applied sciences

MDPI

Article

Psychophysical Validation of Gestalt-Inspired Encoding Strategies for Dynamic Action Recognition in Simulated Prosthetic Vision

Yuxi Wang, Ying Zhao *, Ziyang Fan, Lin Gan and Yuerong Han

School of Digital and Intelligence Industry, Inner Mongolia University of Science and Technology, Baotou 014010, China; wangyuxi0715@163.com (Y.W.); 2024023157@imust.edu.cn (Z.F.); ganlin.g@foxmail.com (L.G.); 17837071857@163.com (Y.H.)
* Correspondence: amengs@imust.edu.cn

Abstract

Visual prostheses provide restricted visual information through phosphene-like percepts, making the organization of spatial and temporal cues important for dynamic action recognition. This study investigated human action recognition under four simulated prosthetic vision (SPV) encoding conditions—NO_EDGE, BASELINE, PCL, and PCO—selected from 11 computationally screened candidates. Twenty normally sighted participants completed open-response action-recognition tasks with nominal 32 × 32 SPV videos. Experiment 1 compared strategies under full, half, and quarter temporal-availability conditions. Experiment 2, by contrast, examined two complementary manipulations—frame-holding and background-suppression—prior to SPV generation. Responses were classified as Correct, Vague, Uncertain, or Wrong, and Correct identification was analyzed using binomial generalized linear mixed models. In Experiment 1, strategy, temporal availability, and their interaction were associated with correct recognition. PCO had the highest overall observed Correct rate (58.6%) and the lowest Uncertain rate (20.8%); BASELINE approached PCO in the Quarter condition, while PCL showed lower overall recognition. Reduced temporal availability decreased correct identification and increased uncertainty. In Experiment 2, frame holding generally reduced recognition, whereas background suppression did not produce a consistent improvement. Model-defined pathway characteristics helped select contrasting strategies, and the behavioral results revealed their distinct, condition-dependent recognition profiles. The findings provide a basis for task-oriented SPV encoding that considers target organization and temporal presentation together.

Keywords: simulated prosthetic vision; phosphene vision; dynamic action recognition; Gestalt principles; encoding strategy; SlowFast model; psychophysical experiment; temporal information

Check for updates

Academic Editor: Thomas Lindner

Received: 20 September 2026
Revised: 30 October 2026
Accepted: 30 October 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Visual prostheses seek to provide partial visual function by stimulating the retina, visual cortex, or other structures along the visual pathway to evoke phosphene-like percepts [1,2]. Implanted systems have supported selected functional tasks, including letter recognition, but the information conveyed remains constrained by stimulation density, spatial resolution, and phosphene characteristics [3,4]. Consequently, the organization of task-relevant information is important when visual input is represented by a limited number of perceptual elements.

Simulated prosthetic vision (SPV) provides a controlled approach to evaluating encoding methods with normally sighted observers [5,6]. Previous work has examined saliency-based processing and visual-information optimization [7,8], structural and semantic segmentation [9], and scene simplification [10]. More recently, task-oriented viewing-point representations have been investigated for object recognition [11]. Dynamic information has also been addressed through foreground-oriented processing for moving-object recognition [12], depth and motion cues [13], and temporally sequenced semantic information for wayfinding [14]. These studies have evaluated different tasks and encoding operations, motivating a common behavioral comparison of representations designed for dynamic action recognition.

Action recognition requires information about body posture, target contours, motion, and changes across frames. Gestalt principles such as proximity, continuity, closure, and figure-ground organization provide a basis for arranging sparse visual elements into recognizable structures [15,16]. Recognition studies using sparse phosphenes and curved segments also indicate the relevance of contour organization and element geometry [17]. In dynamic SPV, however, changes that improve target structure may also alter temporal presentation. The behavioral characteristics of different encoding implementations therefore need to be examined under the same recognition task.

Computational action-recognition research has used multi-pathway video representations, including two-stream networks and SlowFast [18,19]. Eleven candidate SPV encoding conditions were considered for computational screening and subsequent psychophysical stimulus selection. In the present work, computational pathway-input ablation was used to characterize contrasts among candidate SPV encodings and guide the selection of behavioral stimulus conditions. This screening was not intended to establish a correspondence between SlowFast pathways and human spatial or temporal processing. The primary evidence was instead obtained from open-response psychophysical experiments.

The human study used nominal 32 × 32 SPV action videos derived from 24 UCF101 categories [20]. Twenty normally sighted participants identified actions in an open-response task. Experiment 1 compared NO_EDGE, BASELINE, PCL, and PCO under Full, Half, and Quarter temporal-availability conditions. Experiment 2 examined frame holding and background suppression applied before SPV generation, using a separate video set. Responses were categorized as Correct, Vague, Uncertain, or Wrong, with Correct identification as the principal inferential outcome.

The study addressed three behavioral questions: (1) whether the four selected complete encoding implementations differed in human action identification; (2) whether their recognition profiles varied when available frames and presentation duration were jointly reduced; and (3) whether preprocessing before SPV generation changed recognition performance.

The first two questions were examined in Experiment 1 (Exp1) and the third in Experiment 2 (Exp2). Model-defined pathway characteristics served as descriptors for candidate selection and were examined exploratorily in relation to human responses. They were not treated as direct proxies for human perceptual mechanisms. Accordingly, the work evaluates complete displayed representations rather than isolating the causal contributions of individual Gestalt principles. The findings should also be interpreted within the limitations of simulated prosthetic vision, as real implanted vision involves additional stimulation, perceptual, and viewing-related factors [21–23]. Section 2 describes the methods, Section 3 presents the results, Section 4 discusses the findings, and Section 5 provides the conclusions.

2. Materials and Methods

2.1. Study Overview and Psychophysical Design

Figure 1 outlines the design of this study. Two psychophysical experiments provided the primary evaluation of human action recognition under four selected SPV encodings. Experiment 1 compared these encodings across three temporal-availability conditions, whereas Experiment 2 assessed preprocessing before SPV generation using a different set of source videos.

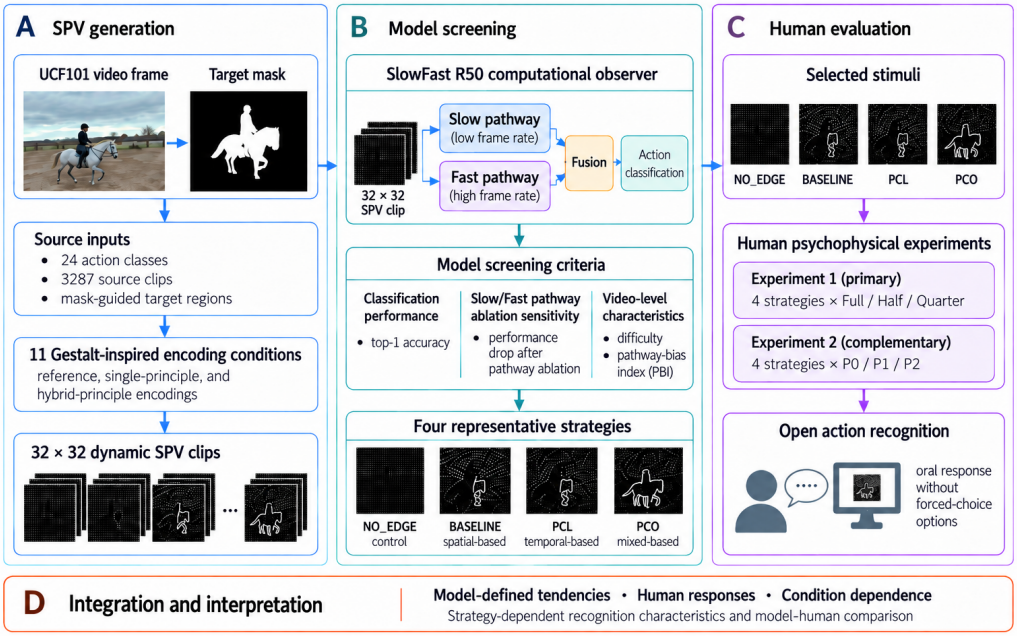


Figure 1. Overview of the psychophysical study and the auxiliary model-guided selection process. (**A**) SPV stimulus generation from UCF101 videos, target masks, and 11 candidate encoding conditions. (**B**) SlowFast-based computational screening and selection of four representative encoding strategies. (**C**) Human evaluation through two open-response psychophysical experiments. (**D**) Integration of model-defined characteristics and condition-dependent behavioral findings.

Twenty normally sighted participants completed open-response action-recognition trials in both experiments. The design and formal trial counts are summarized in Table 1. The encoding candidates and computational screening used to select the four conditions are described in Section 2.3.

Table 1. Design and formal trial counts of the psychophysical experiments.

Experiment	Base Videos	Experimental Factors	Trials/Person	Total Trials
Exp1	36	Four encoding strategies × three temporal-availability conditions	72	1440
Exp2	18	Four encoding strategies × three preprocessing conditions (P0/P1/P2)	36	720

Note. Both experiments included 20 participants and used distinct base-video sets. Strategy-by-condition trial counts were balanced within participants, whereas individual video-by-strategy-by-condition combinations were not fully crossed.

2.2. Psychophysical Experiments

2.2.1. Participants and Apparatus

Twenty normally sighted participants (10 men and 10 women, aged 20–30 years) were recruited from Inner Mongolia University of Science and Technology. All had normal or

corrected-to-normal visual acuity, were native Mandarin Chinese speakers, and had no prior experience with the SPV task. The experiments were conducted under the ethics approval covering the larger research project and in accordance with the Declaration of Helsinki [24]. Written informed consent was obtained from all participants before testing.

Stimuli were presented using PsychoPy [25] at the center of a Philips monitor (model 275V8L/93; MMD China, Shanghai, China) configured to 1920 × 1080 pixels at 60 Hz. The stimuli were viewed within approximately the central 20° of the visual field, at a viewing distance of approximately 40 cm. The experiments were conducted in a dimly lit room with minimal external interference.

2.2.2. Trial Procedure and Response Coding

Participants completed eight practice trials using videos that were not included in the formal stimulus sets. The practice phase familiarized participants with SPV presentation, the action-recognition task, and the key-press procedure. Participants were informed that the stimuli could involve four broad activity groups: sports, musical-instrument performance, everyday interactions, and activities with complex backgrounds. However, the 24 specific action-category labels were not disclosed, and no response options were provided.

