# Multiarrangement: A Plug & Play Geometric Data Collection Package For Video Stimuli

Umur Yıldız[1,2] and Burcu A. Ürgen[1,2,3]

[1]Department of Neuroscience, Bilkent University, Ankara, Turkey
[2]Aysel Sabuncu Brain Research Center and National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey
[3]Department of Psychology, Bilkent University, Ankara, Turkey

**Abstract**

We present Multiarrangement, an offline, open-source Python toolkit for collecting human similarity judgments for video stimuli through multi-arrangement tasks. Participants arrange subsets of stimuli in a 2D arena such that Euclidean distances reflect perceived dissimilarity. The toolkit supports two experimental paradigms: a set-cover scheduling system that uses combinatorial covering designs, aiming to avoid overwhelming the participant for stimulus-rich settings, and an adaptive Lift-the-Weakest scheduler that focuses each new trial on the globally least certain pair and informative neighbors. Across trials, partial distance evidence is fused into a representational dissimilarity matrix. Further refinement is also available with optional reliability-like weighting and inverse MDS which can reduce cross-trial prediction error. We document task design, algorithms, a small within-subject validation, and provide practical guidance for reliable use.

*Keywords:* multiarrangement, video stimuli, representational dissimilarity matrix (RDM)

## Introduction

Video has become a first-class stimulus in many behavioral and neural research paradigms. Naturalistic movies capture rich, time-varying perception and cognition and are now common in neuroimaging (Hanke et al., 2014; Liu et al., 2022), where the shared narrative time-locks cortical responses across viewers (Hasson et al., 2004) and yields more reliable, behavior-predictive functional connectivity than rest (Finn et al., 2020). At the behavioral level, validated film-clip corpora reliably elicit target emotions and support ecologically valid laboratory experiments (Gilman et al., 2017; Gross & Levenson, 1995; Schaefer et al., 2010). Naturalistic videos also strengthen behavioral paradigms by preserving multimodal dynamics, narrative structure, and social context (Baldassano et al., 2017; Nastase et al., 2020; Sonkusare et al., 2019). This results in measurements that generalize better to everyday cognition and exposes processes such as event segmentation and predictive processing that are often not well captured by static stimuli (Baldassano et al., 2017; Nastase et al., 2020; Sonkusare et al., 2019). Furthermore, videos differ from their static counterparts in that they carry a manipulable temporal and audiovisual structure. Researchers can reorder shots or remove sound while keeping content realistic, and these manipulations have measurable effects on reliability and cognition (Hasson et al., 2008; Jääskeläinen et al., 2021; Lerner et al., 2011).

Although the value of video stimuli is better appreciated, most video-rating workflows are sequential, where participants watch a clip and provide a rat-

ing on a scale. Sequential methods are well matched to scalar targets (MOS, valence), but they are inefficient for capturing relational structure among many clips, and paired-comparison protocols or triplet designs scale poorly ($O(N^2)$ or $O(N^3)$). Contemporary continuous-annotation toolchains for video provide frame-level valence/arousal traces, yet still yield unidimensional trajectories per pass (Baveye et al., 2015; Koelstra et al., 2012). What is missing for video-centric studies is a method that allows participants to see and compare multiple clips at once, while scaling to dozens of stimuli without overwhelming either the display or the participant. No off-the-shelf, non-commercial software package currently exists to facilitate this type of data collection for video stimuli.

Multiarrangement addresses this gap. Participants arrange small subsets on a two-dimensional canvas, and we aggregate distances across trials into a single representational dissimilarity matrix (Kriegeskorte et al., 2008; Nili et al., 2014). Each placement yields many informative pairwise constraints, which reduces the quadratic burden of exhaustive pairwise ratings while preserving fine-grained structure. Balanced subset schedules give broad coverage with modest trial counts and reduce scale-use and anchoring artifacts that often appear in scalar ratings. For controlled studies, the resulting stimulus space supports targeted sampling, pre- to post-training comparisons of category structure, and explicit tests of dimensional organization using a single dataset. For cognitive neuroscience the same dissimilarity matrix fits directly into representational similarity analysis and model-based encoding or decoding, enabling comparisons between behavior and brain across images, audio and video, with straightforward reliability checks via split halves and cross-validation. As a library, it provides reproducible scheduling, streamlined interfaces for static and time-varying stimuli, detailed trial logs, and easy integration for downstream analyses.

