CORTEXLINK: RECONSTRUCTING LONG VIDEOS FROM FMRI WITH EVENT-LEVEL SEMANTIC CONSENSUS AND CONTINUOUS GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Current fMRI-to-video systems typically reconstruct continuous visual experience as independent short clips, overlooking the event structure shared by neighboring neural measurements. This mismatch discards repeated semantic evidence and amplifies local uncertainty into content drift and geometric discontinuities across clip boundaries. We introduce CortexLink, an event-aligned framework that reorganizes clip-level predictions within stimulus shots. Its semantic branch treats decoded descriptions as noisy observations of a shared event and distills them into a consensus anchor through Minimum Bayes Risk decoding. Its geometric branch combines shot-level spatial aggregation during training with groupwise shape–pose decomposition at inference, recovering a shared subject template and a continuous pose trajectory. The resulting semantic and geometric anchors jointly guide long-video synthesis and reset at event boundaries. Experiments show consistent improvements in semantic fidelity, motion agreement, and cross-clip continuity over clip-level baselines. On shots lasting at least 10 seconds, CortexLink improves 50-way identification by 15.5%, while reducing Dynamic Degree error by 51.8% and Boundary Seam by 67.3% relative to CortexVideo. Component analyses further verify the complementary contributions of semantic consensus and continuous geometry. These results demonstrate that event structure provides an effective organizing principle for advancing fMRI reconstruction from isolated clips toward coherent long videos.

1 INTRODUCTION

Reconstructing dynamic visual scenes from neural activity has long been a central goal of neuroscience (Naselaris et al., 2011; Nishimoto et al., 2011). Recent advances in multimodal learning (Radford et al., 2021) and generative modeling (Rombach et al., 2022), together with their growing integration into neuroscience, have enabled substantial progress in reconstructing high-fidelity visual stimuli from high-dimensional neural measurements such as functional magnetic resonance imaging (fMRI) (Scotti et al., 2023; Chen et al., 2023; Gong et al., 2024). Nevertheless, major challenges remain, as the organization of semantic and motion information during continuous viewing is still not fully understood.

Early fMRI-based video reconstruction methods directly mapped neural signals into pixel space (Nishimoto et al., 2011; Han et al., 2019; Wen et al., 2018), often producing semantically ambiguous videos with conspicuous visual artifacts. Pretrained diffusion models (Rombach et al., 2022; Guo et al., 2023) have since markedly improved semantic clarity and visual quality (Chen et al., 2023; Gong et al., 2024). Yet recent methods (e.g., MinD-Video (Chen et al., 2023), NeuroClips (Gong et al., 2024) and NEURONS (Wang et al., 2025)) still commonly partition continuous visual experience into independent short clips that are decoded and evaluated separately. For example, in NeuroClips (Gong et al., 2024), continuous test videos are divided into fixed temporal windows and paired with their corresponding neural measurements. However, in cognitive neuroscience community, event segmentation (Zacks et al., 2007; 2010; Baldassano et al., 2017) is a well-established principle. Continuous experience is spontaneously organized into a sequence of events, with cortical activity remaining relatively stable within an event and undergoing concentrated state transitions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

at event boundaries (Baldassano et al., 2017). That means that the imposed clip boundaries are generally misaligned with the intrinsic boundaries of visual events.

Moreover, the temporal dispersion of the hemodynamic response causes the BOLD signal associated with each clip to integrate neural activity over several surrounding seconds (Boynton et al., 1996). Multiple clips within the same event therefore provide repeated noisy observations of a shared semantic state, while clips near a boundary may contain information from adjacent events. Existing decoding paradigm neither exploits complementary evidence within an event nor isolates semantic mixing near its boundaries, leading to content drift across clips in long-video reconstruction. Therefore, we argue that effective reconstruction of continuous visual experiences requires an event-aware perspective that respects the intrinsic temporal organization of visual content.

To this end, We propose CortexLink, an event-aligned framework for reconstructing long videos, which treat each shot as a unit of continuous geometric modeling, extracting a shared subject template from its spatial observations and estimating a pose trajectory that extends across clip boundaries. Event structure consequently determines both the range over which semantic evidence is aggregated and the interval over which subject motion remains continuous. For high-level semantics, this within-event stability implies that multiple clips provide repeated noisy measurements of the same latent semantic state, giving cross-clip aggregation a clear statistical basis. In contrast, prior work (Gong et al., 2024; Wang et al., 2025; Shen et al., 2026) typically treats the sampling grid itself as the decoding unit, overlooking the correspondence between stimulus event structure and the organization of neural responses. This omission aggravates both semantic drift and discontinuous subject motion across clips.