Each formal trial began with a centrally presented SPV video. Participants who formed an action judgment during playback could press the "0" key to end the video and proceed to the oral-response screen; otherwise, that screen appeared when the video ended.

Participants verbally named the perceived action or reported that they could not identify it. A space-bar press confirmed the response; otherwise, the oral-response screen advanced after 20 s.

Responses were coded into four mutually exclusive categories using Table 2: Correct (3), Vague (2), Uncertain (1), and Wrong (0). Correct included an acceptable action synonym; Vague indicated partial or broader action information; Uncertain indicated inability to judge; and Wrong identified an incorrect specific action. One researcher classified participants' verbal responses according to the criteria in Table 2, using the ground-truth action label to assess whether a response correctly identified the target action. During response coding, the researcher was blinded to the encoding-strategy and experimental-condition identifiers associated with the response. Category assignments were checked against the original experimental records before analysis. No independent second rating or formal inter-rater agreement estimate was obtained.

Table 2. Rating criteria for open oral responses.

Rating Value	Category	Criterion	Example Interpretation
3	Correct	Participant clearly identified the target action or an acceptable synonym.	"riding a horse" for HorseRiding
2	Vague	Participant reported partial or superordinate information without naming a wrong specific action.	"someone is playing an instrument"
1	Uncertain	Participant explicitly reported inability to judge.	"uncertain" or "cannot tell"
0	Wrong	Participant named a specific action inconsistent with the target action.	"Tai Chi" for BandMarching

Note. The rating criteria were applied to open oral descriptions rather than forced-choice responses. The rating values were used for trial-level response coding and condition-level summary statistics.

2.2.3. Experiment 1: Temporal Availability and Trial Allocation

Experiment 1 compared NO_EDGE, BASELINE, PCL, and PCO under Full, Half, and Quarter temporal-availability conditions. Full retained the complete encoded SPV sequence, whereas Half and Quarter were generated by retaining every second and every fourth frame, respectively. Temporal subsampling was applied after SPV encoding; therefore, it

also interrupted the original odd–even complementary frame pairings in PCL. The reduced-availability stimuli contained fewer frames and had shorter presentation durations.

Each participant completed 72 formal trials across three blocks, with six trials in each of the 12 strategy-by-temporal-availability combinations (Table 1). Each of the 36 base videos was presented twice per participant under different strategy–condition combinations, with at least five trial positions between repeated presentations. Source-video presentation frequencies were balanced across participants, whereas individual video-by-strategy-by-condition combinations were not equally replicated. Participant-specific trial lists also maintained balanced distributions of model-defined difficulty and pathway-bias categories across the assigned stimuli.

2.2.4. Experiment 2: Preprocessing and Trial Allocation

Experiment 2 compared P0 (no preprocessing), P1 (frame holding), and P2 (background suppression). Each operation was applied to the source video and corresponding mask before strategy-specific SPV generation. Table 3 summarizes the parameter settings.

P0 passed the source frames and masks to the SPV pipeline without additional preprocessing. P1 repeated the content of each selected source frame over k consecutive output frames, preserving the total frame count and nominal video duration while reducing the frequency of visual updates. To maintain alignment between image content and target location, the corresponding mask was held over the same interval. For output frame t, the preprocessed image and mask were defined as

$$F_t^{\mathrm{P1}} = F_{k\left\lfloor \frac{t-1}{k} \right\rfloor + 1}, \tag{1}$$

$$M_t^{\mathrm{P1}} = M_{k\left\lfloor \frac{t-1}{k} \right\rfloor + 1}, \tag{2}$$

where F_t and M_t denote the original frame and its target mask, respectively, and $\lfloor \cdot \rfloor$ is the floor operator. The formal experiment used $k = 6$.

P2 reduced background luminance while retaining the original target-region intensity. The binary target mask $B_t(x, y)$ was defined as

$$B_t(x,y) = \begin{cases} 1, & (x,y) \text{ belongs to the target region,} \\ 0, & (x,y) \text{ belongs to the background region.} \end{cases} \tag{3}$$

The processed frame was obtained as

$$F_t^{\mathrm{P2}}(x,y) = B_t(x,y)F_t(x,y) + [1 - B_t(x,y)]\beta F_t(x,y) \tag{4}$$

where β is the background-luminance retention factor. The formal Experiment 2 used $\beta = 0.5$. P2 did not alter frame order or nominal presentation duration. Unlike P1, it changed the target–background luminance relationship rather than the temporal update pattern.

Preliminary screening considered frame-hold intervals of $k = 4, 6, 8$ and background-retention factors of $\beta = 0.3, 0.4, 0.5$. Stimuli generated using these candidate settings were rated by several participants, and the settings receiving higher overall ratings were selected for the formal experiment ($k = 6$ and $\beta = 0.5$). This was an exploratory parameter-selection procedure rather than a formal comparison of candidate settings. The preliminary ratings were not included in the formal behavioral analyses.

Experiment 2 followed Experiment 1 after a break and additional instructions. Each participant completed 36 formal trials using 18 base videos distinct from those in Experiment 1. The 12 strategy-by-preprocessing combinations each contributed three trials per participant (Table 1). Each base video was presented twice per participant under different combinations, with at least seven trial positions between repeated presentations. The response procedure

was unchanged. The same stimulus-balancing principles were applied to the Experiment 2 trial lists while preserving the three trials per strategy-by-preprocessing combination.

Table 3. Candidate and selected preprocessing parameters for Experiment 2.

Condition	Operation	Candidate Values	Selected Value
P0	No preprocessing	–	–
P1	Frame holding (interval k)	$k = 4, 6, 8$	$k = 6$
P2	Background suppression (retention β)	$\beta = 0.3, 0.4, 0.5$	$\beta = 0.5$

Note. Candidate values were considered during preliminary parameter screening. The selected values were used in the formal Experiment 2. Preliminary screening data were not included in the formal behavioral analysis.

2.3. Stimulus Preparation and Model-Guided Selection

The stimulus-generation procedure included preparation of source videos, acquisition of target masks, and strategy-dependent phosphene rendering. The encoding operations are described in Section 2.3.4.

2.3.1. Source Videos, Target Masks, and SPV Generation

Action videos were obtained from UCF-101 [20]. A subset of 3287 videos from 24 action categories was prepared for computational screening and SPV generation. The smaller video sets used in the human experiments were selected subsequently, as described in Section 2.3.5. The 24 categories covered sports (15 classes), musical-instrument performance (3 classes), everyday interactions (3 classes), and activities with complex backgrounds (3 classes).

For mask-dependent strategies, target-region masks were estimated using the pre-trained InSPyReNet salient-object segmentation model [26]. The masks delineated the target regions used for boundary extraction, phosphene allocation, and target–background luminance modulation. For each source video, the same target-region definition was applied across encoding conditions.

Source frames were spatially aligned with their target masks, converted to grayscale using OpenCV, and spatially resized as required by the corresponding encoding.

The selected rendering paths did not include an additional explicit per-frame or per-video min–max intensity-normalization step. The nominal SPV resolution was 32 × 32. All four encoding conditions used 1024 phosphene elements per frame, although their spatial locations and assigned intensities could differ across strategies. The element count included zero-intensity elements and therefore did not imply an equal number of visibly illuminated phosphenes across strategies. NO_EDGE used a regular low-resolution array, whereas strategies involving proximity-based sampling redistributed phosphene positions toward target regions. Depending on the encoding, target contours, phosphene shapes, or frame-wise visibility were modified. The encoded frames retained the source-video temporal order.

The encoded SPV clips were used to prepare the psychophysical stimuli.

2.3.2. Candidate Encoding Conditions

Eleven candidate encodings were evaluated: two reference conditions, four single-principle strategies, and five hybrid strategies. These operations modified the SPV representation rather than the network architecture. Table 4 summarizes the candidate strategies and identifies the four conditions retained for human evaluation.

The conceptual design and computational evaluation of the 11 Gestalt-inspired SPV encoding conditions were reported in our earlier study [27]. The present work focuses on

the human psychophysical evaluation of four selected complete implementations; SlowFast-based screening was used here as an auxiliary selection step, not as a substitute for human behavioral evidence.

Table 4. Candidate SPV encoding strategies and their roles in the human experiments.

Group	Strategy	Main Operation	Role
Reference	NO_EDGE	Grayscale SPV rendering without edge enhancement	Control
Reference	BASELINE	Target-contour enhancement	Spatial-biased
Single	PROXIMITY	Target-aware non-uniform sampling	–
Single	CONTINUITY	Directional contour rendering	–
Single	CLOSURE	Temporally complementary target display	–
Single	FIGURE_GROUND	Target–background luminance modulation	–
Hybrid	PROX-CONT (PCO)	Proximity + continuity rendering	Mixed
Hybrid	PROX-CLOS (PCL)	Proximity + complementary display	Temporal-biased
Hybrid	PROX-FG	Proximity + figure–ground modulation	–
Hybrid	CLOS-CONT	Closure + continuity rendering	–
Hybrid	PROX-CLOS-CONT	Proximity + closure + continuity	–

Note. The role column refers to the four conditions retained for human testing; a dash indicates that the candidate was screened computationally but not selected. Bias labels describe model-defined tendencies, not human visual pathways.

2.3.3. Model-Guided Selection of Four Conditions

SlowFast R50 was used as a computational observer [19]. Its Slow and Fast pathways sampled the input video at different temporal rates and exchanged features through lateral connections (Figure 2). The network was adapted to 24 classes, and its input configuration is summarized in Table 5.

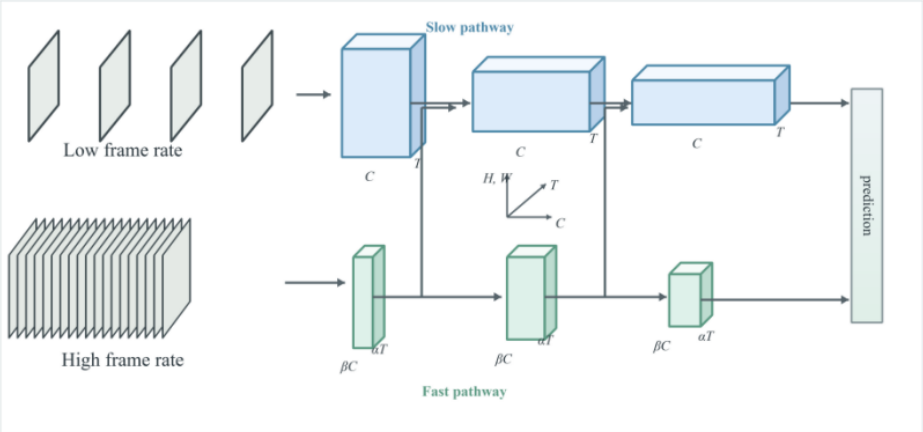


Figure 2. Schematic structure of the SlowFast network used as the computational observer. Blue and green blocks represent the Slow and Fast pathways, respectively.