# Features and Functionality

The design of Multiarrangement focuses on providing a turn-key solution for geometric data collection by combining a straightforward user interface with an efficient back-end for trial management. The goal is to offer a tool that is both accessible for the researcher and effective for data collection. This is accomplished through two main components, which we detail in the following sections:

## User Interface

The user interface of Multiarrangement adheres to a design similar to prior multi-arrangement tools (Kriegeskorte & Mur, 2012), but is tailored for the specific demands of dynamic media. Multiarrangement offers a workspace in which a single circular arrangement area is present. Stimulus tokens initially appear in a seating area outside the circle, from which

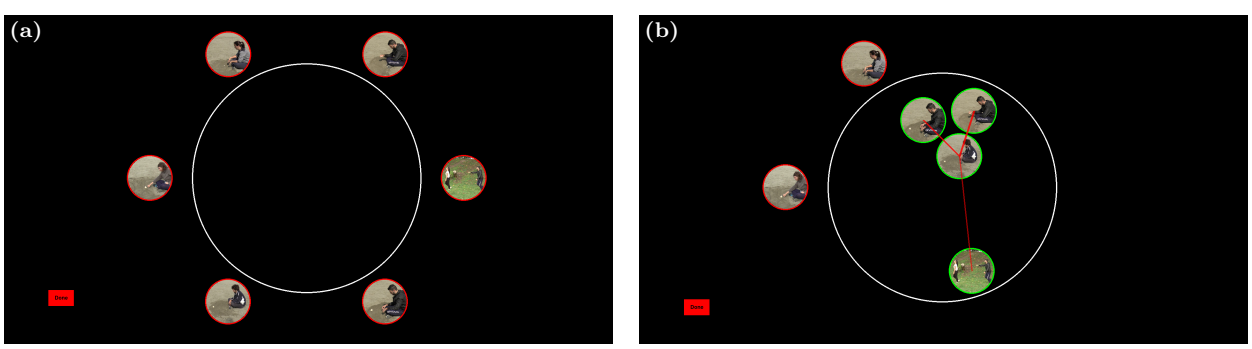


**Figure 1.** *Multiarrangement user interface.* The workspace at the start of a trial (a) shows the central circular arrangement area and the peripheral seating area. During placement (b), dragging a token displays transient guidance lines whose thickness/opacity scale with proximity.

the user must move them. This two-region design is a deliberate choice, adhering to the paradigm established in seminal multi-arrangement research. The circular workspace is specifically employed to minimize layout biases, as it lacks the corners or implicit axes of a square that might otherwise influence participant placements (Figure 1, panel a). For large-subset trials, an optional, center-locked zoom allows local magnification for precise placement.

The primary mode of interaction is direct manipulation via drag-and-drop. Participants use the mouse to move stimulus tokens (typically thumbnails for videos and images, or icons for audio) from the seating area into the circular workspace. A critical feature, designed specifically for dynamic media, is the ability to inspect stimuli on demand. By double-clicking any stimulus token, the participant can play the associated video in a pop-up window. This allows for repeated review and fine-grained comparisons, a necessary function when judging the similarity of complex, time-varying stimuli.

While a stimulus token is being dragged, transient guidance lines connect it to every other token currently in the circle. Line thickness/opacity scale with proximity, providing a quick visual sense of relative distances (Figure 1, panel b). By default, a trial can be submitted only after every token in the current trial's subset has been inspected at least once and all tokens are inside the circle.

The interface also uses colored rims around stimulus tokens to signal compliance status: red indicates the requirements for submission pertaining to the specific token have not yet been fulfilled, and green confirms the token satisfies the per-trial requirement. Default instruction sets are provided in Turkish and English, and the interface supports user-specified instructional text and labels via simple templates to enable customization.

## Scheduling Algorithms

The scheduler selects, on each trial, a subset of stimuli that balances simultaneous comparison against per-trial effort while accumulating pairwise evidence toward a reliable RDM (Representational Dissimilarity Matrix) within a fixed session budget. We provide two complementary strategies. First is a Set-Cover mode that precomputes small batches that together cover all unordered pairs and yields a fixed, between-subject-comparable schedule. In this approach we seed from published $(v, k, t{=}2)$ coverings in the La Jolla Covering Repository (Gordon, 2025; Gordon et al., 1995) and try to provide more balanced batches. The second strategy is an adaptive mode that maintains an evidence matrix, selects the globally weakest pair as anchors, and expands to a bounded subset until an evidence threshold or time limit is reached. We follow the "Lift-the-Weakest" formulation of Kriegeskorte and Mur (2012) with some modifications.

