i, j)$ with multiple regression using model distances as predictors. For models $X_1, X_2, \ldots$, we fit

$$\hat{d}(i,j) \;=\; \beta_0 \;+\; \sum_m \beta_m \, d_{X_m}(i,j)$$

over all video pairs in the test split, and report R^2. Unique and shared contributions are obtained by comparing nested models (e.g., unique X_1 is $R^2_{X_1,X_2} - R^2_{X_2}$). We use the best language model as one predictor, and the pretrained and fine-tuned V-JEPA 2.1 as the other predictors.

F.5 Split–half reliability

We estimate a noise ceiling for the human RSM with a split–half procedure that respects unequal judgments per pair. In each of 1,000 iterations we: (1) restrict to lower-triangle pairs with at least two ratings; (2) reconstruct binary votes ("similar"/"dissimilar") for each pair using its observed proportion and count, shuffle, and split the votes into two halves; (3) compute the proportion "similar"

in each half for every pair and take the Spearman correlation across pairs between halves; (4) average these correlations over iterations and apply the Spearman–Brown correction to estimate full-sample reliability. We report this corrected average as the split–half noise ceiling for the human judgments. In figures, we label this as *split–half* R^2, i.e., the squared Spearman–Brown–corrected split–half correlation.

G Code and data availability

All code, sentence captions, human odd-one-out similarity judgments (with the 200/50 train–test split), human RSMs, and trained LoRA adapters are publicly available at `https://github.com/garciakathy/bgs-finetuning`. The social-affective attribute ratings are available on OSF (`https://osf.io/4j29y/`) under a CC-BY-4.0 license. The videos shown to participants are from the Moments in Time (MiT) dataset (`http://moments.csail.mit.edu`); the MiT license prohibits public redistribution, but instructions for obtaining the videos are provided in McMahon et al. [2023], and the mapping from MiT filenames to our video IDs (0–249) is included in the repository.

H Social-Affective Behavioral Encoding

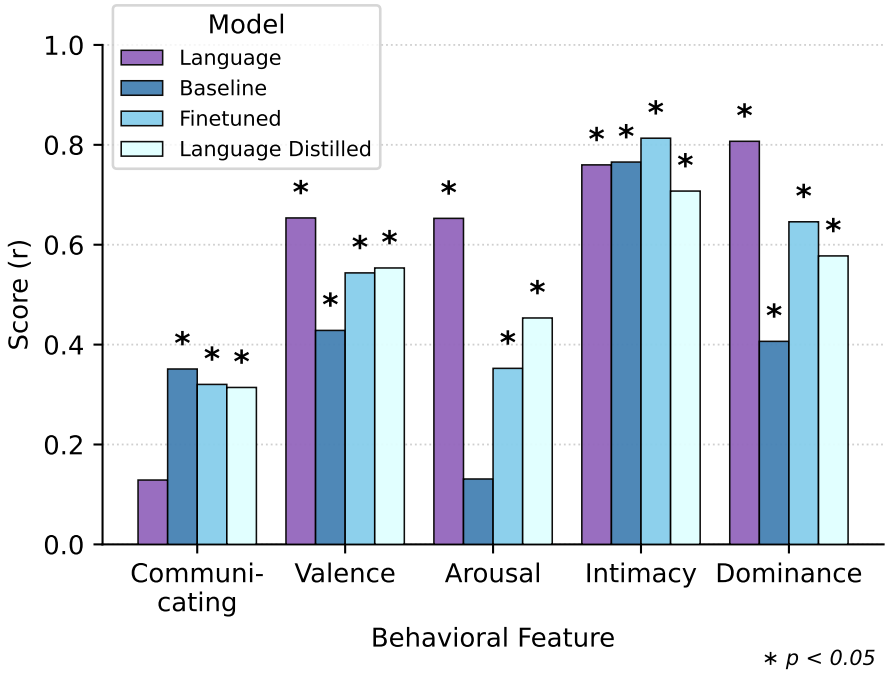

Figure 5: Ridge-regression prediction of social-affective attributes from stimulus embeddings, reported as Pearson correlation (r). MPNet provides the language-model comparison; V-JEPA-2.1 variants include baseline, fine-tuned, and language-distilled models. *: $p < 0.05$, permutation test with 5,000 shuffled held-out ratings.

I Action Recognition Performance

We include here the full results of the UCF101 linear-probe evaluation. All backbone parameters were frozen, and a linear classifier was trained on top of features extracted from the pretrained and fine-tuned TimeSformer and V-JEPA 2.1 models (`[CLS]` token for TimeSformer; mean-pooled patch tokens for V-JEPA 2.1). Training was repeated across three random seeds, and Top-1 accuracy is reported as mean $\pm$ standard deviation.

Table 1: Linear probe Top-1 accuracy (%) on UCF101 split1 with frozen backbones. Reported as mean $\pm$ standard deviation over 3 seeds.

Backbone	Top-1 (%)
Pretrained TimeSformer	95.75 ± 0.18
Fine-tuned TimeSformer	95.70 ± 0.14
Pretrained V-JEPA-2.1	78.17 ± 0.03
Fine-tuned V-JEPA-2.1	80.31 ± 0.06

J Kinetics-400 Action Recognition Performance

We additionally verify that behavior-guided fine-tuning does not degrade performance on Kinetics-400 [Kay et al., 2017]. The two backbones require different evaluation protocols, and we report each accordingly.

TimeSformer is distributed already fine-tuned for K400 classification, so its retention can be measured *training-free*: we merge the LoRA weights into the backbone, retain the original classification head, and read off accuracy. This evaluation is deterministic, so we report Wilson 95% confidence intervals rather than a seed distribution.

V-JEPA 2.1 is self-supervised and ships *no* K400 classification head, and our fine-tuned checkpoints carry a 128-d similarity readout rather than a classifier. K400 accuracy for V-JEPA 2.1 is therefore measured with a linear probe, following exactly the UCF101 protocol of Appendix I: the backbone is frozen, patch tokens are mean-pooled to a 768-d vector, and a linear classifier is trained on K400 with three random seeds. As in Table 1, the 128-d similarity head is stripped so that the frozen representation is the only difference between the pretrained and fine-tuned rows. V-JEPA-2.1 absolute accuracies are below published attentive-probe results because we use a linear probe on mean-pooled tokens; only the pretrained-vs-fine-tuned comparison is relevant here.