Candidate screening compared clip-level Top-1 accuracy under Full, Slow-only, and Fast-only inference. The single-pathway conditions used the same trained checkpoint, with the inactive pathway input set to zero and no additional parameter updates. Because full-model accuracies were relatively similar across the candidate encodings, selection emphasized contrasting pathway-input sensitivity profiles rather than full-model accuracy alone.

Table 5. SlowFast configuration and pathway-related metrics.

Model Configuration			
Backbone	SlowFast R50; 24 classes	Weights	Pretrained; PyTorchVideo
Dataset	UCF101; 24 classes; 3287 videos	SPV grid	11 strategies; 32 × 32
Model input	32 frames; 224 × 224; pseudo-RGB	Pathways	$\alpha = 4$; Slow 8; Fast 32
Split	80/20 within each class, by video order; no group isolation	Batch size	4 (some runs: 2)
Optimizer	SGD; momentum 0.9	Learning rate	10^{-5}; constant
Weight decay	10^{-4}	Epochs	20
Evaluation	Clip-level Top-1 accuracy	Input ablation	Full, Slow-only, Fast-only
Pathway-Related Metrics			
Spatial sensitivity	$S_{\text{spatial},s} = A_{\text{Full},s} - A_{\text{Fast},s}$		
Temporal sensitivity	$S_{\text{temporal},s} = A_{\text{Full},s} - A_{\text{Slow},s}$		
Normalized bias	$B_s = \frac{S_{\text{temporal},s} - S_{\text{spatial},s}}{S_{\text{temporal},s} + S_{\text{spatial},s}}$		
Synergy (descriptive)	$G_s = \frac{S_{\text{spatial},s} + S_{\text{temporal},s}}{2}$		

Note. $A_{\text{Full},s}$, $A_{\text{Slow},s}$ and $A_{\text{Fast},s}$ are clip-level Top-1 accuracies. In inactive-pathway tests, the other pathway input was set to zero without retraining. Sensitivities are in percentage points; B_s is dimensionless; and G_s is a descriptive average, not an interaction effect.

Figure 3 summarizes the pathway-ablation profiles of the 11 candidate encodings. BASELINE showed a spatial-biased profile (79.3 and 12.4 percentage points for spatial and temporal sensitivity, respectively). PCL showed a temporal-biased profile (4.6 and 87.7 percentage points), whereas PCO showed a mixed profile (40.1 and 41.8 percentage points).

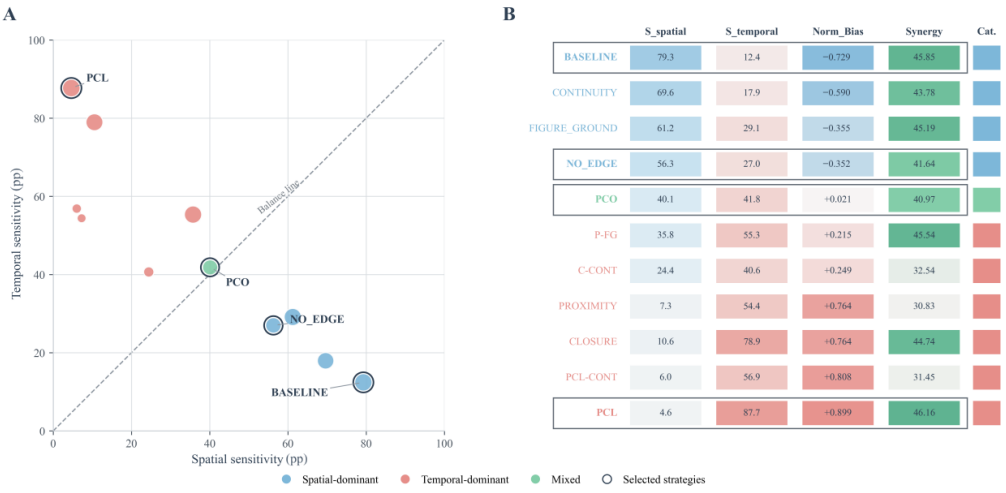


Figure 3. Model-based spatiotemporal bias analysis and representative strategy selection. (**A**) Distribution of candidate encoding strategies in the spatial–temporal sensitivity space. (**B**) Summary of spatial score, temporal score, normalized bias, synergy, and model-defined category for each strategy.

NO_EDGE was retained as the conventional SPV control. Together, the four selected conditions provided contrasting complete encoding implementations for the psychophysical experiments. The computational profiles guided stimulus selection rather than serving as direct measures of human spatial or temporal processing.

2.3.4. Implementation of the Four Selected Encodings

The four selected strategies were implemented for a nominal 32 × 32 SPV representation.

Let M_t denote the binary target mask for frame t, $I_t^{\mathrm{NO}}(p)$ the grayscale-derived digital intensity evaluated at phosphene position p, and $I_{\max} = 255$ the maximum 8-bit digital intensity.

In the completed experimental stimuli, each encoded frame comprised 1024 assigned phosphene positions, including positions with zero or low digital intensity; this count does not imply that all 1024 elements were visibly illuminated.

NO_EDGE used a fixed 32 × 32 sampling lattice without additional target-boundary enhancement and served as the control condition.

For BASELINE, the inner target boundary was represented by the difference between the target mask and its erosion with a 3 × 3 square structuring element:

$$E_t^{\mathrm{BASE}} = M_t - \mathrm{erode}_{3\times3}(M_t). \tag{5}$$

The equivalent implementation used a 3 × 3 binary-neighborhood test: a target-grid location was designated as a boundary if at least one location within its neighborhood was outside the processed target mask. This equivalence presumes the same binary mask and boundary-handling convention.

The boundary elements were enhanced while the remaining intensities were unchanged:

$$I_t^{\mathrm{BASE}}(p) = \begin{cases} I_{\max}, & p \in E_t^{\mathrm{BASE}}, \\ I_t^{\mathrm{NO}}(p), & p \notin E_t^{\mathrm{BASE}}. \end{cases} \tag{6}$$

The operation increased the digital intensity assigned to target-boundary elements without introducing frame-wise complementary presentation. BASELINE retained fixed-grid phosphene positions; boundary enhancement changed their assigned intensities rather than their spatial coordinates.

PCL combined proximity-oriented phosphene allocation with temporally complementary presentation of target-region information. After preferential sampling of the target region, a binary visibility field was used to control the presentation of target elements.

Let $t = 1, 2, \ldots, T$ denote the frame index and let $q = \lceil t/2 \rceil$ identify the corresponding pair of successive frames. For each pair, $R_q(p)$ was generated by bilinearly upsampling an 8 × 8 uniform random field and thresholding it at 0.5.

The visibility mask applied to frame t was

$$C_t(p) = \begin{cases} 1 - R_q(p), & t \text{ is odd}, \\ R_q(p), & t \text{ is even}. \end{cases} \tag{7}$$

For the inspected PCL implementation, the random field for frame pair q was generated using `default_rng` with seed $42 + q - 1$. The frame index restarted for each source video.

The two visibility masks were complementary in the common image coordinate system, although the nonuniform phosphene positions were resampled separately for each frame.

Let E_t^{PCL} denote the target-side edge region, defined as the intersection of the binary target mask and the edge region obtained using a 5 × 5 morphological-gradient operation. The PCL intensities were assigned as

$$I_t^{\mathrm{PCL}}(p) = \begin{cases} 0, & p \in M_t,\ C_t(p) = 0, \\ I_{\max}, & p \in E_t^{\mathrm{PCL}},\ C_t(p) = 1, \\ I_t^{\mathrm{NO}}(p), & p \in M_t \setminus E_t^{\mathrm{PCL}},\ C_t(p) = 1, \\ I_t^{\mathrm{NO}}(p), & p \notin M_t. \end{cases} \tag{8}$$

Thus, enabled boundary elements were highlighted, enabled nonboundary target elements retained their grayscale-derived values, and masked-out target elements remained unilluminated.

The complementary visibility masks changed in successive frame pairs, while PCL sampling positions were regenerated for each frame.

PCO combined proximity-based non-uniform sampling with continuity-oriented phosphene rendering. A variable-radius Poisson-disk-like sampler was applied in the 512×512 processing space. The target mask was used to distinguish edge, target-interior, and background regions, with the edge region identified by Canny contour extraction.

Canny edge detection with low and high thresholds of 100 and 200, respectively, was applied to the binary target mask.

The resulting edge candidates were dilated using a 5×5 square kernel. The region-dependent minimum sampling distance was defined as

$$r_{\min}(p) = \begin{cases} 10, & p \in \Omega_{\text{edge}}, \\ 12, & p \in \Omega_{\text{int}}, \\ 16, & p \in \Omega_{\text{bg}}, \end{cases} \tag{9}$$

where the distances are expressed in pixels. For the nominal 32×32 SPV representation, the target sampling budget was $N = 1024$. Edge samples were allocated according to contour length, while the remaining sampling budget was distributed between the target interior and background according to their areas, with an interior weighting factor of 2.0.

For contour-oriented rendering, the target mask was smoothed with a 5×5 Gaussian kernel using OpenCV's `sigmaX=0` setting.

Scharr derivatives were then used to obtain the local contour-tangent direction:

$$\theta_t(p) = \operatorname{atan2}\left(g_{y,t}(p), g_{x,t}(p)\right) + \frac{\pi}{2}. \tag{10}$$

Edge phosphenes were rendered as filled, antialiased ellipses with their major axes aligned to the local contour tangent rather than as Gaussian intensity-distribution ellipses.

With a grid-cell reference size of 16 px, the ellipse scale factor 0.8 and nominal major-to-minor aspect ratio 3.0 yielded integer drawing semiaxes of 6 and 2 pixels, respectively.

Non-edge elements were rendered as filled, antialiased circles with a radius of 5 pixels.