### Set-Cover Implementation

**Problem setup** Let $\mathcal{I} = \{1, \ldots, N\}$ index the stimuli. A trial presents a batch $B \subset \mathcal{I}$ of size $k$, covering the intra–batch pairs $\mathcal{P}(B) = \{(i, j) : i < j,\ i, j \in B\}$. We seek a small family $\mathcal{B} = \{B_m\}_{m=1}^{M}$ such that $\bigcup_{m=1}^{M} \mathcal{P}(B_m) = \{(i, j) : i < j\}$ while bounding $k$ to avoid on-screen overload and keeping $M$ small enough to fit the session budget.

**Construction** When available, we seed from published $(v{=}N, k, t{=}2)$ coverings in the La Jolla Covering Repository, otherwise we initialize greedily by iteratively adding a batch that covers the largest number of yet-uncovered pairs. We then apply a light refinement pipeline consistent with the released code: *repair/prune* to guarantee full coverage while removing batches that add no unique pairs, followed by *local search + group DFS* to smooth per-item load via item swaps and small group moves without breaking coverage. A shrink-only variant (*flex*) may reduce some late batches to $k_{\min}$ when this closes residual gaps or trims $M$ without increasing per-trial burden.

**Execution controls for video** We show one batch per trial. For video, we keep $k$ in a narrow range (e.g., $6 \leq k \leq 10$) to bound playback and interaction time, randomize initial token positions on the screen to reduce placement bias, and interleave batches so that no stimulus appears in adjacent trials beyond

```
# C_ij: pair-coverage counts; deg(i): per-item load
Input: N, k, k_min, seed, flex
rng <- init_rng(seed)

if LJCR_has(N, k) then
   B <- LJCR_seed(N, k) # published (v,k,2) blocks
else
   B <- greedy_seed(N, k, rng) # add batches
        covering most uncovered pairs
end if

if flex then
   B <- flex_shrink(B, k_min) # shrink late
        batches to k_min if helpful
end if

B <- repair_prune(B) # repair: cover all uncovered
     pairs; prune: drop batches with no unique
     pairs (keep C_ij >= 1)

B <- local_search_groupDFS(B, rng) # swaps/one-out-
     one-in/small-group moves; accept only if
     coverage preserved (C_ij >= 1) and var(deg)
     decreases

return B # final batch schedule
```

**Figure 2.** Set-Cover scheduler.[1] Initialization uses a published $(v, k, 2)$ covering when available or a greedy seed otherwise. `flex_shrink` can trim late batches to $k_{\min}$. `repair_prune` enforces full pair coverage and removes batches contributing no unique pairs, and `local_search_groupDFS` preserves coverage while reducing exposure variance via small swaps/group moves. Returns ordered batches and coverage diagnostics.

a user-specified limit. The schedule, RNG seed, and coverage diagnostics are recorded for reproducibility.

**Set-Cover for video** Set-Cover fixes the per-trial set size $k$ and, for fixed $k$, reduces total trials relative to naive pairwise enumeration by approximately a factor of $\binom{k}{2}$. Participants compare a small, consistent number of clips per trial while every unordered pair is co-presented at least once. To control duplication, we refine the schedule to remove redundant batches and balance per-item exposure $\deg i$ via local swaps and prune & repair iterations, which minimizes repeated co-occurrences and distributes necessary repeats across items and pairs. For remaining repeats, we aggregate distances using either an RMS-matched estimator, which aligns each trial's scale to the running estimate, a hybrid estimator, which keeps that alignment while giving more weight to clearer separations, or a maximum distance scaled variant, which normalizes each trial by its largest distance for bounded, simple updates. We do not renormalize the global matrix between batches. Each trial is aligned to the current scale using only overlapping already observed pairs. A single final rescaling to unit off-diagonal RMS is applied after all pairs are observed. These estimators also come with optional robust Winsor or Huber reweighting and a subsequent optional inverse MDS step. Together, these steps preserve the time efficiency and manageable on-screen complexity of a fixed-batch design while yielding a smoother, globally consistent RDM.