Table 2: Kinetics-400 validation accuracy (%) before and after behavior-guided fine-tuning, on the full 19,877-clip validation split. The two backbones use different protocols and the absolute values are *not* comparable across panels; the relevant quantity is the within-panel retention. V-JEPA 2.1 rows are mean $\pm$ standard deviation over 3 probe seeds; TimeSformer rows are a single deterministic pass with Wilson 95% CIs on Top-1.

Backbone	Top-1 (%)	Top-5 (%)	Retention (Top-1)
Reused K400 head (training-free)			
Pretrained TimeSformer	77.28 [76.69, 77.85]	93.03	—
Fine-tuned TimeSformer	76.32 [75.73, 76.91]	92.60	98.8%
Linear probe on frozen features (3 seeds)			
Pretrained V-JEPA-2.1	50.05 ± 0.12	75.18 ± 0.10	—
Fine-tuned V-JEPA-2.1	53.30 ± 0.04	77.79 ± 0.09	106.5%

K PHASE frozen-feature transfer accuracy

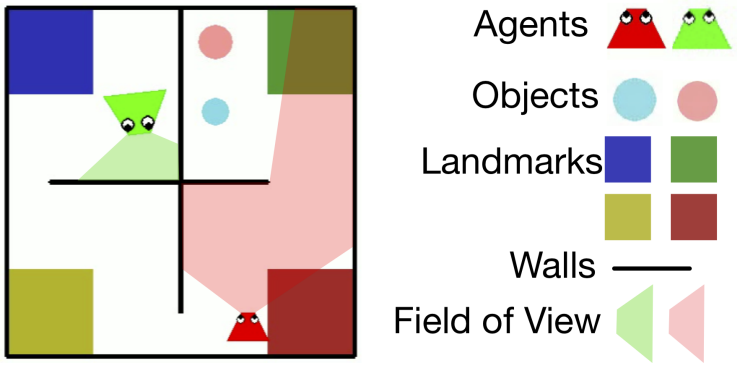

Figure 6: Representative stimuli from the PHASE dataset [Netanyahu et al., 2021]. Example frame shows two agents in a trial (colored triangles) interacting within a 2D environment with walls, obstacles, and goal objects. Trials span a range of social interaction categories including collaborate, help, hinder, follow, chase, and independent, which differ in the agents' goals and the contingency between their movements.

Prior baselines. The three prior baselines reported in Fig. 4 are: **SocialGNN** [Malik and Isik, 2023], a graph neural network that operates on *privileged* symbolic inputs (ground-truth agent positions, velocities, and pairwise distances) extracted from the PHASE simulator rather than on raw pixels; **VisualRNN** [Malik and Isik, 2023], a recurrent visual baseline that processes PHASE clips from raw pixels (no access to the underlying scene graph) and serves as the published pixel-level point of comparison; and **SiMPLE** [Netanyahu et al., 2021], a hand-designed, physics-grounded inverse-planning model introduced alongside the PHASE benchmark that infers latent goals and relations via Bayesian inverse simulation rather than learned representations. Reported numbers for SocialGNN and VisualRNN are taken from Malik and Isik [2023]; the SiMPLE number and the human agreement estimate are from Netanyahu et al. [2021].

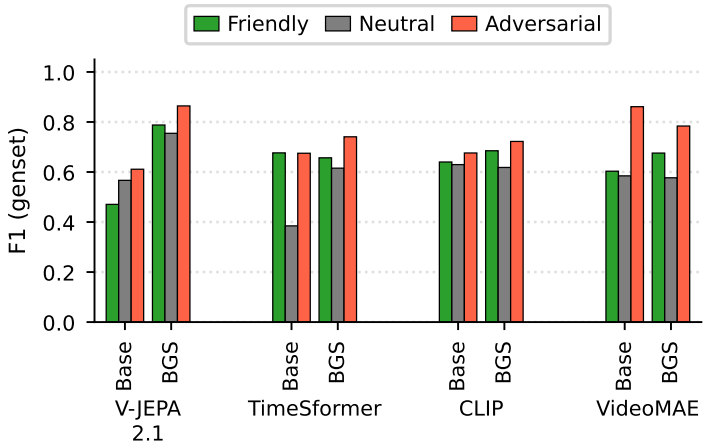

Figure 7: Phase 3-way classification (100 held-out generalization set videos), per-class F1 (Friendly, Neutral and Adversarial) for each pretrained (Base) and fine-tuned (BGS) backbone.

L TimeSformer Fine-tuned Variance Partitioning

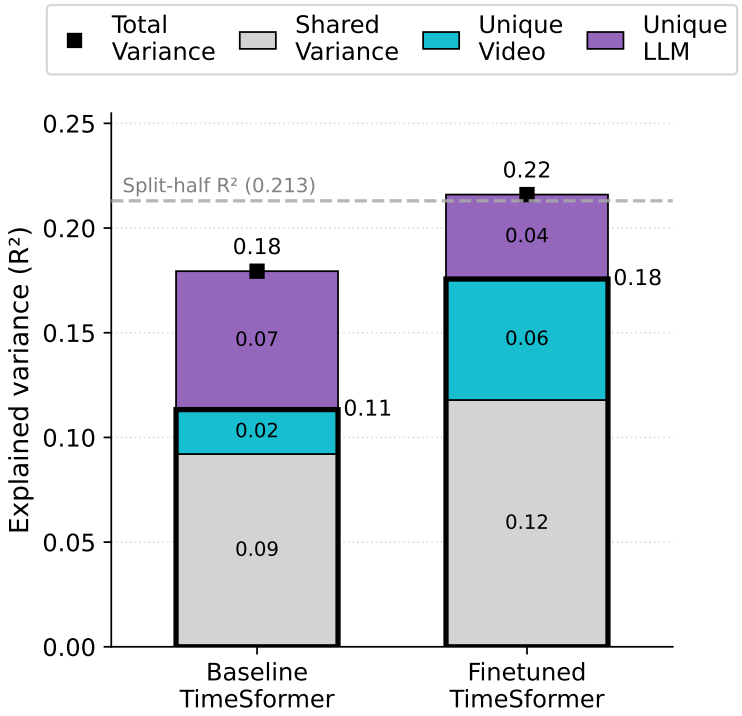

Figure 8: Variance partitioning between TimeSformer and MPNet before and after finetuning. Finetuning shifts explained variance from language-specific to shared and video-specific structure, bringing total variance near the split-half reliability ceiling (split-half ceiling is an estimate and can be exceeded). Black outlines indicate total variance explained by video-model.