Technically, CortexLink reconstructs long videos from fMRI through two complementary forms of information. As illustrated in Figure 1(a), fixed sampling windows can either remain within an event or straddle an event boundary, yet conventional pipelines decode every clip independently. This design discards repeated evidence within an event and amplifies local uncertainty into semantic drift and spatial discontinuities. CortexLink instead elevates the unit of inference from the sampling grid to the visual event, as shown in Figure 1(b). On the semantic side, independently decoded descriptions within an event are treated as noisy observations of a shared latent meaning, and Minimum Bayes Risk (MBR) decoding (Eikema & Aziz, 2020) distills them into an event-level semantic anchor. On the geometric side, frame-wise masks are jointly decomposed into a shared subject template and time-varying pose, from which a continuous trajectory is estimated over the complete shot. The semantic anchor and continuous geometry then jointly guide shot-level video synthesis. Event boundaries provide their common scope, allowing evidence and trajectories to be integrated within a shot and reinitialized at shot transitions. Our main contributions are threefold.

- We introduce an event-aligned decoding paradigm that identifies the mismatch between independent clip decoding and continuous visual events. By treating event structure as an inference constraint, it aggregates repeated observations within events and prevents information propagation across boundaries, improving semantic consistency and subject-motion continuity in long-video reconstruction.
- We propose event-level semantic consensus decoding, which treats clip descriptions within an event as repeated observations of a shared semantic state and aggregates them into an event-level semantic anchor using MBR. Without modifying the upstream decoder, it improves semantic accuracy and cross-clip consistency.
- We develop event-level geometric motion modeling that combines shot-level spatial aggregation during training with shape–pose decomposition during inference. It recovers a shared subject template and continuous pose trajectory from multiple event observations, improving subject localization and reducing motion discontinuities without test-time ground truth or additional inference networks.

2 Related Work

Static image reconstruction. Early studies primarily learned direct mappings from fMRI signals to pixels or low-dimensional visual features (Miyawaki et al., 2008; Fang et al., 2020), often producing semantically ambiguous reconstructions with conspicuous artifacts. With the development of CLIP (Radford et al., 2021) and diffusion models (Rombach et al., 2022), contemporary approaches