[1] Algorithm pseudocode listings are framed in blue, and runnable code examples are framed in black.

### Lift-the-Weakest Implementation

**Problem setup** Adaptive scheduling maintains an evidence matrix $W \in \mathbb{R}^{N \times N}$ that accumulates support for pairwise dissimilarities across trials. At each step the scheduler targets the least supported region while keeping per-trial effort bounded, following the Lift-the-Weakest principle (Kriegeskorte & Mur, 2012).

**Policy** Lift-the-weakest begins with an initial rating step (Trial 0) in which all $N$ items are displayed simultaneously. Each subsequent trial anchors on the globally weakest pair $(a, b) = \arg\min_{i<j} W_{ij}$. Starting with $S = \{a, b\}$, items are added greedily up to a size bound to increase expected information, preferring candidates that create low-evidence links to members of $S$. After a layout is collected, the trial is off-diagonal RMS-matched to the current estimate. Weights are computed from the raw on-screen distances and increase quadratically with distance, while the numerator uses the RMS-matched scaled distances. The run stops when the minimum utility $u(W) = 1 - \exp(-\kappa W)$ exceeds a threshold $u^\star$ or, equivalently, when the minimum evidence threshold

```
# Inputs: N, k_min, k_max, target_time, kappa,
     u_star, w_star, max_trials, time_budget, seed,
     alpha=2.0
rng<-init_rng(seed);
t0<-now()
W,Num,D_hat <- zeros(N,N), zeros(N,N), zeros(N,N);
     viewed,recent<-zeros(N),zeros(N); subsets<-[]

def fuse_update(S, D_obs):
  r_D <- RMS_offdiag(D_hat[S,S]); r_O <- max(
       RMS_offdiag(D_obs), eps)
  s <- (r_D if r_D>eps else 1) / r_O # Trial-0 guard:
        if r_D<=eps, use s=1/r_O to seed D_hat from
       first layout
  for (i,j) in pairs(S): w<-max(D_obs[i,j],0)^alpha;
       d<-s*D_obs[i,j];
    W[i,j]+=w; W[j,i]+=w; Num[i,j]+=w*d; Num[j,i]+=w*
         d;
    Dij<-Num[i,j]/max(W[i,j],eps); D_hat[i,j]=Dij;
         D_hat[j,i]=Dij
  D_hat<-rescale_offdiag_RMS1(D_hat); set_diag_zero(
       D_hat)

S<{1..N};fuse_update(S,pairwise_distances(
     present_and_collect(S))); t0<-now()

for t in 1..max_trials:
  if now()-t0>=time_budget or min(1-exp(-kappa*
       W_offdiag))>=u_star or min(W_offdiag)>=w_star:
        break
  (a,b)<-argmin_offdiag(W); S<-{a,b}
  while |S|<k_max and est_cost(S)<=target_time:
    x<-argmax_x( (sum_{(i,j) in aff(x,S)} [u(W[i,j]+
         dW[i,j]) - u(W[i,j])]) / time_cost( S{x})
               - penalties(x,S,viewed,recent) );
    if x==None: break; S<- S{x}
  if |S|<k_min: S<-pad_with_cheapest(S)
  fuse_update(S, pairwise_distances(
       present_and_collect(S)))
  viewed[S]<-True; recent<-decay(recent); recent[S
       ]+=1; subsets.append(S)

return {W, D_hat, subsets}
```

**Figure 3.** Lift-the-Weakest with initial rating step. Trial 0 shows all $N$ items to orient participants. Each subsequent step anchors the current weakest-evidence pair under $W$, grows subset $S$ within size/-time bounds to maximize expected gain, collects a layout, RMS-matches to $\hat{D}$, updates $W, \mathrm{Num}, \hat{D}$ using weights based on raw on-screen distances$^{\alpha}$ with the numerator computed from RMS-matched scaled distances, renormalizes to off-diagonal RMS $= 1$, and stops at utility threshold $u_\star$, minimum evidence threshold $w_\star$, or the time cap.

is reached $\left(\min_{i<j} W_{ij} \geq w^\star = -\ln(1-u^\star)/\kappa\right)$, or finally when a trial or time cap is reached (Kriegeskorte & Mur, 2012).