M Model Performance and Supervision Budget

Table 3: Model performance and supervision constraints budget (— indicates not applicable).

Model UID	Explained Variance (R^2)	OOO Accuracy	Constraints/epoch
Fine-tuned Models			
vjepa2-vitl-k710-ft	0.172117	71.67%	12978
timesformer-ft-hybrid	0.162023	74.46%	12978
vjepa2.1-vitb-dist-ft†	0.160068	70.65%	12978
timesformer-ft-triplet-match	0.156857	66.58%	12978
clip-vit-b-32-hybrid-finetuned	0.155953	67.84%	12978
timesformer-ft-triplet	0.145600	70.65%	12240
videomae-hybrid-finetuned	0.123535	64.56%	12978
vjepa2.1-triplet-only-ft†	0.122122	69.29%	12978
timesformer-ft-rsa	0.121153	63.86%	13038
vjepa2.1-rsa-only-ft†	0.117029	65.48%	13038
vjepa2.1-mpnet-distilled†	0.088731	67.21%	12978
Video Models			
x3d-m	0.123559	68.48%	—
x3d-s	0.105202	64.67%	—
x3d-xs	0.103721	64.95%	—
timesformer-base	0.102408	63.59%	—
i3d-r50	0.094969	67.66%	—
c2d-r50	0.090121	65.76%	—
vjepa2-vitl-k710	0.086683	60.05%	—
videomae-base	0.086675	66.75%	—
slow-r50	0.086501	67.93%	—
slowfast-r50	0.085466	64.95%	—
vjepa1-vitl-meanpool-224-16f	0.052311	57.60%	—
vjepa2.1-vitb-dist†	0.042462	63.85%	—
Language Models			
paraphrase-multilingual-mpnet-base-v2	0.134374	70.38%	—
mxbai-embed-2d-large-v1	0.122445	66.58%	—
paraphrase-multilingual-MiniLM-L12-v2	0.120615	67.39%	—
distiluse-base-multilingual-cased-v1	0.110899	64.95%	—
paraphrase-MiniLM-L6-v2	0.102647	65.49%	—
all-distilroberta-v1	0.101303	63.04%	—
stsb-distilroberta-base-v2	0.098953	64.13%	—
mxbai-embed-large-v1	0.090592	67.39%	—
all-roberta-large-v1	0.088598	63.04%	—
all-mpnet-base-v1	0.086371	66.58%	—
all-mpnet-base-v2	0.085562	64.67%	—
all-MiniLM-L6-v1	0.078124	65.22%	—
all-MiniLM-L6-v2	0.077037	65.49%	—
multi-qa-MiniLM-L6-cos-v1	0.068142	64.40%	—
all-MiniLM-L12-v2	0.065997	67.39%	—
LaBSE	0.052770	61.96%	—
clip-ViT-B-32-multilingual-v1	0.052506	62.77%	—
FacebookAI/roberta-base	0.025612	59.24%	—
FacebookAI/xlm-roberta-base	0.022418	49.46%	—
FacebookAI/roberta-large-mnli	0.016395	47.83%	—
FacebookAI/xlm-roberta-large	0.010090	57.07%	—
Image Models			
clip-vit-b-32	0.126510	69.38%	—
dino-dino-vitb16	0.075451	60.77%	—
Vision-Language Models			
Qwen3-VL-2B-Instruct	0.128127	68.75%	—
lfm2-5-vl-450m	0.077825	65.76%	—

†Mean over 5 training seeds; other fine-tuned rows are single runs (seed 0).

Table 4: Subset – Fine-tuned and Baseline Models along with best baseline Video, Language, Image and Vision-Language model performance.

Model UID	**Explained Variance** (R^2)	**OOO Accuracy**
Fine-tuned		
vjepa2-vitl-k710-ft	0.172117	71.67%
timesformer-ft-hybrid	0.162023	74.46%
vjepa2.1-vitb-dist-ft†	0.160068	70.65%
timesformer-ft-triplet-match	0.156857	66.58%
clip-vit-b-32-hybrid-finetuned	0.155953	67.84%
timesformer-ft-triplet	0.145600	70.65%
videomae-hybrid-finetuned	0.123535	64.56%
vjepa2.1-triplet-only-ft†	0.122122	69.29%
timesformer-ft-rsa	0.121153	63.86%
vjepa2.1-rsa-only-ft†	0.117029	65.48%
vjepa2.1-mpnet-distilled†	0.088731	67.21%
Baseline Models		
clip-vit-b-32	0.126510	69.38%
timesformer-base	0.102408	63.59%
vjepa2-vitl-k710	0.086683	60.05%
videomae-base	0.086675	66.75%
vjepa1-vitl-meanpool-224-16f	0.052311	57.60%
vjepa2.1-vitb-dist†	0.042462	63.85%
Best Video Model		
x3d-m	0.123559	68.48%
Best Language Model		
paraphrase-multilingual-mpnet-base-v2	0.134374	70.38%
Best Image Model		
clip-vit-b-32	0.126510	69.38%
Best Vision-Language Model		
Qwen3-VL-2B-Instruct	0.128127	68.75%

†Mean over 5 training seeds; other fine-tuned rows are single runs (seed 0).

Matching Constraints. Despite the same number of optimizer steps across all approaches, the hybrid objective includes an additional RSA term, introducing a modest number of extra supervision signals (≈ 738 pairwise constraints per epoch) beyond the triplet loss ($12{,}240$ pairwise constraints). To ensure a fair comparison, we trained a *triplet-only (budget-matched)* variant by adding the same number of extra triplet constraints each epoch. This budget-matched triplet model slightly outperforms standard triplet-only training, confirming that more constraints help. Yet, it still under-performs compared to the hybrid model, indicating that the RSA term contributes qualitatively different information by enforcing global structure beyond what can be achieved by simply adding more triplet comparisons.