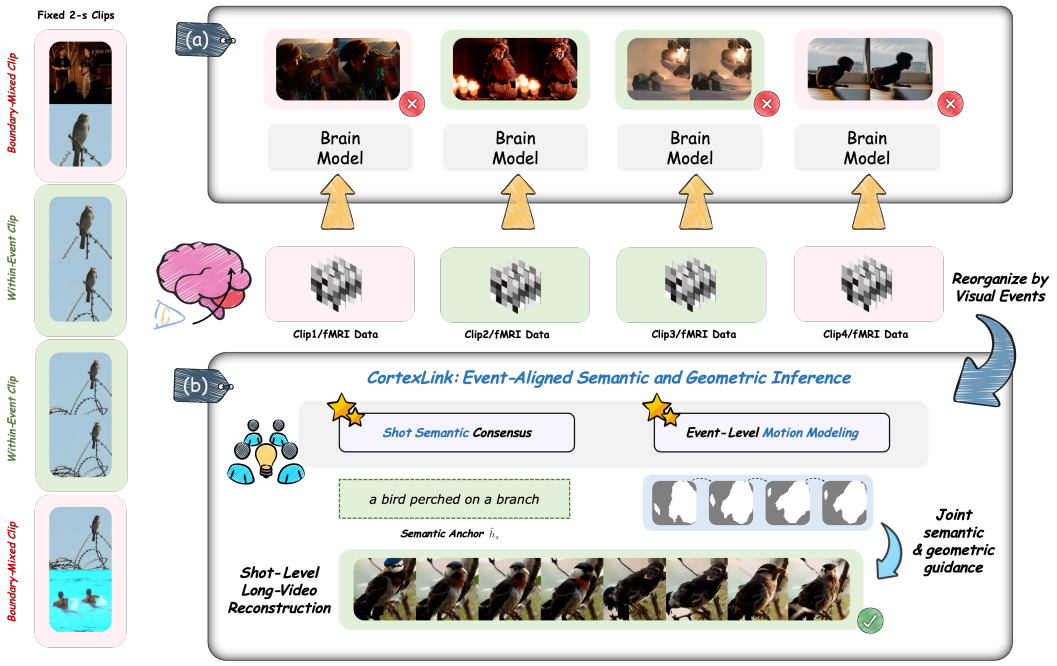

Figure 1: **From independent clips to event-aligned long-video reconstruction.** (a) Fixed sampling windows may fall within a visual event or mix content across an event boundary. Decoding each clip independently fails to exploit repeated within-event evidence and produces semantic drift and spatial discontinuities. (b) CortexLink reorganizes clip-level predictions by visual event, distills an event-level semantic anchor from candidate descriptions, and recovers continuous geometry from frame-wise masks. The two anchors jointly guide shot-level long-video synthesis. Visual shots serve as the operational approximation to events in our implementation.

increasingly align fMRI representations with image or text embeddings that encode rich semantic and spatial information, and subsequently recover visual content using pretrained generative models (Scotti et al., 2023; 2024). This paradigm has substantially improved static-image reconstruction at both semantic and pixel levels.

Dynamic video reconstruction. Video reconstruction additionally requires recovering dynamic appearance and motion. Existing methods mainly follow two directions. The first directly predicts latent representations or conditioning signals for video generation from fMRI. MinD-Video (Chen et al., 2023) aligns neural representations with CLIP space to drive a temporal diffusion model, while Mind-Animator (Lu et al., 2024) and CineNeuron (Wei et al., 2026) further introduce disentangled semantic, structural, motion, and multimodal priors. These methods reduce the gap between neural signals and generation conditions. However, due to the low SNR of fMRI (Wei et al., 2026), appearance and motion details remain underconstrained, making reconstruction highly dependent on pretrained generative priors and limiting low-level fidelity and stability.

The second direction, represented by NeuroClips (Gong et al., 2024), first reconstructs interpretable intermediate visual representations and then conditions video generation. NeuroClips combines semantic keyframes with blurry video streams to balance semantics and temporal smoothness. NEURONS (Wang et al., 2025) improves this framework with hierarchical subtasks and decoupled auxiliary losses, while CortexVideo (Shen et al., 2026) further introduces direct semantic decoding and spatially aware supervision. By providing explicit appearance and spatiotemporal constraints, this paradigm achieves more controllable reconstruction and motivates our framework.

Nevertheless, existing approaches remain based on independent fixed-length clips. Although blurry video streams provide temporal information, they entangle motion and appearance in a low-resolution representation, preventing explicit motion measurement and long-term composition. Instead of adding another clip-level branch, we address two key questions: how to aggregate com-

plementary evidence across fMRI observations within the same event, and how to organize local dynamics into a consistent and composable long-term motion trajectory.

Continuous event modeling. MindShot (Zeng et al., 2025) predicts shot boundaries in fMRI sequences elicited by multi-shot videos and decodes keyframe descriptions from the separated shot components, demonstrating the importance of isolating signals mixed across boundaries. NeuroWorld (Dong et al., 2026) instead recurrently predicts future fMRI responses driven by continuous stimuli in a latent neural-state space, showing that long-range transitions in neural activity during natural viewing can be modeled. MindShot focuses on cross-shot decomposition and text-based recovery of individual shots, whereas NeuroWorld focuses on predicting neural activity rather than reconstructing visual stimuli. Neither approach exploits the repeated evidence provided by multiple fMRI observations within the same event. Studies in cognitive neuroscience indicate that cortical activity remains relatively stable within an event during continuous narrative viewing and changes collectively at event boundaries (Hasson et al., 2008; Zacks et al., 2010; Baldassano et al., 2017), while the hemodynamic response causes adjacent measurements to share neural activity over several surrounding seconds (Boynton et al., 1996). Observations within an event are therefore suitable for aggregation, whereas mixed responses near boundaries should be isolated. Although shot boundaries do not coincide exactly with cognitive event boundaries (Magliano & Zacks, 2011), they provide a stable and detectable approximation to visual events. Inspired by Minimum Bayes Risk decoding (Eikema & Aziz, 2020) and self-consistency (Wang et al., 2022), we treat multiple fMRI decoding outputs within the same event as repeated noisy observations of shared semantics and aggregate them into an event-level consensus.