**Execution controls for video** Subset size is bounded $k_{\min} \leq |S| \leq k_{\max}$ to keep playback and interaction time practical. Light diversity constraints limit consecutive reuse of the same anchors and discourage high Jaccard overlap between successive subsets. A limited-iteration inverse MDS refinement can be applied intermittently or at the end to improve cross-trial consistency without a large runtime (Kriegeskorte & Mur, 2012).

**Video-specific adaptations** We provide an adaptive Lift-the-Weakest (LtW) mode (Kriegeskorte & Mur, 2012) and introduce small, pragmatic modifications for dynamic video stimuli. Expansion uses a cost-aware utility that divides expected evidence gain by an estimated per-item review cost based on clip duration and recent playback behavior, aiming to keep per-trial time near a target. To avoid starving long clips, we cap the per-item cost estimate and enforce a small minimum inclusion rate for long-duration items. Subset size is chosen within bounds to satisfy a per-trial time target rather than a fixed $k$ (i.e., $k_t \in [k_{\min}, k_{\max}]$). Light diversity constraints use soft penalties to limit anchor reuse and discourage excessive Jaccard overlap between successive subsets to reduce fatigue without overriding information gain. Boundary-focused expansion prioritizes candidates that increase evidence around locally uncertain regions by combining low-evidence links with a simple local stress score. Evidence updating supports three modes. The default hybrid mode RMS matches each trial to the current estimate, computes residuals on the RMS scale, and weights pair updates by raw on-screen distance raised to $\alpha$. Winsorization applies to these raw-distance weights and Huber reweighting applies to residuals. The RMS-only mode RMS matches each trial and uses uniform pair weights. Huber reweighting still applies and Winsorization has no effect in this mode. The max-scaled mode rescales each trial so its largest pair spans the layout. It supports either uniform weights or raw-distance weights

with the same options. In the hybrid and RMS-only modes we renormalize the RDM after each fuse so that the off-diagonal RMS equals 1. In the max-scaled mode this renormalization is optional and is off by default in order to preserve the interpretation that the largest pair spans the layout. An inverse-MDS update can be applied intermittently or at the end; we use a small, fixed number of iterations in practice to balance runtime and stability. Defaults and exact operationalizations of these parameters (cost model, stress score, diversity penalties) and their usage are documented in our GitHub repository.[2]

# Validation

**Protocol** We compared multi-arrangement RDMs against one-by-one comparison RDMs within-subject (n=2) on the 58-item set of natural human action videos from a validated corpus (Urgen et al., 2023). Multi-arrangement trials followed a fixed Set-Cover schedule with $N = 58$ and batch size $k = 8$, guaranteeing at least one co-presentation of every unordered pair while bounding per-trial load. Fusion used the per-trial scaled estimator with $\alpha = 0$ (max-normalized distances, no distance-based weighting) and inverse-MDS disabled. Convergence was quantified using Pearson and Spearman correlations of the raw off-diagonal entries (vectorized lower triangle; $N = 58 \Rightarrow 1{,}653$ pairs). Agreement on the raw scale was assessed with the concordance correlation coefficient (CCC; Lin 1989), Deming regression with variance ratio $\lambda = 1$ and 95% confidence intervals (Carstensen, 2010; Linnet, 1993), RMSE, and Bland–Altman bias and 95% limits of agreement (Bland & Altman, 1986; Giavarina, 2015). For range-standardized error and visualization only, we additionally report metrics after separate per-method min-max scaling to $[0, 1]$. Summary metrics are reported in Table 1.

**Interpretation** Subject 1 shows high correlation and substantial concordance; Subject 2 shows moderate correlation and concordance. Deming slopes below 1.0 with small negative intercepts indicate that multi-arrangement yields slightly compressed dissimilarities relative to one-by-one overall, consistent with a mild range contraction in simultaneous layouts. Bland–Altman analysis shows a small positive bias for one-by-one over multi-arrangement with reasonably narrow limits of agreement.