Table 5: Paired bootstrap of V-JEPA 2.1 ViT-B training objectives over the 637 observed test pairs (10,000 resamples; all models scored on identical resamples). Fine-tuned R^2 values are averaged over 5 training seeds per resample. Read-out layers were selected on the training-set RSM and held fixed; no test data informed layer selection.

Model	R^2	**95% CI**	
Hybrid (BGS)	0.160	[0.109, 0.216]	
Triplet-only	0.122	[0.076, 0.176]	
RSA-only	0.117	[0.074, 0.168]	
MPNet language-distillation	0.089	[0.050, 0.137]	
Pretrained	0.042	[0.016, 0.080]	
Paired contrast (hybrid − variant)	ΔR^2	**95% CI**	p
vs. Pretrained	$+0.118$	$[+0.065, +0.173]$	$< 10^{-4}$
vs. MPNet language-distillation	$+0.073$	$[+0.026, +0.120]$	0.002
vs. RSA-only	$+0.044$	$[+0.007, +0.083]$	0.020
vs. Triplet-only	$+0.039$	$[-0.005, +0.083]$	0.077

N Hyperparameter Sensitivity and Sample Efficiency

To test robustness, we swept the BGS hyperparameters on V-JEPA 2.1 ViT-B (single seed per configuration unless noted; 32 runs total). Across all runs, held-out test R^2 ranges from 0.101 to 0.176; every run exceeds the pretrained baseline (0.042) and 29/32 exceed the strongest language baseline (MPNet, 0.134). Alignment is insensitive to RSA batch size K, gradient-bearing subset M, LoRA rank, and β schedule shape (ranges ≤ 0.016). Larger RSA weight ($\beta_{\max}$) and triplet margin (γ) yield modest gains, and RSA cadence is the main sensitivity: reducing RSA updates per epoch degrades alignment, indicating that the global RSA objective requires sufficient updates to contribute. Several configurations exceed the reported one; we retain the pre-specified configuration rather than selecting post hoc, since the sweep is a robustness check rather than a tuning procedure.

Table 6: Loss weighting, fully crossed ($\alpha \times \beta_{\max}$; β ramps linearly from 0.3 to $\beta_{\max}$). Held-out test R^2. Bold: configuration reported in the paper.

	$\beta_{\max}$			
α	0.3	0.5	0.7	1.0
0.3	0.148	0.157	0.170	0.170
0.5	0.153	0.160	0.166	0.170
0.7	0.148	0.156	**0.163**	0.163
0.9	0.150	0.154	0.160	0.154

Table 7: Remaining axes, each varied from the reported configuration (bold). Held-out test R^2.

Axis	Values	Held-out test R^2
β schedule	fixed 0.7 / **ramp 0.3→0.7**	0.156 / **0.163**
RSA steps per epoch	2 / 6 / 12 / 24 / **106**	0.105 / 0.101 / 0.136 / 0.151 / **0.163**
RSA batch K	12 / **24** / 48	0.148 / **0.163** / 0.164
Gradient-bearing subset M	3 / **6** / 12	0.160 / **0.163** / 0.169
LoRA rank	4 / **8** / 16 / 32	0.154 / **0.163** / 0.148 / 0.150
Triplet margin γ	0.1 / **0.2** / 0.4	0.131 / **0.163** / 0.176
Seed (reported config)	42 / 43 / 44	0.163 / 0.160 / 0.155 (SD 0.004)

Sample efficiency. To quantify how much behavioral data BGS requires, we trained on random subsets of the 12,239 training triplets (3 seeds per size). Alignment improves monotonically with data, exceeds the pretrained baseline by 1,000 triplets, and recovers roughly half of the full-data gain with 2,500–5,000 triplets.

Table 8: Sample efficiency: held-out test R^2 (mean $\pm$ SD over 3 seeds) as a function of training triplets.

Training triplets	Held-out test R^2
500	0.046 ± 0.013
1,000	0.071 ± 0.035
2,500	0.098 ± 0.041
5,000	0.124 ± 0.016
10,000	0.142 ± 0.012
12,239 (full)	0.159 ± 0.004
Pretrained baseline	0.042

O Causal Occlusion of Social Regions

Table 9: Model–human alignment (R^2) under causal occlusion of social regions. Each cell re-scores the 50-video test set RSA (637 observed pairs; seed-0 checkpoint) with that region masked out (gray fill, ImageNet mean). Regions are ordered by mean masked area.

	Occluded region						
Model	none	hands	control	face	body	person	background
Mean masked area (% frame)	0.0	0.8	3.0	3.0	30.1	34.4	65.7
V-JEPA 2.1 (BGS fine-tuned)	0.1649	0.1639	0.1624	0.1154	0.0835	0.0467	0.0498
V-JEPA 2.1 (Pretrained)	0.0583	0.0564	0.0553	0.0442	0.0305	0.0256	0.0183

Note. `control` is the face mask translated to a person-free location, matched to `face` in size, shape and motion. `body` is the person mask excluding the face.

P Additional Methods

P.1 Sentence captioning of videos

Sentence captions were used from a publicly available dataset [see Garcia et al., 2025, for full details]. A group of 150 participants was recruited on Prolific to provide sentence captions. Eligibility criteria for this online study required participants to be native English speakers, 18 years or older (M=39.72, SD=13.24), with normal or corrected-to-normal vision. The cohort consisted of 63 females and 87 males. Self-reported race and ethnicity were as follows: 114 white, 14 black, 10 Asian, 9 mixed race, 2 other, and 3 who declined to report.

The task required each participant to write a single-sentence caption for 12 videos presented in a random order (10 standard and 2 catch trials). Participants typed their responses into a text box that initially showed a placeholder prompt: “Description of the actions and interactions of the people in the video in a single sentence...” (Appendix Fig. 9).