PCO sampling positions were regenerated for each frame, without an explicit fixed seed for the randomized nonuniform coordinate sampler in the inspected implementation.

As spatial positions were regenerated on successive frames for PCL and PCO, additional apparent positional variation or flicker was possible; this component was not experimentally isolated from the complete encoding implementation.

The encoding was performed offline, and the resulting saved videos were played back without online stimulus regeneration.

PCO therefore differed from BASELINE in both phosphene allocation and contour-element geometry. These descriptions characterize the complete displayed encodings rather than isolating the causal contributions of individual Gestalt principles.

Additional stimulus-rendering and temporal-update details for the four selected SPV encoding conditions are summarized in Table 6.

Table 6. Additional stimulus-rendering and temporal-update details for the four selected SPV encoding conditions.

Implementation Detail	Setting and Scope
Grayscale and intensity	OpenCV BGR-to-grayscale conversion; digital output intensities in the 8-bit range 0–255. No explicit additional per-frame or per-video min–max, gamma, or percentile intensity normalization was identified in the selected rendering paths.
Circular phosphene rendering	NO_EDGE, BASELINE, and PCL used filled, antialiased circular phosphene elements with a radius of 5 px on a 512 × 512 rendering canvas. Non-edge elements in PCO used the same circular profile, whereas contour elements were rendered as filled ellipses with semiaxes of 6 and 2 px. The circular elements were not generated using an individual Gaussian point-spread function.
Boundary and contour operations	BASELINE: 3 × 3 binary-neighborhood inner-boundary criterion, equivalent to square-kernel erosion subtraction under matched boundary handling. PCL/PCO: Canny thresholds 100/200 for sampling candidates; 5 × 5 dilation of candidate edges. PCO: 5 × 5 Gaussian smoothing (`sigmaX=0`) and Scharr derivatives for tangent estimation; filled ellipses with semiaxes 6/2 px.
Randomness and visibility	PCL: complementary fields thresholded at 0.5; one field pair per two successive frames; field-generator seed $42 + q - 1$ for pair q. No explicit fixed seed was identified for PCL/PCO randomized nonuniform position sampling in the selected implementations.
Phosphene positions across frames	NO_EDGE and BASELINE: fixed grid positions. PCL and PCO: nonuniform sampling positions regenerated at each frame. Encoding and rendering were performed offline; participants viewed saved video stimuli, without online resampling at playback.

2.3.5. Selection and Balancing of Behavioral Stimuli

Following computational screening, video-level descriptors were calculated under NO_EDGE as a common reference for psychophysical stimulus selection. Video difficulty was defined using the SlowFast recognition accuracy obtained from NO_EDGE-encoded videos. Video-level recognition accuracy was calculated as the proportion of correctly classified clips among all evaluated clips from each NO_EDGE-encoded source video. This accuracy was used to categorize videos as Easy (>0.85), Medium ($0.60 < S_v \leq 0.85$), or Hard (≤0.60). Video-level pathway-bias index (PBI) descriptors were derived from the same Full, Slow-only, and Fast-only inference comparisons used for the strategy-level profiles, applied to individual source videos under NO_EDGE. These video-level labels were distinct from the strategy-level profiles in Figure 3.

Stimulus selection considered action-category diversity, visible action content, target-mask suitability, model-defined difficulty, and pathway-bias categories. Different activity types were included to reduce repeated exposure to similar action content. The resulting sets comprised 36 base videos for Experiment 1 and 18 distinct base videos for Experiment 2. During selection of videos for the human experiments, clips from the same original UCF101 recording group were excluded to reduce overlap in source content. This criterion concerned behavioral stimulus selection and was separate from the model training–validation partition. Some source videos used in the psychophysical experiments had also been included in computational model training. Accordingly, the exploratory model–human comparisons were not treated as an independent held-out validation.

Participant-specific trial lists were constructed by jointly considering activity groups, model-defined difficulty, and PBI categories to maintain balanced stimulus composition across experimental conditions. Strategy-by-condition trial counts were kept equal, although individual video-by-strategy-by-condition combinations were not fully crossed. Different processed versions of the same source video were not presented consecutively, with repetition spacing specified in Sections 2.2.3 and 2.2.4. Individual video-by-strategy-by-condition combinations were partially sampled rather than fully crossed. These stimulus

sets were intended for behavioral evaluation, not as an independent confirmatory test of model-to-human prediction.

2.4. Data Analysis

Correct identification was the main outcome for statistical inference. The four original response categories were retained for descriptive summaries, with Uncertain responses used to characterize explicit inability to identify an action. Vague and Wrong rates and the coded mean rating were auxiliary measures. Response and timing calculations are summarized in Table 7.

Table 7. Behavioral measures and statistical analyses used in the psychophysical experiments.

Measure or Analysis	Definition or Calculation	Role
Response rates	$R_{i,c}^{(r)} = n_{i,c}^{(r)}/N_{i,c}$	Descriptive
Correct outcome	$Y_t = \mathbb{1}(r_t = \text{Correct})$	Primary outcome
Binomial GLMM	$\text{logit}(p_t)$; fixed effects and participant/video intercepts	Primary inference
Mean rating	$M_{i,c} = (3n_{i,c}^{\text{Correct}} + 2n_{i,c}^{\text{Vague}} + n_{i,c}^{\text{Uncertain}})/N_{i,c}$	Auxiliary summary
Full-to-quarter drop	$D_s = R_{s,\text{Full}}^{(\text{Correct})} - R_{s,\text{Quarter}}^{(\text{Correct})}$	Exp1 description
Timing measures	$E_{i,c} = n_{i,c}^{\text{early}}/N_{i,c}$; $T_{\text{oral}} = t_{\text{confirm}} - t_{\text{onset}}$	Auxiliary
Group summary	$\bar{x}_c = N_p^{-1}\sum_i x_{i,c}$; $\text{SEM}_c = \text{SD}(x_{i,c})/\sqrt{N_p}$	Mean ± SEM
Friedman; Kendall W	$W = \chi^2/[N_p(k-1)]$	Complementary
Holm adjustment	Step-down adjustment of pairwise p-values	Multiple testing

Note. i, c, r, s, and t denote participant, condition, response category, strategy, and trial; N_p is the number of participants and $N_{i,c}$ the trial count. Categories are Correct, Vague, Uncertain, and Wrong. Only Correct versus all others enters the binary GLMM; the four categories remain separate in descriptive summaries.

For trial-level inference, Correct was coded as 1 and all other responses as 0. Separate binomial generalized linear mixed models (GLMMs) were fitted to the two experiments. Each model included encoding strategy, experimental condition, and their interaction as fixed effects, with participant and source-video random intercepts. The experimental condition was temporal availability in Experiment 1 and preprocessing in Experiment 2. The general model was

$$\text{logit}[\Pr(Y_t = 1)] = \beta_0 + \beta_{S(t)} + \beta_{C(t)} + \beta_{SC(t)} + u_{\text{participant}(t)} + v_{\text{video}(t)} \quad (11)$$

Here, Y_t indicates correct identification in trial t, $S(t)$ denotes the encoding strategy, and $C(t)$ denotes the experimental condition; u and v are participant- and video-specific random intercepts. Sum-to-zero contrasts and Type III Wald χ^2 tests were used for fixed effects. Estimated correct-response probabilities and 95% confidence intervals were obtained for each strategy-by-condition combination.

Models were fitted in R using the `lme4` package. Participant random intercepts accounted for repeated responses from the same observer, whereas source-video random intercepts accounted for repeated presentations of the same base video. We examined convergence warnings, singular fits, random-intercept variances, overdispersion, and potential complete separation.

As a data-quality sensitivity check, the GLMMs were refitted after excluding trials associated with anomalous stimulus-file durations (seven trials in Experiment 1 and six in Experiment 2) while retaining the same fixed- and random-effects structure.

Participant-level Correct and Uncertain rates were also compared using Friedman tests, with Kendall's W as the effect-size measure [28,29]. Where applicable, Wilcoxon signed-rank tests with Holm adjustment were used for pairwise comparisons [30,31]. Early key presses and oral completion times were analyzed descriptively. Video-level PBI cate-

gories were examined as exploratory participant-level groupings. The experiments were analyzed separately.

3. Results

The behavioral results were analyzed at the trial and participant levels. Table 8 summarizes the fixed-effect tests from the binomial GLMMs, and Table 9 reports the complementary participant-level Friedman tests.

Table 8. Type III Wald tests of fixed effects in the binomial GLMMs.

Exp.	Fixed Effect	df	χ^2	p Value
Exp1	Strategy	3	54.245	9.95×10^{-12}
Exp1	Temporal availability	2	63.407	1.70×10^{-14}
Exp1	Strategy × temporal availability	6	13.127	0.0411
Exp2	Strategy	3	11.771	0.0082
Exp2	Preprocessing	2	13.240	0.0013
Exp2	Strategy × preprocessing	6	9.128	0.1665

Note. Correct = 1 and all other response categories = 0. Separate models were fitted to Exp1 (n = 1440 trials) and Exp2 (n = 720 trials), with participant and source-video random intercepts.

Both primary GLMMs converged without `lme4` convergence warnings, and neither was singular. The participant and source-video random-intercept variances were 0.5381 and 4.8880, respectively, in Experiment 1, and 0.1256 and 4.7560 in Experiment 2. The corresponding overdispersion ratios were 0.8424 and 0.7759, and no complete-separation condition cells were identified.

The sensitivity analysis retained 1433 Experiment 1 trials and 714 Experiment 2 trials. After exclusion of anomalous-duration trials, the strategy-by-temporal-availability interaction in Experiment 1 remained significant ($\chi^2(6) = 13.473$, $p = 0.0361$), while the strategy-by-preprocessing interaction in Experiment 2 remained nonsignificant ($\chi^2(6) = 9.337$, $p = 0.1555$).

These findings indicate that the interaction conclusions did not change under this particular data-quality exclusion; they do not establish robustness to every alternative model specification.

Table 9. Friedman omnibus tests of participant-level behavioral summaries.