Importantly, we do not expect perfect alignment between the two methods. Multi-arrangement elicits context-infused judgments where each item is evaluated relative to the specific subset presented, whereas sequential ratings may invoke implicit comparison against different internal anchors. Nevertheless, the systematic relationship indicates that both methods converge on the similar underlying similarity structure (see Table 1 and Figure 4).

| | **Correlation and error metrics** | | | |
|---|---|---|---|---|
| **Subject** | Pearson | Spearman | CCC | $\mathrm{RMSE}_{\mathrm{norm}}$ |
| Subject 1 | 0.846 | 0.744 | 0.778 | 0.207 |
| Subject 2 | 0.694 | 0.593 | 0.578 | 0.200 |
| Mean | 0.770 | 0.669 | 0.678 | 0.203 |
| | **Deming regression** | | | |
| **Subject** | Slope | $\mathrm{CI}_{\mathrm{low}}$ | $\mathrm{CI}_{\mathrm{high}}$ | Intercept |
| Subject 1 | 0.934 | 0.914 | 0.956 | −0.073 |
| Subject 2 | 0.872 | 0.827 | 0.911 | −0.024 |
| | **Bland–Altman** | | | |
| **Subject** | Mean diff | $\mathrm{LoA}_{\mathrm{low}}$ | $\mathrm{LoA}_{\mathrm{high}}$ | |
| Subject 1 | 0.123 | −0.204 | 0.450 | |
| Subject 2 | 0.120 | −0.175 | 0.415 | |

Table 1: Agreement between one-by-one and multi-arrangement dissimilarities.

# Usage of the Package

## Installation and setup

Install from PyPI (Python 3.12+) and ensure common multimedia dependencies are available.

```
pip install multiarrangement
# or: uv pip install multiarrangement
```

The package bundles demo media (videos, images, audio), default instruction clips, and a cache of cov-

[2] https://github.com/UYildiz12/Multiarrangement-for-videos

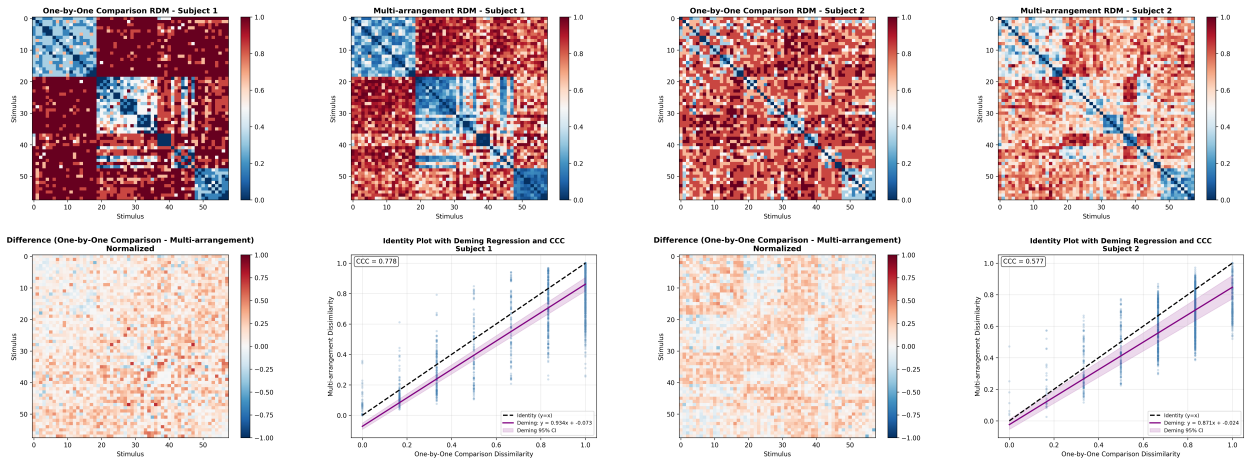


**Figure 4.** RDM agreement for two subjects. Panels correspond to Subject 1 (left) and Subject 2 (right). Within each panel: top row shows the one-by-one and multi-arrangement RDMs. Bottom left shows the normalized differences (one-by-one minus multi-arrangement) while the bottom right shows the identity plot with Deming regression and CCC.

ering designs (LJCR) for offline use.

## Preparing stimuli

Place your files (videos, audio, or images) in a single input folder. The type of modality is detected automatically, and requirements for submissions are adaptively set, and mixed modalities are supported. Filenames serve as token labels for easy tracking of entries.