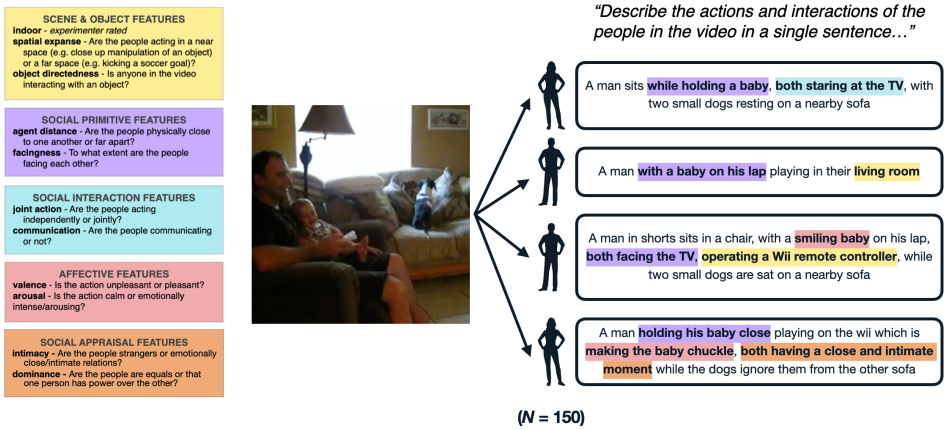

Figure 9: Example video and participant descriptions from the publicly available dataset from Garcia et al. [2025]. Participants viewed a short video clip (center) and were asked to describe the actions and interactions of the people in the scene using a single sentence. Example responses (right) show how different aspects of the same video can emphasize distinct feature categories: scene and object features (yellow), social primitive features (purple), social interaction features (blue), affective features (red), and social appraisal features (orange) (Modified from McMahon et al. [2023]). Each highlighted phrase corresponds to the feature dimension it represents.

P.2 Participant Demographics for Odd-One-Out Similarity Dataset.

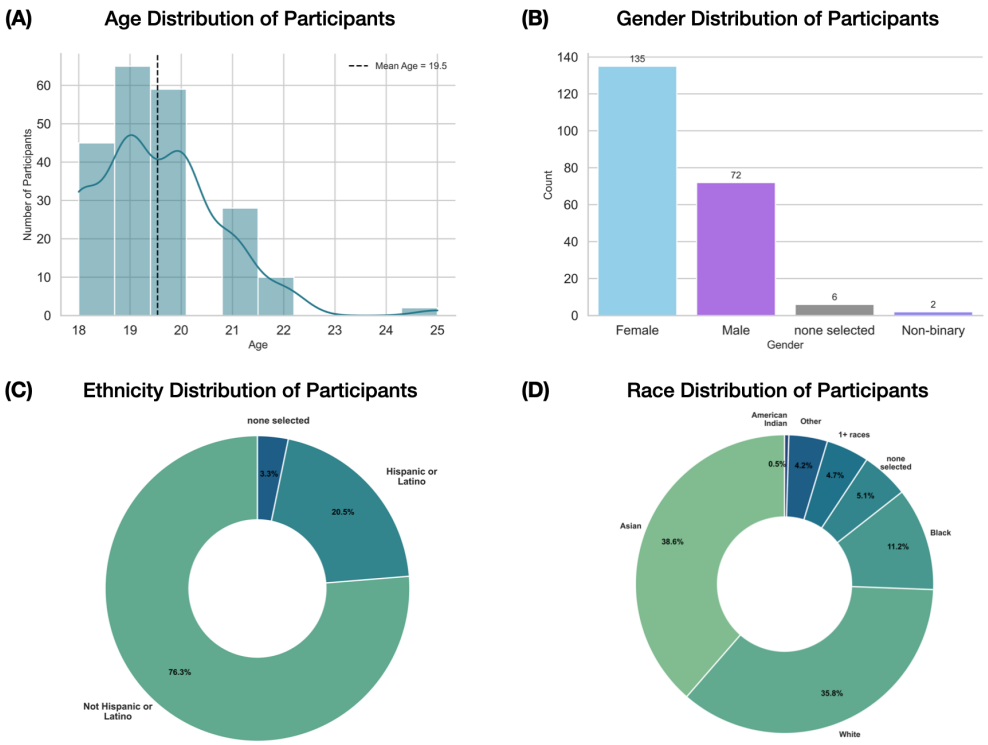

Figure 10: Participant Demographics. (A) Age distribution (mean = 19.5 years). (B) Gender distribution. (C) Ethnicity distribution. (D) Race distribution. All participants were recruited online via our University's psychological research platform, and participated in the study on Meadows Research (`https://meadows-research.com`). All participants gave informed consent in accordance with our internal Institutional Review Board, who provided explicit approval of all protocols and procedures discussed. All participants were 18 or older, with normal or corrected-to-normal vision, and at least academically proficient English speakers.

P.3 Video Action Category Distribution

Table 10: Action category frequencies across all 250 videos, showing the full distribution of annotated behaviors from the most common everyday actions to the least frequent, diversity-focused categories.

category	count	category	count
crying	21	building	2
laughing	18	boating	2
drumming	18	baking	2
brushing	18	closing	2
fishing	11	wrestling	2
dancing	10	bowing	2
giggling	9	gardening	2
cooking	9	shopping	1
discussing	8	digging	1
clapping	7	pushing	1
driving	6	unloading	1
smoking	6	exercising	1
working	6	drilling	1
eating	6	applauding	1
singing	5	spitting	1
planting	5	catching	1
reading	5	camping	1
bathing	5	barbecuing	1
playing	5	hugging	1
kicking	5	riding	1
mowing	4	chewing	1
hiking	4	cleaning	1
throwing	4	speaking	1
skating	3	playing+videogames	1
drinking	3	drawing	1
walking	3	skiing	1
dipping	3	jogging	1
hunting	3	studying	1
knitting	3	bowling	1
		unpacking	1

P.4 Proportions of Nouns and Verbs Across Video Captions

Table 11: Most frequent nouns and verbs across all captions, showing the proportion of word occurrence across captions.