3 METHOD

Overview of CortexLink. To recover long-video conditions with coherent semantics and continuous subject locations, CortexLink first organizes a continuous stimulus into visual events and regards the neural responses within each event as complementary observations of shared latent content. It then constructs two conditions: a semantic anchor that captures event-level content consistency and provides a stable textual condition for video generation, and a geometric anchor that represents shared subject shape together with its time-varying pose to support continuous spatial control across clips. As summarized in Figure 2, the two branches share the same event scope while aggregating complementary semantic and geometric information.

Specifically, CortexLink builds on CortexVideo (Shen et al., 2026) as its clip-level video reconstruction framework. Let E_f, E_f^{ASD}, and D_{sem} denote its spatial (dorsal-stream) fMRI encoder, semantic (ventral-stream) fMRI encoder, and text decoder, respectively. For the i-th neural response,

$$\mathbf{Z}_i = E_f(\mathbf{x}_i), \qquad h_i = D_{\mathrm{sem}}\big(E_f^{\mathrm{ASD}}(\mathbf{x}_i)\big)\,. \tag{1}$$

The visual conditioning representation $\mathbf{Z}_i$ is used to reconstruct the keyframe, blurry video stream, and subject masks, while h_i is the clip description generated directly from fMRI. CortexLink builds event-level inference on top of these predictions, organizing otherwise independent clip outputs into shared semantic and continuous spatial conditions.

Let a continuous stimulus contain N video clips $\mathbf{v}_i$ and their corresponding fMRI responses $\mathbf{x}_i$, with each clip comprising F frames. A clip-level neural decoder predicts a candidate description h_i and a subject-mask sequence $\mathbf{M}_i = \{\mathbf{M}_{i,f}\}_{f=1}^{F}$ from $\mathbf{x}_i$. A stimulus-side shot partition $\mathcal{S} = \{[a_s, b_s)\}_{s=1}^{S}$ then reorganizes these independent outputs into visual events. For each shot s, the semantic branch selects a consensus description $\hat{h}_s$ from the candidate set, while the geometric branch decomposes all predicted masks into a shared shape template $\mathbf{T}_s$ and a continuous pose trajectory $\boldsymbol{\pi}_s(t)$. The consensus description serves as a shot-level text prompt, and the template and trajectory jointly define the spatial condition for continuous video generation within the shot.

CortexLink operates at both training and inference time. During training, an event-aligned MaskBank provides the segmentation decoder with a low-variance spatial prior, encouraging stable subject localization within each shot. During inference, all mask observations in a shot jointly participate in shape–pose decomposition and cross-clip trajectory estimation. This formulation reduces the need to regress precise frame-wise motion directly from noisy neural measurements and introduces no additional trainable network at inference time.

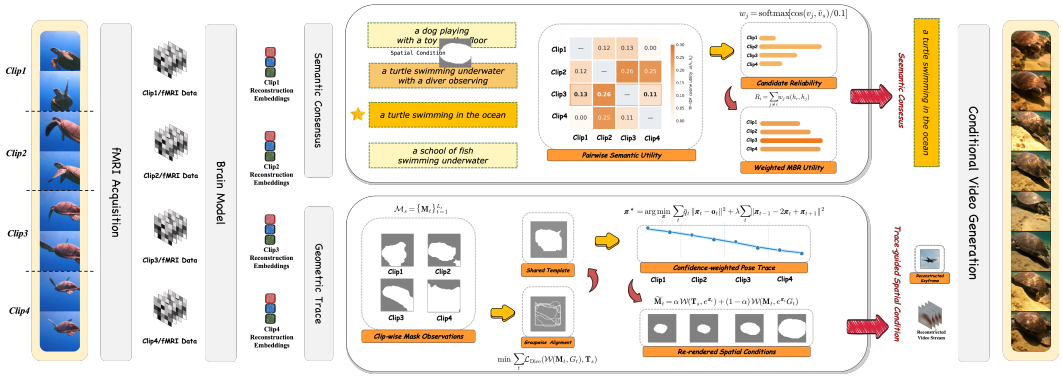

Figure 2: **Overview of the CortexLink framework.** Given a visual event comprising consecutive clips, a shared brain model maps each clip-wise fMRI response to reconstruction embeddings. (i) The *semantic consensus* branch treats clip descriptions as noisy observations of a shared event, computes pairwise semantic utilities and candidate reliability, and uses weighted Minimum Bayes Risk decoding to select an event-level semantic anchor. (ii) The *geometric trace* branch jointly aligns clip-wise mask observations, aggregates a shared subject template, estimates a confidence-weighted pose trajectory across clip boundaries, and re-renders continuous spatial conditions. (iii) The semantic anchor and trace-guided spatial conditions are jointly injected into the conditional video generator to reconstruct a coherent long video. The illustrated example contains four consecutive clips from the Subject 3 turtle shot.