## Fixed-batch (Set-Cover) Usage

The following example demonstrates how to use the Set-Cover mode in our API. Firstly we precompute a fixed-$k$ schedule of covering batches, seeding for reproducibility and using La Jolla coverings when available, then launch the session with that schedule to collect placements and automatically save the resulting RDM and logs. Batch size $k$ can be tuned and shrink-only refinement can be enabled with `flex=True`. Weight modes and alphas can be tuned as parameters based on preferences and assumptions. The run writes per-trial logs, coverage diagnostics, and the final RDM to the output folder for use in visualization or further downstream analysis.

```
import multiarrangement as ma
input_dir = "./videos" # your stimuli
output_dir = "./results"
batches = ma.create_batches(
    ma.auto_detect_stimuli(input_dir),
    k=8, seed=42, flex=False
)
results = ma.multiarrangement(
    input_dir="./videos", #Where your videos or audios
        are
    batches=batches,
    output_dir="./results", #Where your results will
        appear
    show_first_frames=True,
    fullscreen=False,
    language="en", # Or tr if you'd like Turkish
        instructions
    instructions="default", # or None, or ["Custom", "
        lines"]
    setcover_weight_alpha=2.0,
    setcover_weight_mode='max', # 'max' (d/max) or '
        rms' (-RMSmatched) or 'k2012' for hybrid LtW
        style
    use_inverse_mds=False,
    robust_method=None, # 'winsor' or 'huber'
)
results.vis(title="-SetCover RDM")
results.savefig("results/rdm_setcover.png", title="
    -SetCover RDM")
```

### Adaptive (Lift-the-Weakest) Usage

The following example demonstrates how to run the adaptive Lift-the-Weakest mode within the Multiarrangement library: set subset size bounds and a target per-trial time or evidence threshold, optionally enable inverse MDS, then run to produce the evolving RDM, evidence matrix, and logs. The LtW approach is more dynamic and fine-grained than Set-Cover in how it revisits stimulus pairings, but can take longer and may overwhelm participants when N is large, both in total time and during the initial arrangement phase.

```
import multiarrangement as ma

res = ma.multiarrangement_adaptive(
    input_dir="./videos",
    output_dir="./results",
    participant_id="S01",
    language="en",
    fullscreen=True,
    min_subset_size=4, max_subset_size=6,
    evidence_weight_mode="k2012", # "rms" or "max"
    evidence_alpha=2.0,
    stop_on_utility=False, # use raw-evidence
         threshold
    evidence_threshold=0.35,
    instructions="default",
)
res.vis(title="Adaptive LTW RDM")
res.savefig("./results/rdm_adaptive.png", title="
     Adaptive LTW RDM")
```

### Custom instructions and localization

Default instructions are provided in English and Turkish for convenience. To override, pass a list of strings to be shown as paginated screens before the task, per the following example:

```
custom = [
  "Welcome to the study.",
  "Drag each item into the white circle.",
  "Double-click a token to play/replay the video.",
  "Press SPACE to continue."
]
ma.multiarrangement(input_dir, batches, "./results",
     instructions=custom)
# ...or for adaptive:
ma.multiarrangement_adaptive("./videos", "./results",
     instructions=custom)
```

For further detailed documentation, see our GitHub page[2]. We release all code, experiment templates, and analysis scripts used in this paper.

## Conclusions

In this paper, we introduce Multiarrangement, a turn-key, offline, open-source software package for collecting human similarity judgments with dynamic visual stimuli. The system unifies a simple, robust interface with a circular arena, on-demand playback, compliance cues, optional zoom, and bilingual defaults with scheduling back-ends that make large stimulus sets practical. A fixed Set-Cover mode produces deterministic, between-subject-comparable schedules with bounded per-trial load, while an adaptive Lift-the-Weakest mode targets the least-supported region of the space under explicit time and subset-size constraints. Both schedulers offer choices for distance fusion, stopping rules, and robustness & refinement, allowing researchers to align the pipeline with their design goals and assumptions. We also report a small within-subject validation indicating that multiarrangement approximates the structure captured by one-by-one comparisons while using far fewer trials, enabling context-aware judgments at lower participant burden.