Noun	Noun Proportion	Verb	Verb Proportion
man	0.255	play	0.173
baby	0.203	sit	0.083
woman	0.151	hold	0.067
child	0.133	look	0.065
girl	0.083	talk	0.052
boy	0.081	cry	0.051
people	0.058	watch	0.046
drum	0.052	stand	0.046
adult	0.044	brush	0.043
hand	0.041	laugh	0.038
tooth	0.039	sing	0.025
toddler	0.037	dance	0.024
water	0.032	appear	0.024
mother	0.031	smile	0.024
kid	0.026	try	0.022
toy	0.025	help	0.021
fishing	0.025	feed	0.021
person	0.024	make	0.020
car	0.024	use	0.019
lady	0.022	lie	0.019
kitchen	0.022	have	0.019
bed	0.021	walk	0.018
book	0.019	fish	0.018
hair	0.018	show	0.017
father	0.018	do	0.017
dance	0.017	seem	0.017
guy	0.017	put	0.017
food	0.017	eat	0.017
friend	0.017	lay	0.016
front	0.017	throw	0.014
floor	0.016	move	0.013
son	0.015	take	0.013
bath	0.015	read	0.013
arm	0.015	face	0.013
game	0.015	wear	0.012
male	0.014	speak	0.012
brother	0.014	explain	0.012
mouth	0.014	get	0.011
guitar	0.014	catch	0.010
ball	0.014	work	0.009
side	0.013	learn	0.009
dog	0.013	prepare	0.009
playing	0.012	smoke	0.009
room	0.012	cook	0.009
dad	0.012	interact	0.009
ice	0.011	discuss	0.009
microphone	0.011	drink	0.009
boat	0.011	clean	0.009
time	0.011	drive	0.009
face	0.011	practice	0.008

Q Supplemental Qualitative Interpretability

Q.1 Model-Human Agreement for the Same Set of Triplets

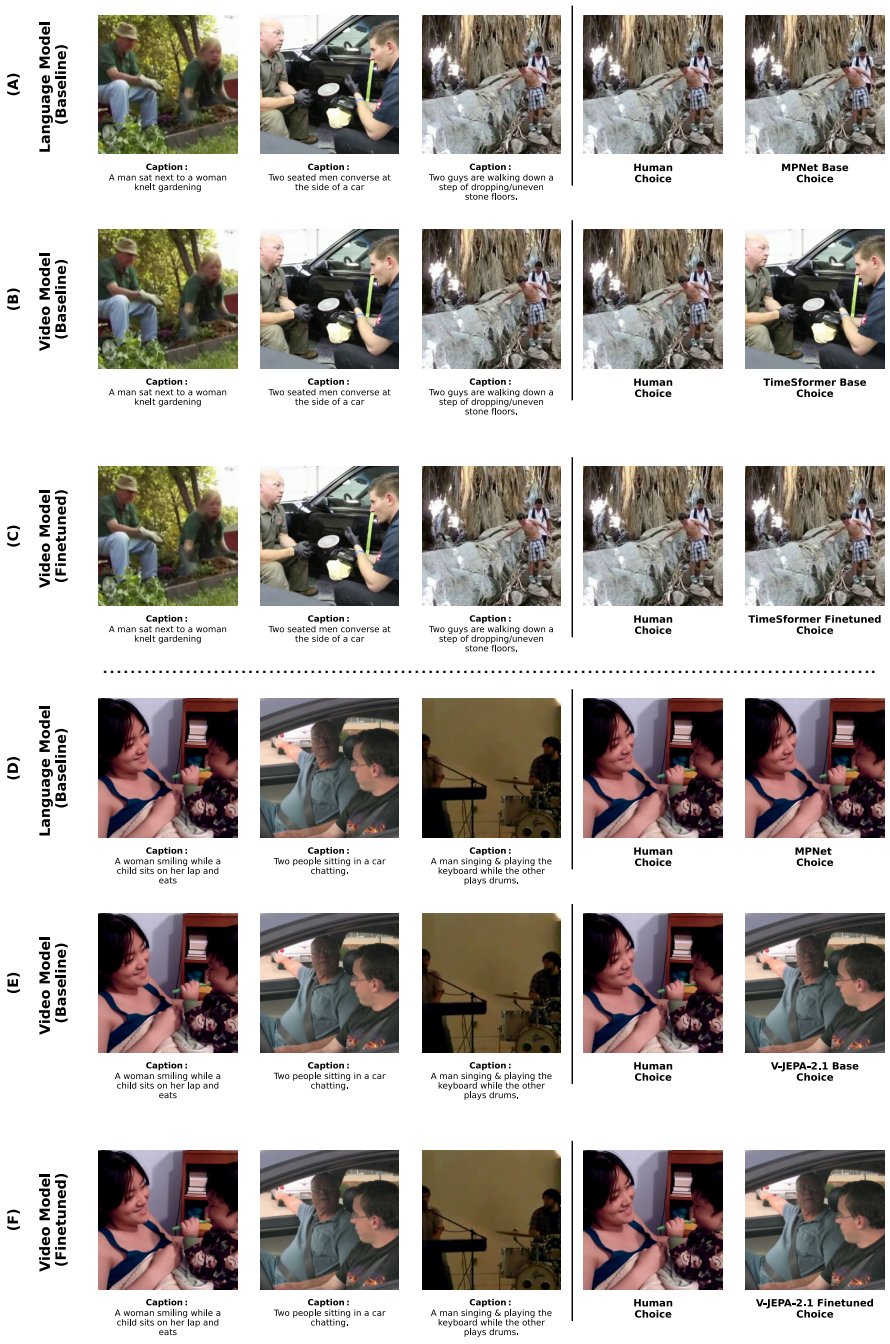

Figure 11: Human–model agreement and disagreement on triplet odd-one-out judgments across three modalities (A, B, C) baseline MPNet, baseline V-JEPA-2.1 and fine-tuned V-JEPA-2.1 respectively where V-JEPA-2.1 agrees with human judgment after finetuning. (D, E, F) baseline MPNet, baseline V-JEPA-2.1 and fine-tuned V-JEPA-2.1 respectively where V-JEPA-2.1 disagrees with human judgment after finetuning.