Exp.	Outcome	Factor	χ^2	df	p	W
Exp1	Correct	Strategy	28.681	3	2.61×10^{-6}	0.478
Exp1	Uncertain	Strategy	33.247	3	2.86×10^{-7}	0.554
Exp1	Correct	Temporal availability	29.325	2	4.29×10^{-7}	0.733
Exp1	Uncertain	Temporal availability	24.320	2	5.24×10^{-6}	0.608
Exp1	Correct-rate drop	Strategy (Full–Quarter)	1.467	3	0.6899	0.024
Exp1	Correct	Video PBI category	5.840	2	0.0539	0.146
Exp1	Uncertain	Video PBI category	7.707	2	0.0212	0.193
Exp2	Correct	Preprocessing	7.861	2	0.0196	0.197
Exp2	Uncertain	Preprocessing	6.594	2	0.0370	0.165

Note. These are participant-level analyses complementary to the trial-level GLMMs; W is Kendall's effect size. Video PBI categories were Slow, Mixed, and Fast, as defined by the model.

3.1. Experiment 1: Encoding Strategies and Temporal Availability

The four encoding strategies showed different response profiles in Experiment 1 (Figure 4). Friedman tests indicated differences among strategies in both Correct ($p = 2.61 \times 10^{-6}$) and Uncertain response rates ($p = 2.86 \times 10^{-7}$; Table 9).

Across temporal-availability conditions, PCO had the highest observed Correct rate (58.6%), followed by BASELINE (48.3%), NO_EDGE (39.7%), and PCL (37.8%). The corresponding Uncertain rates were 20.8%, 35.3%, 41.7%, and 40.3%, respectively. Response composition also differed: PCO had a Wrong rate of 14.4%, compared with 9.7% for BASELINE (Figure 4C).

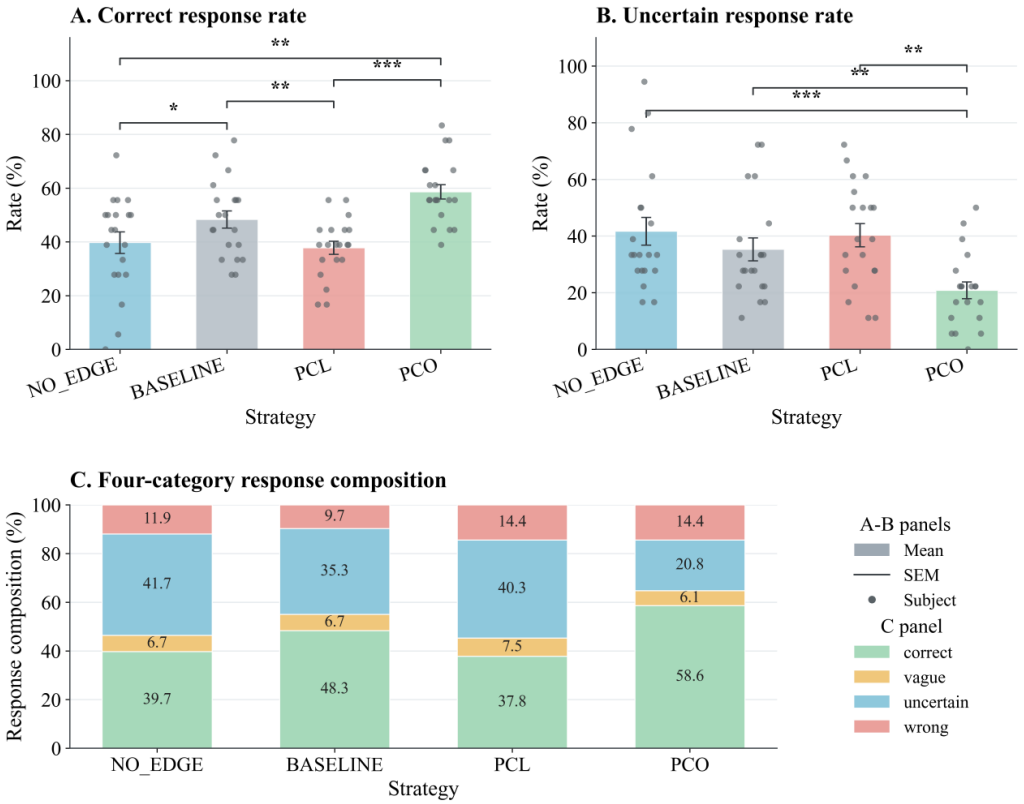


Figure 4. Effects of encoding strategy on open-response performance in Exp1. (**A**) Correct response rate. (**B**) Uncertain response rate. (**C**) Four-category response composition. For bar plots, bars represent participant-level means, error bars represent SEM, and dots represent individual participants. Significance markers indicate Holm-corrected Wilcoxon post-hoc comparisons. Note. Friedman tests were used for within-subject comparisons. Kendall's W is reported as the effect size. Post-hoc comparisons were performed using Wilcoxon signed-rank tests with Holm correction. Asterisks indicate Holm-adjusted significance: * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$.

Recognition changed across temporal-availability conditions (Figure 5). Correct response rates decreased from 55.8% in Full to 47.5% in Half and 35.0% in Quarter, while Uncertain rates increased from 25.4% to 31.9% and 46.2%, respectively (Table 10).

Friedman tests indicated differences across conditions for both Correct ($p = 4.29 \times 10^{-7}$) and Uncertain responses ($p = 5.24 \times 10^{-6}$; Table 9).

Table 10. Response-category composition across temporal-availability conditions in Experiment 1.

Condition	Correct (%)	Vague (%)	Uncertain (%)	Wrong (%)	Mean Rating
Full	55.8	7.9	25.4	10.8	2.09
Half	47.5	6.0	31.9	14.6	1.87
Quarter	35.0	6.2	46.2	12.5	1.64

Note. Values are participant-level means. The four response categories are mutually exclusive; percentages may not sum to exactly 100% because of rounding.

Figure 6 presents the observed and GLMM-estimated Correct response rates across strategy and temporal-availability conditions. The trial-level GLMM identified effects of strategy ($\chi^2(3) = 54.245$, $p = 9.95 \times 10^{-12}$) and temporal availability ($\chi^2(2) = 63.407$, $p = 1.70 \times 10^{-14}$), together with a strategy-by-condition interaction ($\chi^2(6) = 13.127$, $p = 0.0411$; Table 8).

PCO showed higher observed Correct rates in Full and Half, whereas BASELINE (43.3%) and PCO (45.0%) showed similar observed rates in Quarter. NO_EDGE (26.7%) and PCL (25.0%) had lower observed Correct rates in Quarter (Table 11). The interaction indicated differences in the strategy-related patterns across the three temporal-availability conditions.

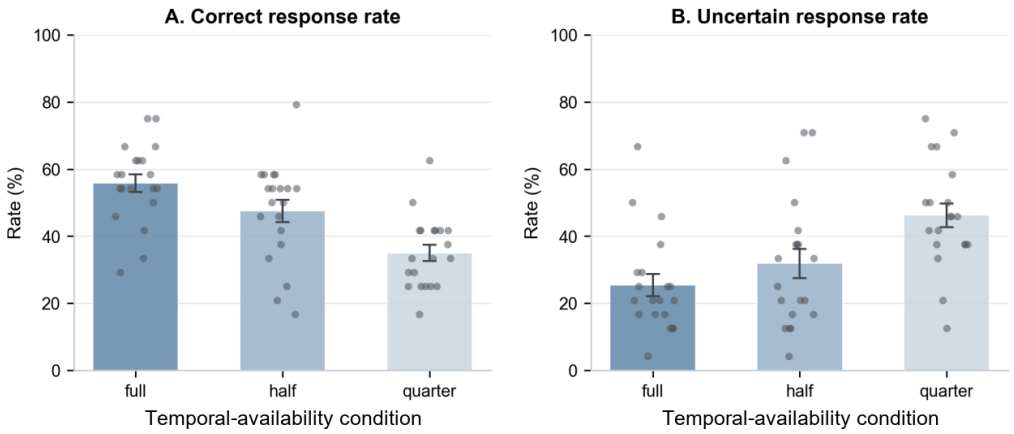


Figure 5. Effects of temporal-availability conditions on open-response performance in Experiment 1. (**A**) Correct response rate. (**B**) Uncertain response rate. Bars represent participant-level means, error bars indicate SEM, and points represent individual participants.

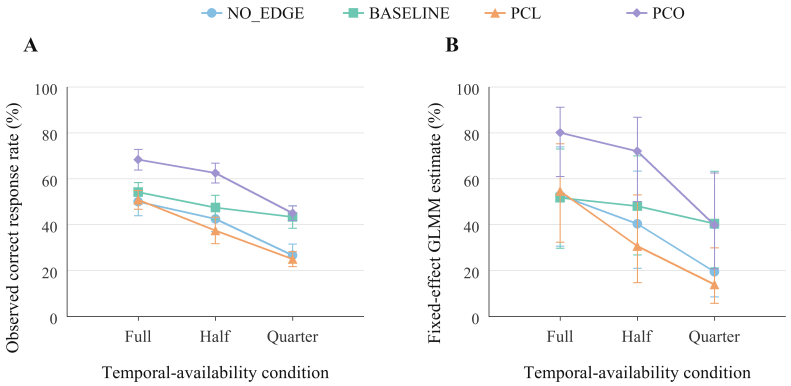


Figure 6. Observed and GLMM-estimated correct response rates across encoding strategies and temporal-availability conditions in Experiment 1. (**A**) Participant-level mean correct response rates with standard errors of the mean (SEM). (**B**) Fixed-effect GLMM estimates of correct response probabilities with 95% confidence intervals. The temporal-availability conditions also differed in stimulus presentation duration.

The observed Full-to-Quarter decrease was smallest for BASELINE (10.83 percentage points) and largest for PCL (25.83 percentage points; Table 11). However, the participant-level comparison of these paired decreases was not significant ($\chi^2(3) = 1.467$, $p = 0.6899$, $W = 0.024$; Table 9).

This endpoint comparison differed from the GLMM interaction, which evaluated the complete strategy-by-condition pattern.

Early key-press rates decreased from 24.58% in Full to 9.17% in Half and 1.67% in Quarter, while median oral completion times increased from 1.76 to 2.13 and 2.39 s, respectively (Figure 7).

Table 11. Full-to-Quarter decrease in correct response rate across encoding strategies.

Strategy	Full (%)	Quarter (%)	Drop (pp)
NO_EDGE	50.00 ± 6.05	26.67 ± 4.90	23.33 ± 6.99
BASELINE	54.17 ± 4.17	43.33 ± 4.90	10.83 ± 6.31
PCL	50.83 ± 4.10	25.00 ± 3.31	25.83 ± 5.73
PCO	68.33 ± 4.51	45.00 ± 3.22	23.33 ± 5.72

Note. Values are participant-level mean ± SEM. Drops are paired Full-minus-Quarter differences in percentage points (pp); drop SEMs were calculated from participant-level paired differences. The strategy comparison of these drops was not significant (Friedman $\chi^2(3) = 1.467$, $p = 0.6899$).