Beyond video, the same workflow supports images and audio with no changes to the analysis path, and results are written in standard formats to facilitate downstream use in behavioral and neuroimaging pipelines. The package is designed to be easily extended. Researchers can customize instructional text, enforce additional compliance rules, adjust scheduler parameters, and integrate new quality-control or weighting schemes while preserving reproducibility via recorded seeds and schedules. We hope this plug-and-play framework lowers the barrier to deploying multi-arrangement studies at scale and stimulates new experimental designs that leverage simultaneous comparison to probe perceptual and cognitive structure in richer, more naturalistic settings.

### Declarations

**Authors' contributions** Both authors jointly conceptualized the study. U.Y. implemented the software, con-

ducted testing & validation, and collected the data. Both authors co-wrote the first draft, critically revised subsequent drafts, and approved the final manuscript for submission.

**Funding** This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

**Competing interests** The authors declare no competing interests.

**Ethics approval** The study was approved by, and conducted along the guidelines of, the Bilkent University Ethics Committee.

**Consent to participate** All participants provided written informed consent prior to participation. Data were analyzed anonymously.

**Consent for publication** All participants provided written informed consent for publication of anonymized data.

**Open Practices Statement** Materials (stimulus lists and templates), validation data, and analysis code are available at 10.5281/zenodo.17463843. Further details on package use are available in our GitHub repository at github.com/UYildiz12/Multiarrangement-for-videos. None of the reported studies were preregistered.

# References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Baveye, Y., Dellandréa, E., Chamaret, C., & Chen, L. (2015). LIRIS-ACCEDE: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, *6*(1), 43–55. https://doi.org/10.1109/TAFFC.2015.2396531

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Carstensen, B. (2010). *Comparing clinical measurement methods: A practical guide*. Wiley. https://doi.org/10.1002/9780470683019

Finn, E. S., Glerean, E., Khojandi, A. Y., Nielson, D., Molfese, P. J., Handwerker, D. A., & Bandettini, P. A. (2020). Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, *215*, 116828. https://doi.org/10.1016/j.neuroimage.2020.116828

Giavarina, D. (2015). Understanding bland–altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Gilman, T. L., Shaheen, R., Nylocks, K. M., Halachoff, D., Chapman, J., Flynn, J. J., Matt, L. M., & Coifman, K. G. (2017). A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation. *Behavior Research Methods*, *49*(6), 2061–2082. https://doi.org/10.3758/s13428-016-0842-x

Gordon, D. M. (2025, October 24). *La jolla covering repository tables* [Last updated 2025-10-24]. Retrieved October 28, 2025, from https://ljcr.dmgordon.org/cover/table.html

Gordon, D. M., Kuperberg, G., & Patashnik, O. (1995). New constructions for covering designs. *Journal of Combinatorial Designs*, *3*(4), 269–284. https://doi.org/10.1002/jcd.3180030404

Gross, J. J., & Levenson, R. W. (1995). Emotion elicitation using films. *Cognition & Emotion*, *9*(1), 87–108. https://doi.org/10.1080/02699939508408966

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2014). A high-resolution 7-tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*, 140003. https://doi.org/10.1038/sdata.2014.3

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience

of film. *Projections*, *2*(1), 1–26. https://doi.org/10.3167/proj.2008.020102

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Jääskeläinen, I. P., Sams, M., Glerean, E., & Ahveninen, J. (2021). Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage*, *224*, 117445. https://doi.org/10.1016/j.neuroimage.2020.117445

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., & Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*, 245. https://doi.org/10.3389/fpsyg.2012.00245

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*(1), 255–268. https://doi.org/10.2307/2532051

Linnet, K. (1993). Evaluation of regression procedures for method comparison studies. *Clinical Chemistry*, *39*(3), 424–432. https://doi.org/10.1093/clinchem/39.3.424

Liu, X., Dai, Y., Xie, H., & Zhen, Z. (2022). A studyforrest extension, MEG recordings while watching the audio-visual movie Forrest Gump. *Scientific Data*, *9*, 206. https://doi.org/10.1038/s41597-022-01299-1

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLOS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition & Emotion*, *24*(7), 1153–1172. https://doi.org/10.1080/02699930903274322

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Urgen, B. A., Nizamoğlu, H., Eroğlu, A., & Orban, G. A. (2023). A large video set of natural human actions for visual and cognitive neuroscience studies and its validation with fMRI. *Brain Sciences*, *13*(1), 61. https://doi.org/10.3390/brainsci13010061