Q.2 Qualitative Examples of Human–Model Agreement



Figure 12: Human–model agreement and disagreement on triplet odd-one-out judgments across three modalities *(A) fine-tuned TimeSformer*, *(B) baseline TimeSformer*, and *(C) baseline MPNet*, where each row shows the three candidate videos, the human-selected odd one out, the model-selected odd one out, and the human-written captions, with horizontal separators marking modality groups and row labels indicating agreement or disagreement.

Q.3 Attention Rollout Comparison - V-JEPA-2.1



Figure 13: Spatial attention rollout for V-JEPA 2.1 ViT-B before and after BGS fine-tuning. We select the five stimuli with the largest attention divergence between models, quantified as $1 - \bar{r}$: one minus the mean Pearson correlation between baseline and fine-tuned per-frame saliency maps, averaged over the model's 8 temporal slices. **Left:** original frame. **Center:** pretrained model, which concentrates attention on a single focal region (typically one agent's torso or limbs). **Right:** fine-tuned model, which distributes attention across socially informative regions—faces, gaze direction, interacting hands, and both agents—without any spatial supervision during training. $1 - \bar{r}$ values for the five videos shown are 0.30, 0.29, 0.61, 0.29, and 0.37, compared to the dataset-wide mean of 0.18 ± 0.03. Even among these maximally divergent cases, the shift is consistently toward socially informative regions rather than arbitrary redistribution; the narrow dataset-wide variance confirms systematic reweighting of spatial priority across the full 250-video stimulus set.

Q.4 Attention Rollout Comparison - TimeSformer

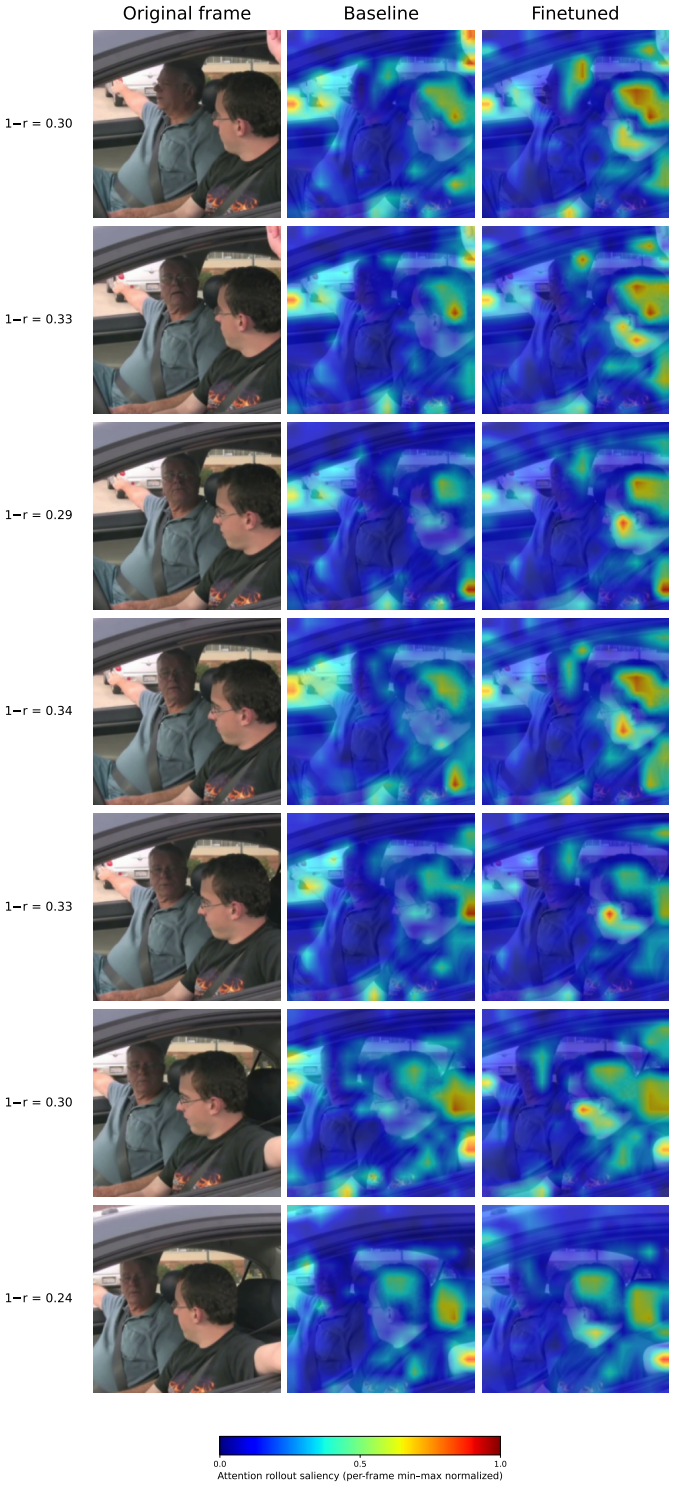

Figure 14: Spatial attention rollout for TimeSformer before and after BGS fine-tuning. We select a representative stimuli and show attention over the model's 8 temporal slices. **Left:** original frame. **Center:** pretrained model. **Right:** fine-tuned model, which shifts attention across socially informative regions—heads, faces, gaze direction, and both agents—without any spatial supervision during training.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 (Contributions paragraph) state three specific claims: (i) a ∼49k-judgment human OOO benchmark on 250 social videos showing a lang-vision gap across 14 pretrained models (see Fig. 2) (ii) BGS, a hybrid triplet+RSA fine-tuning objective that uses RSA as a differentiable training signal trained on four ViT backbones (§4.2) and (iii) empirical gains in human alignment and show transfer, with frozen features, to abstract social interactions (§4.4), reorganizes spatial attention to socially informative regions (§4.5) and preserves UCF101 action recognition (Appendix I).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §5.3 explicitly discusses three limitations: dataset coverage (single source corpus, no cross-dataset transfer tests), evaluator subjectivity (aggregate-consensus only, no personalization), and task scope (alignment evaluated on similarity plus five attribute probes; broader task trade-offs untested beyond UCF101).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper contributes an empirical method (behavior-guided fine-tuning) and a dataset; it does not claim any theoretical results requiring formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The Methods section (§3) specifies data splits (200 train / 50 test videos; $\sim$24k / $\sim$368 triplets), backbone (TimeSformer + LoRA rank=16, $<$2% trainable params), optimizer (AdamW, lr=1×10^{-4}), schedule (50 epochs, batch size 4, mixed precision, gradient-checkpointing), loss weights ($\alpha = 0.7$, β ramped $0.3 \rightarrow 0.7$, $\gamma = 0.2$), RSA sampling ($K = 24$, $M = 6$, 6 RSA steps/epoch), and layer-selection protocol (5-fold CV on train set). Appendix §E describes the triplet-selection algorithm used to build the human RSMs. The Reproducibility Statement (Appendix §A) summarizes all details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: As stated in the Reproducibility Statement (Appendix §A), we release the 250-video stimulus list, captions, ∼49k OOO judgments with train/test splits, per-split human RSMs, the triplet selection script, the full fine-tuning configuration, evaluation scripts, and the trained LoRA adapters for all backbones, code and data for the UCF101 and social-affective linear probes, available at `https://github.com/garciakathy/bgs-finetuning`. The Moments in Time video subset, themselves, cannot be redistributed under the MiT license; we provide the ID-to MiT mapping that users can obtain directly from the sources (Appendix §G)