Stimulus-file durations also decreased across these conditions (Table 12), reducing the time available for an early key press.

Table 12. Stimulus durations across experimental conditions.

Experiment	Condition	Duration, Median (Range), s
Exp1	Full	5.00 (2.04–5.04)
	Half	2.50 (1.03–2.54)
	Quarter	1.26 (0.53–1.28)
Exp2	P0	5.00 (2.64–5.02)
	P1	5.01 (2.64–5.03)
	P2	5.01 (2.64–5.03)

Note. Durations were summarized for unique stimulus video files within each condition. Values represent video file durations rather than actual viewing times, which could be shortened by early key presses.

Additional timing measures, including early key-press rates and oral completion times, are summarized in Table 13.

Table 13. Auxiliary timing measures in Experiments 1 and 2.

Exp.	Condition	Formal Trials	Early Presses, *n* (%)	Valid Oral Times, *n*	Oral Completion, Median (IQR), s
Exp1	Full	480	118 (24.58)	472	1.76 (1.02–2.82)
Exp1	Half	480	44 (9.17)	473	2.13 (1.40–3.24)
Exp1	Quarter	480	8 (1.67)	476	2.39 (1.60–3.81)
Exp2	P0	240	98 (40.83)	235	1.35 (0.72–2.12)
Exp2	P1	240	87 (36.25)	236	1.39 (0.79–2.17)
Exp2	P2	240	102 (42.50)	233	1.29 (0.67–2.05)

Note. Early-press percentages use all formal trials within each condition as the denominator. Valid oral-time counts exclude trials without usable timing records; oral completion times are descriptive medians and interquartile ranges.

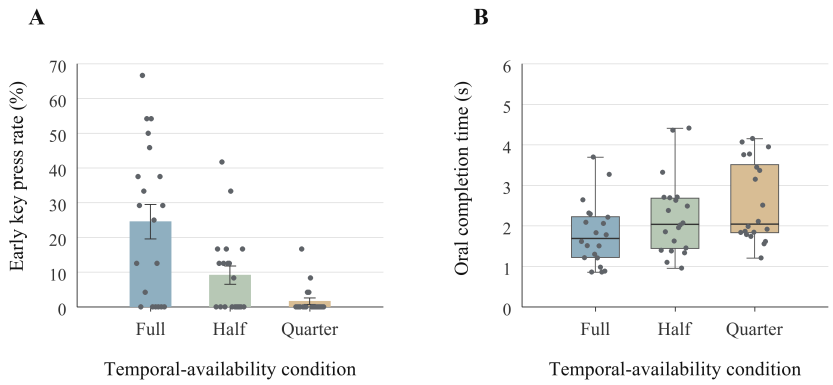


Figure 7. Auxiliary timing measures across temporal-availability conditions in Experiment 1. (**A**) Early key press rates. Bars represent participant-level means, error bars indicate standard errors of the mean (SEM), and points represent individual participants. (**B**) Oral completion times. Boxes represent the interquartile range, horizontal lines indicate the median, and points represent individual participants.

3.2. Experiment 2: Effects of Preprocessing on Recognition

Experiment 2 examined the effects of preprocessing on Correct and Uncertain responses (Figure 8). Participant-level Friedman tests indicated differences among preprocessing conditions for Correct ($p = 0.0196$) and Uncertain rates ($p = 0.0370$; Table 9).

P1 showed lower overall Correct rates than P0 and P2. Holm-adjusted comparisons identified differences between P0 and P1 and between P1 and P2 for Correct rates, and between P0 and P1 for Uncertain rates (Figure 8).

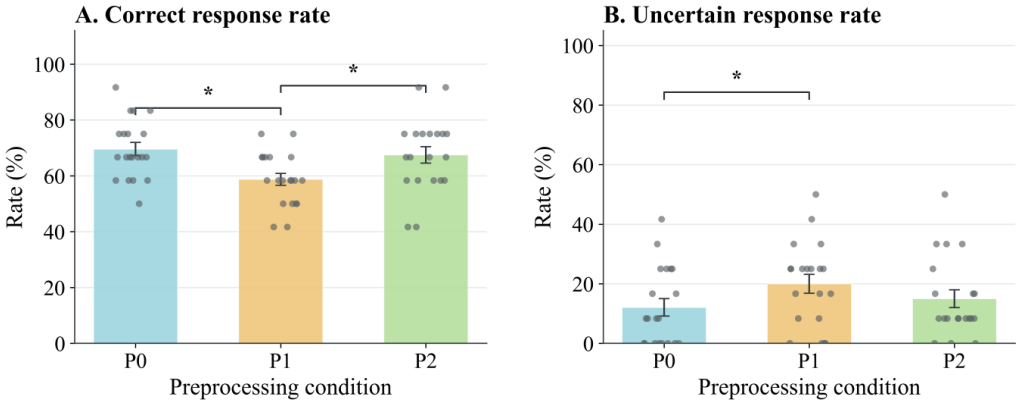


Figure 8. Effects of preprocessing on open-response performance in Experiment 2. (**A**) Correct response rate. (**B**) Uncertain response rate. Bars represent participant-level means, error bars indicate standard errors of the mean (SEM), and points represent individual participants. Brackets and asterisks indicate Holm-adjusted paired Wilcoxon comparisons (* $p_{\text{Holm}} < 0.05$): P0–P1 and P1–P2 in panel A, and P0–P1 in panel B.

The trial-level GLMM showed effects of strategy ($\chi^2(3) = 11.771$, $p = 0.0082$) and preprocessing ($\chi^2(2) = 13.240$, $p = 0.0013$), while their interaction was not significant ($\chi^2(6) = 9.128$, $p = 0.1665$; Table 8).

The model-estimated Correct response probabilities for each combination are presented in Figure 9.

Strategy-specific response rates are summarized in Table 14. Under P0, PCL had the highest observed Correct rate (75.0%). Under P1, Correct rates decreased across all four strategies, but the observed decreases differed: 20.0 percentage points for NO_EDGE, 13.3 for BASELINE, 6.7 for PCL, and 3.3 for PCO.

Consequently, PCL (68.3%) and PCO (66.7%) had higher observed Correct rates under P1 than BASELINE (53.3%) and NO_EDGE (46.7%). Under P2, PCO had the highest observed Correct rate (73.3%) and the lowest Uncertain rate (5.0%). These descriptive differences are considered separately from the non-significant factorial interaction reported above.

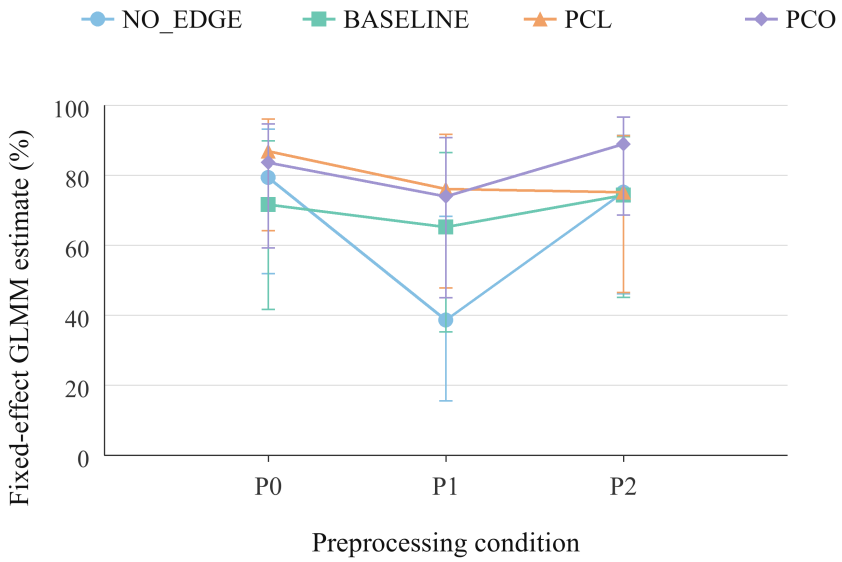


Figure 9. Fixed-effect GLMM estimates of correct response probabilities across preprocessing conditions and encoding strategies in Experiment 2. Estimates are presented on a percentage scale, and error bars indicate 95% confidence intervals. P0 denotes no preprocessing, P1 denotes frame holding ($k = 6$), and P2 denotes background suppression ($\beta = 0.5$).

Table 14. Correct and uncertain response rates under preprocessing × strategy conditions in Exp2.

Preprocessing	Strategy	Correct Rate, Mean ± SEM (%)	Uncertain Rate, Mean ± SEM (%)
P0	NO_EDGE	66.7 ± 7.6	18.3 ± 7.0
P0	BASELINE	66.7 ± 5.9	11.7 ± 4.4
P0	PCL	75.0 ± 6.3	6.7 ± 3.1
P0	PCO	70.0 ± 4.8	11.7 ± 3.6
P1	NO_EDGE	46.7 ± 8.9	31.7 ± 6.6
P1	BASELINE	53.3 ± 5.6	20.0 ± 5.6
P1	PCL	68.3 ± 7.4	13.3 ± 5.6
P1	PCO	66.7 ± 5.9	15.0 ± 4.5
P2	NO_EDGE	65.0 ± 6.2	16.7 ± 3.8
P2	BASELINE	65.0 ± 6.6	20.0 ± 5.6
P2	PCL	66.7 ± 5.9	18.3 ± 6.6
P2	PCO	73.3 ± 5.2	5.0 ± 2.7

Note. Values are participant-level mean ± SEM. This table reports descriptive statistics only.

Early key-press rates ranged from 36.25% to 42.50% across P0–P2, and median oral completion times ranged from 1.29 to 1.39 s (Table 13). Median stimulus-file durations were approximately 5 s under all three preprocessing conditions (Table 12).

3.3. Exploratory Associations Between Model-Defined Pathway Bias and Human Responses

The exploratory analysis first compared human responses across model-defined video PBI categories (Slow, Mixed, and Fast; Figure 10). The Correct response rate was descrip-

tively lower for Fast videos, but the Friedman test did not reach significance ($\chi^2(2) = 5.840$, $p = 0.0539$; Table 9).