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In Methods (§3.3) and the Reproducibility Statement (Appendix §A) list the train/val/test splits (200 train videos with a 20% held-out validation slice for checkpoint selection; 50-video test set), optimizer (AdamW), learning rate, batch size, epochs, LoRA rank, triplet margin, RSA sampling parameters, and the loss-weighting schedule. Layer selection for evaluation used 5-fold CV on the training set with Spearman's ρ as the selection criterion.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The UCF101 sanity check (Table 1) reports mean $\pm$ standard deviation over 3 random seeds; Explained variance between models and human RSMs in Fig. 2 and social-affective linear-probe results in Fig. 5 mark significance with $*: p < 0.05$; the test-set reliability ceiling $R^2 = 0.213$ is derived from human split-half correlations. For the V-JEPA 2.1 training objectives (hybrid, triplet-only, RSA-only, MPNet distillation, pretrained), we train 5 seeds and report paired bootstrap 95% CIs and p-values over the 637 observed test pairs (10,000 resamples; Tab. 5, §4.2); Fig. 2 plots these as 5-seed means, without error bars. Other fine-tuned and pretrained rows in Table 3 are single-run point estimates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report a rough estimate of a little over 150 GPU-hours of compute across all 48 models (21 language encoders, 2 image encoders, 2 VLMs, 12 pretrained video backbones, and 11 finetuned video models). Across this compute, the predominant cost is associated with our finetuning process, with a single V-JEPA 2.1 ViT-B + LoRA hybrid training taking $\sim$2.5 GPU-hours alone (12 epochs on a single A100, `num-epochs=50`, `early-stopping=5`). When we include all V-JEPA 2.1 and TimeSformer ablations, and the hybrid finetuning process for the baselines of CLIP and VideoMAE, finetuning as a whole consumed about 110 GPU-hours. Additional seeds and data sampling runs cost around 17.5 GPU-h. The remaining compute can be attributed to the checkpoint evaluation across different tasks, including behavioral RSA ($\sim$17 GPU-h), OOO triplet accuracy ($\sim$2 GPU-h),

the PHASE out-of-distribution probe (∼3 GPU-h), and attention-rollout visualizations (∼0.5 GPU-h). All experiments ran on HPC cluster on a node of 4x NVIDIA A100 (80 GB) GPUs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research followed the NeurIPS Code of Ethics. Human similarity judgments were collected under Institutional Review Board approval with informed consent from all participants (see §3.3 and the Ethics Statement, Appendix §B), and participants were compensated through the online platform used for recruitment.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Discussion (§5) and Conclusion (§6 sections highlight the positive impact of aligning AI representations with human social perception for interpretability and human-centric applications. On the negative side, the Limitations section (§5.3) notes that the dataset reflects aggregate-consensus judgments from a specific participant pool, so models fine-tuned on it could encode cultural or demographic biases in social perception if deployed without the subgroup-level alignment discussed there.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifacts are (a) aggregated odd-one-out similarity judgments over publicly available short social video clips from an existing, IRB-cleared stimulus set and (b) a discriminative LoRA adapter for a video encoder used for representation alignment—not a generative model. These assets do not pose a foreseeable high-risk-of-misuse profile comparable to large pretrained generative models or scraped image/text datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used are cited: the Moments in Time source corpus [Monfort et al., 2019] (with a non commercial research license; videos not redistributed) and the social-video subset and annotations [McMahon et al., 2023]; the TimeSformer backbone [Bertasius et al., 2021] and other evaluated video encoders (X3D, SlowFast, I3D, C2D); the language models listed in Table 3; DeepJuice [Conwell et al., 2024]; LoRA [Hu et al., 2022]; and UCF101 [Soomro et al., 2012]. We use each asset under its original license and release only the judgments we collected plus our LoRA adapter weights on top of the publicly released TimeSformer checkpoint.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets, i.e., ∼49k OOO triplet judgments with train/test splits, per-split human RSMs, the triplet-selection code, fine-tuning configuration, and LoRA adapter weights, are described under Methods (§3.1 and the Reproducibility Statement (Appendix §A), and will be released with a documentation/README describing data schema, collection protocol, splits, and intended use.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: §3.1 reports the participant count ($n = 245$), the online recruitment platform (Meadows Research, `https://meadows-research.com/`), the verbatim task instruction given to participants ("focus on what the people are doing and choose the odd-one-out"), and the triplet-design structure (∼220 trials per participant). Appendix E details the triplet-selection procedure used to construct trials. Full instruction text and a screenshot will be included in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Under Methods (§3.1), it states that all participants gave informed consent under JHU Homewood Institutional Review Board (IRB) that provided explicit approval of all protocols and procedures. The study involved viewing short everyday social video clips and making similarity judgments, which carries minimal risk. (Appendix §B).

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a core methodological component of this work. Pretrained sentence/caption embedding models (e.g., MPNet, MiniLM, RoBERTa variants) are used only as off-the-shelf baselines for caption-based representational similarity and are cited and listed in Table 3; they are not used to propose or implement the fine-tuning method itself.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.