The Uncertain response rate differed across video PBI categories ($\chi^2(2) = 7.707$, $p = 0.0212$), with a higher rate for Fast than Mixed videos in the Holm-adjusted comparison ($p_{\text{Holm}} = 0.0168$).

At the strategy level, the selected encodings showed different relationships between model-defined profiles and human recognition patterns.

PCO had a mixed pathway-ablation profile and the highest overall observed Correct rate in Experiment 1. BASELINE was Slow-biased and maintained a Correct rate close to PCO under Quarter, whereas PCL was Fast-biased in the computational evaluation but showed lower overall recognition in Experiment 1.

These findings describe the combined model and behavioral characteristics of the selected encodings rather than a direct numerical mapping between pathway-ablation scores and human response rates.

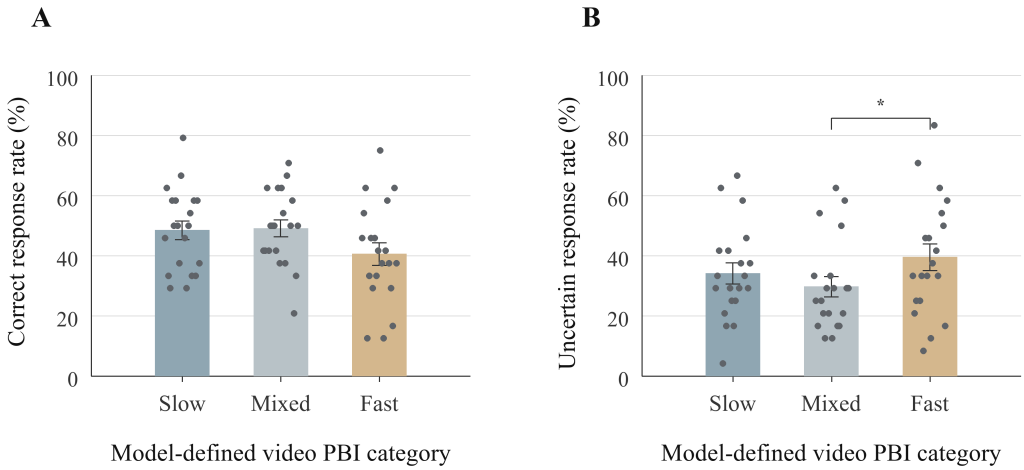


Figure 10. Exploratory association between model-defined video pathway-bias categories and human behavioral responses in Experiment 1. (**A**) Correct response rates. (**B**) Uncertain response rates. Bars represent participant-level means, error bars indicate standard errors of the mean (SEM), and points represent individual participants. The asterisk (*) indicates a statistically significant Mixed–Fast comparison ($p_{\text{Holm}} < 0.05$; observed $p_{\text{Holm}} = 0.0168$), based on a paired Wilcoxon signed-rank test with Holm correction.

4. Discussion

The selected SPV encodings differed in how they represented target boundaries, phosphene arrangements, and information across frames. Their behavioral profiles in Experiment 1 are considered below in relation to these design operations, followed by the temporal-availability findings, complementary preprocessing results, and model-based comparisons.

4.1. Gestalt-Inspired Organization and Behavioral Characteristics of the Four Strategies

Gestalt grouping principles provide a basis for interpreting how sparse phosphene elements form recognizable target structures [15,16]. The four selected encodings implemented different combinations of contour visibility, element allocation, and temporal presentation. Their effects are therefore assessed at the level of the complete displayed representations.

NO_EDGE retained a coarse grayscale representation without additional target-boundary enhancement. Its relatively low Correct rate and frequent Uncertain responses in Experiment 1 (Figure 4) suggest that target–background separation and action-bearing con-

tours were not consistently available from the unenhanced array. As the reference condition, NO_EDGE provided a basis for assessing the effects of introducing structural emphasis within the same SPV framework.

BASELINE enhanced the target boundary, making shape, orientation, and posture more accessible despite the limited internal detail available in the coarse SPV representation. This operation is consistent with the perceptual roles of good continuation and figure-ground separation: an identifiable outline can support grouping when individual elements are sparse. BASELINE had a higher overall observed Correct rate than NO_EDGE and approached PCO in the Quarter condition (Figure 6; Table 11). Its comparatively favorable performance under restricted temporal availability is consistent with a possible benefit of retaining a recognizable contour within individual frames. A contour alone, however, may remain insufficient when actions share similar silhouettes or depend on changes within the target.

PCO combined proximity-oriented target sampling with directional rendering of contour elements. Concentrating elements around the target increased the representation of task-relevant regions, while locally aligned elliptical phosphenes supported the organization of discontinuous boundary fragments. PCO showed the highest overall observed Correct rate and the lowest Uncertain rate in Experiment 1 (Figure 4). This pattern is consistent with the joint use of target-focused allocation and contour continuity to support action identification. Its Wrong rate was nevertheless higher than that of BASELINE, indicating that improved identification and fewer uncertain reports did not imply fewer incorrect specific responses in every category. The response composition is therefore relevant alongside the Correct rate when comparing these encodings.

PCL allocated phosphenes toward the target and displayed complementary information across successive frames. Its closure-inspired design may require observers to integrate fragments shown at different moments rather than rely on a continuously available contour. Such presentation may help complete an incomplete target representation, but it also introduces frame-to-frame changes in visible elements that need not reflect the action itself. PCL showed lower overall recognition in Experiment 1, particularly under Quarter (Figure 6; Table 11). This pattern may reflect the demands of integrating temporally distributed target information, while post-encoding subsampling in Half and Quarter also disrupted the intended odd–even complementary frame pairs. The relative contributions of these factors cannot be distinguished in the present experiment.

Together, the four profiles distinguish the benefits of a readily visible boundary from those of redistributed or temporally complementary information. The relevant comparison is how each complete encoding organizes the cues available for action recognition rather than the effect of any one Gestalt principle in isolation.

4.2. Temporal Availability and Recognition in the Primary Experiment

Reducing temporal availability was accompanied by fewer Correct identifications and more explicit Uncertain responses (Figure 5; Table 10). The observed shift is consistent with reduced temporal evidence and shorter viewing opportunities, making action identification more difficult; these components were not independently manipulated. Vague and Wrong responses showed less consistent changes, emphasizing that the main behavioral shift involved identification success and explicit uncertainty.

The strategy-by-temporal-availability interaction (Figure 6; Table 8) showed that recognition patterns differed across the three conditions. PCO had the highest observed Correct rates in Full and Half, whereas BASELINE approached PCO in Quarter (Figure 6). The visible target boundary in BASELINE could remain informative within a reduced set of frames, whereas temporal subsampling in PCL disrupted the intended complementary

frame pairings. Although BASELINE had the smallest observed Full-to-Quarter decline, the between-strategy comparison of endpoint declines was not significant (Tables 9 and 11). This endpoint result does not replace the three-level interaction; the two analyses describe different aspects of the response profiles.

Half and Quarter also shortened presentation duration, so these findings concern the combined reduction in available frames and viewing time. The decline in early key presses and increase in oral completion time (Figure 7; Tables 12 and 13) provide auxiliary descriptions of responding under these conditions without separating their temporal components.

4.3. Complementary Findings from Preprocessing in Experiment 2

Experiment 2 examined preprocessing before SPV encoding using a separate set of source videos. Frame holding (P1) generally reduced Correct responses relative to the other preprocessing conditions (Figure 8). Repeating source frames may preserve viewing duration while reducing changes in posture, movement, and target–object relations that help distinguish actions.

The observed P1 profiles were not identical across encodings. PCL and PCO had higher observed Correct rates than NO_EDGE and BASELINE under P1 (Table 14). For PCL, frame holding was applied before SPV generation, whereas complementary fragment presentation was produced during encoding. Successive encoded frames could therefore differ despite repeated source-frame content. This processing order offers a possible explanation for the observed PCL profile. However, the strategy-by-preprocessing interaction was not significant (Table 8), so these condition-specific differences remain descriptive.

Background suppression (P2) reduced scene intensity relative to the target and thus modified figure-ground organization. Unlike P1, P2 did not show a consistent overall Correct-rate improvement over P0. While target separation can be useful when background interference limits recognition, scene context may also assist action identification. The P2 results therefore illustrate how preprocessing changes the information presented by an encoding, without supporting a general preference for background suppression across strategies (Figure 9; Table 14). These findings complement, rather than rank against, the results of Experiment 1.

4.4. Computational Screening, Human Responses, and Related SPV Research

SlowFast pathway-input ablation distinguished the candidate encodings by their computational sensitivity profiles and guided the selection of representative strategies (Figure 3). Human recognition did not follow those profiles as a direct numerical mapping: PCO combined a mixed model-defined profile with the highest overall observed Correct rate in Experiment 1, BASELINE remained comparatively effective under Quarter, and PCL showed strong Fast-pathway dependence in the model but lower overall human recognition in the primary task. The model thus supported selection of contrasting inputs, while psychophysical responses characterized their behavioral consequences.

The exploratory video-level PBI analysis addressed a different question. Videos classified as Fast showed more Uncertain responses than Mixed videos in the adjusted comparison, whereas the corresponding Correct-rate omnibus test did not reach significance (Figure 10; Table 9). Unlike strategy-level sensitivity, video-level PBI described differences among source-video inputs; the two classifications should not be interpreted as interchangeable measures.

Previous SPV studies have investigated saliency enhancement, structural segmentation, scene simplification, and foreground processing for different tasks [4,7,9,10,12]. Other work has examined sparse contour geometry, task-oriented object representations, and temporally presented semantic information [11,14,17].

More specifically, Wang et al. [12] reported improvements in moving-object recognition using background-subtraction-based processing, whereas background suppression (P2) in our study did not produce a consistent overall improvement. Scialom et al. [17] investigated sparse contour segments for object recognition, providing a relevant comparison for PCO's contour-oriented representation. However, PCO also modified spatial sampling, so its recognition advantage cannot be attributed to element geometry alone.

Unlike these studies, the present work focuses on open-response dynamic action recognition under four complete SPV encoding implementations and different temporal presentation conditions. Differences in tasks, stimuli, and experimental settings preclude direct comparisons of recognition accuracy across studies.

4.5. Implications and Study Scope

For SPV action recognition, the findings indicate a need to preserve coherent target contours while maintaining changes relevant to the action. PCO illustrates target-focused allocation with oriented boundary elements; BASELINE illustrates the utility of an explicit target outline; and PCL suggests that recognition from temporally distributed fragments may benefit from greater persistence and more stable spatial cues. Background suppression may be useful selectively, depending on the action and its contextual cues.

The experiments involved normally sighted participants viewing simulated 32 × 32 SPV stimuli. Implanted visual percepts may differ in phosphene shape, persistence, and viewing behavior [21–23]. Moreover, differences in boundary enhancement, phosphene allocation, and temporal presentation were constitutive features of the compared encodings, not independently manipulated components. Although response coding followed uniform criteria and was blinded to strategy and experimental-condition identifiers, all responses were classified by one researcher. Independent inter-rater reliability could therefore not be estimated. In particular, the classification of ambiguous verbal responses may remain subject to assessor-related variation, which could affect the Correct-versus-other outcome used in the GLMMs. This limitation should be considered when interpreting the behavioral effects, and future studies should incorporate independent, condition-blinded scoring. Trial allocation was balanced at the strategy-by-condition and source-video levels, although not every full stimulus combination was replicated equally. The mixed models accounted for participant and source-video clustering; potential residual effects of action category, source-video difficulty, and repeated exposure remain. Further work can examine these components separately and assess their effects under more realistic prosthetic viewing conditions.

5. Conclusions

This study evaluated human dynamic action recognition under four selected Gestalt-inspired SPV encodings in two psychophysical experiments. Auxiliary SlowFast-based screening was used to select the four conditions from 11 candidates. In the primary psychophysical experiment, the encodings produced different profiles of correct identification and explicit uncertainty at a nominal 32 × 32 resolution. PCO, combining target-focused allocation and continuity-oriented contour rendering, showed the highest overall observed Correct rate (58.6%) and the lowest Uncertain rate (20.8%). BASELINE approached PCO in Quarter, whereas the complementary presentation of PCL showed lower overall recognition in Experiment 1.

Conditions with fewer available frames and shorter presentation durations showed lower Correct rates and higher Uncertain rates, with strategy-dependent response profiles. These findings do not isolate frame-rate effects. The complementary preprocessing experiment showed an overall reduction in recognition under frame holding and no consistent

improvement from background suppression over no preprocessing. The strategy-specific patterns were descriptive, as the strategy-by-preprocessing interaction was not statistically significant. Experiment 2 was not intended as a second ranking of the encoding strategies.

The computational profiles informed strategy selection, while human responses established the recognition characteristics of the selected encodings. Together, the findings emphasize the organization of target structure, temporal changes, and contextual information when designing SPV representations for dynamic action recognition.

Author Contributions: Y.W. designed and conducted the study, performed the experiments and data analysis, and wrote the manuscript. Y.Z. supervised the research, provided guidance on key research decisions, and revised the manuscript. Z.F., L.G., and Y.H. reviewed the manuscript and provided feedback. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (81460279 and 61841204); the Inner Mongolia Natural Science Foundation (2024LHMS08045 and 2018LH08066); and the Inner Mongolia Autonomous Region High School Science Research Foundation (NJZY145).

Institutional Review Board Statement: The study was approved by the Medical Ethics Committee of Xilingol League Mongolian Medicine Hospital, Inner Mongolia, China (approval date: 5 June 2026).

Informed Consent Statement: Written informed consent was obtained from all participants involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. The original UCF-101 dataset can be obtained from its public repository. The processed SPV stimuli and anonymized behavioral data supporting the conclusions of this article will be made available by the corresponding author upon reasonable request.

Acknowledgments: The authors thank all participants.

Conflicts of Interest: The authors declare no conflicts of interest. The funders provided financial support for the research but had no role in the study design, data collection, analysis, interpretation, manuscript preparation, or decision to publish.

References

1. Brindley, G.S.; Lewin, W.S. The sensations produced by electrical stimulation of the visual cortex. *J. Physiol.* **1968**, *196*, 479–493. [CrossRef]
2. Chen, S.C.; Suaning, G.J.; Morley, J.W.; Lovell, N.H. Simulating prosthetic vision: I. Visual models of phosphenes. *Vis. Res.* **2009**, *49*, 1493–1506. [CrossRef]
3. da Cruz, L.; Coley, B.F.; Dorn, J.D.; Merlini, F.; Filley, E.; Christopher, P.; Chen, F.K.; Wuyyuru, V.; Sahel, J.; Stanga, P.; et al. The Argus II epiretinal prosthesis system allows letter and word reading and long-term function in patients with profound vision loss. *Br. J. Ophthalmol.* **2013**, *97*, 632–636. [CrossRef]
4. van Rheede, J.J.; Kennard, C.; Hicks, S.L. Simulating prosthetic vision: optimizing the information content of a limited visual display. *J. Vis.* **2010**, *10*, 32. [CrossRef]
5. Chen, S.C.; Suaning, G.J.; Morley, J.W.; Lovell, N.H. Simulating prosthetic vision: II. Measuring functional capacity. *Vis. Res.* **2009**, *49*, 2329–2343. [CrossRef]
6. Beyeler, M.; Boynton, G.M.; Fine, I.; Rokem, A. pulse2percept: A Python-based simulation framework for bionic vision. In Proceedings of the 16th Python in Science Conference, Austin, TX, USA, 10–16 July 2017; pp. 81–88. [CrossRef]
7. Parikh, N.; Itti, L.; Weiland, J. Saliency-based image processing for retinal prostheses. *J. Neural Eng.* **2010**, *7*, 016006. [CrossRef]
8. Guo, F.; Yang, Y.; Gao, Y. Optimization of visual information presentation for visual prosthesis. *Int. J. Biomed. Imaging* **2018**, *2018*, 3198342. [CrossRef]
9. Sanchez-Garcia, M.; Martinez-Cantin, R.; Guerrero, J.J. Semantic and structural image segmentation for prosthetic vision. *PLoS ONE* **2020**, *15*, e0227677. [CrossRef]
10. Han, N.; Srivastava, S.; Xu, A.; Klein, D.; Beyeler, M. Deep learning-based scene simplification for bionic vision. In Proceedings of the Augmented Humans International Conference 2021 (AHs '21), Rovaniemi, Finland, 22–24 February 2021; pp. 45–54. [CrossRef]

11. Nejad, A.; Küçükoğlu, B.; de Ruyter van Steveninck, J.; Bedrossian, S.; Heutink, J.; de Haan, G.A.; Cornelissen, F.W.; van Gerven, M. Point-SPV: end-to-end enhancement of object recognition in simulated prosthetic vision using synthetic viewing points. *Front. Hum. Neurosci.* **2025**, *19*, 1549698. [CrossRef]
12. Wang, J.; Lu, Y.; Gu, L.; Zhou, C.; Chai, X. Moving object recognition under simulated prosthetic vision using background-subtraction-based image processing strategies. *Inf. Sci.* **2014**, *277*, 512–524. [CrossRef]
13. Perez-Yus, A.; Bermudez-Cameo, J.; Lopez-Nicolas, G.; Guerrero, J.J. Depth and motion cues with phosphene patterns for prosthetic vision. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 1516–1525. [CrossRef]
14. Kasowski, J.M.; Varshney, A.; Beyeler, M. Static or Temporal? Semantic Scene Simplification to Aid Wayfinding in Immersive Simulations of Bionic Vision. In Proceedings of the 2025 31st ACM Symposium on Virtual Reality Software and Technology (VRST '25), Montreal, QC, Canada, 12–14 November 2025; pp. 1–11. [CrossRef]
15. Wertheimer, M. Laws of organization in perceptual forms. In *A Source Book of Gestalt Psychology*; Ellis, W.D., Ed.; Routledge & Kegan Paul: London, UK, 1938; pp. 71–88. Original work published 1923.
16. Wagemans, J.; Elder, J.H.; Kubovy, M.; Palmer, S.E.; Peterson, M.A.; Singh, M.; von der Heydt, R. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychol. Bull.* **2012**, *138*, 1172–1217. [CrossRef]
17. Scialom, E.; Ernst, U.A.; Rotermund, D.; Herzog, M.H. Object recognition from sparse simulated phosphenes and curved segments. *Vis. Res.* **2026**, *238*, 108718. [CrossRef]
18. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
19. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211. [CrossRef]
20. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human action classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
21. Hou, Y.; Pullela, L.; Su, J.; Aluru, S.; Sista, S.; Lu, X.; Beyeler, M. Predicting the temporal dynamics of prosthetic vision. In Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 15–19 July 2024; pp. 1–4. [CrossRef]
22. de Ruyter van Steveninck, J.; Nipshagen, M.; van Gerven, M.A.J.; Güçlü, U.; Güçlütürk, Y.; van Wezel, R.J.A. Gaze-contingent processing improves mobility, scene recognition and visual search in simulated head-steered prosthetic vision. *J. Neural Eng.* **2024**, *21*, 026037. [CrossRef]
23. van der Grinten, M.; de Ruyter van Steveninck, J.; Lozano, A.; Pijnacker, L.; Rueckauer, B.; Roelfsema, P.; van Gerven, M.; van Wezel, R.; Güçlü, U.; Güçlütürk, Y. Towards biologically plausible phosphene simulation for the differentiable optimization of visual cortical prostheses. *eLife* **2024**, *13*, e85812. [CrossRef]
24. World Medical Association. Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants. 2024. Available online: https://www.wma.net/policies-post/wma-declaration-of-helsinki/ (accessed on 29 October 2026).
25. Peirce, J.W.; Gray, J.R.; Simpson, S.; MacAskill, M.R.; Höchenberger, R.; Sogo, H.; Kastman, E.; Lindeløv, J.K. PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **2019**, *51*, 195–203. [CrossRef]
26. Kim, T.; Kim, K.; Lee, J.; Cha, D.; Lee, J.; Kim, D. Revisiting Image Pyramid Structure for High Resolution Salient Object Detection. In Proceedings of the Asian Conference on Computer Vision (ACCV), Macau, China, 4–8 December 2022; pp. 108–124.
27. Wang, Y.; Zhao, Y.; Fan, Z.; Gan, L.; Han, Y. Encoding strategies for simulated prosthetic vision based on gestalt psychology. *Signal Image Video Process.* **2026**, *in press*.
28. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [CrossRef]
29. Kendall, M.G.; Babington Smith, B. The problem of m rankings. *Ann. Math. Stat.* **1939**, *10*, 275–287. [CrossRef]
30. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
31. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.