Semantic Branch and Event-Level Consensus Decoding. A shot typically corresponds to a sustained visual event in which adjacent clips share the principal subjects, scene, and action context. Local neural noise and the temporal mixing induced by the hemodynamic response can nevertheless cause clip-wise descriptions to drift in subject, attributes, or actions. This effect is particularly pronounced near shot transitions, where the slow hemodynamic response causes nearby neural measurements to contain information from both preceding and subsequent visual events (Boynton et al., 1996). Our semantic branch adopts the semantic decoder from the ventral stream of CortexVideo (Shen et al., 2026) to generate one candidate description for each clip. The candidate set $\mathcal{H}_s$ within shot s is then treated as a collection of noisy observations of an underlying event-level semantic state $y_s^\star$. We use Minimum Bayes Risk (MBR) decoding (Eikema & Aziz, 2020) to select the description that is most consistent with the overall semantics of the shot,

$$\hat{h}_s = \arg\max_{h \in \mathcal{H}_s} \sum_{h' \in \mathcal{H}_s \setminus \{h\}} w(h')\, u(h, h'). \tag{2}$$

We define the semantic utility $u(h, h')$ as the cosine similarity between TF–IDF representations (Salton & Buckley, 1988). TF–IDF downweights generic words that recur across candidate descriptions and assigns greater importance to terms associated with the specific subjects, actions, and scenes of the current event. The resulting utility therefore favors candidates that receive broad support within the shot while retaining event-specific information.

Candidate descriptions may have different levels of reliability. We accordingly introduce adaptive candidate weights. Let v_h denote the TF–IDF representation of description h, and let $\bar{v}_s$ be the centroid of the candidate representations in shot s. We define

$$w(h) \propto \exp\left(\frac{\cos(v_h, \bar{v}_s)}{\tau}\right), \qquad \sum_{h \in \mathcal{H}_s} w(h) = 1. \tag{3}$$

These weights increase the contribution of candidates that agree with the global semantics of the shot and reduce the influence of semantic mixing near boundaries and occasional decoding errors. When a shot contains fewer than $K_{\min}$ candidates, the available evidence is insufficient to support a reliable event-level consensus, and we retain the original clip-wise descriptions. For all remaining shots, the selected description $\hat{h}_s$ is shared across their constituent clips as the event-level semantic anchor for subsequent video generation.

Geometric Branch with a Shared Subject and Continuous Pose. An event-level geometric representation must capture both the temporal consistency of subject shape and the dynamic evolution of its spatial state. We therefore decompose each frame-wise mask into a shot-level shared shape template and a time-varying pose, explicitly modeling stable subject structure and continuous positional change. During training, MaskBank aggregates spatial evidence within each shot to provide the segmentation decoder with a stable localization prior. During inference, shape–pose decomposition connects adjacent clips and establishes continuous subject motion across clip boundaries.

Event-aligned MaskBank. During the second training stage, MaskBank stores the most recent mean prediction for each clip, $\bar{\mathbf{M}}_i = F^{-1} \sum_f \mathbf{M}_{i,f}$. When constructing a prior for clip i, the bank reads only the other clips from the same shot, $\mathcal{N}(i) = \{j \neq i \mid s(j) = s(i)\}$. Directly averaging these masks would convolve within-shot positional variation with subject shape, producing a blurred and enlarged prior. We instead compute the centroid $\mathbf{c}_j$ of each mask and translate it to the common centroid $\bar{\mathbf{c}}_{-i}$ of its shot mates. The aligned shot prior is then

$$\mathbf{P}_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathcal{W}\big(\bar{\mathbf{M}}_j, \bar{\mathbf{c}}_{-i} - \mathbf{c}_j\big)\,, \tag{4}$$

where $\mathcal{W}$ denotes spatial warping. The prior $\mathbf{P}_i$ aggregates repeated spatial evidence from the same shot and provides the segmentation decoder with a stable reference for subject location, thereby suppressing random localization deviations across clips. MaskBank primarily improves within-shot spatial consistency; cross-clip motion continuity is subsequently established by the trajectory model.

Groupwise shape–pose decomposition. At inference time, we jointly register all L_s masks in a shot, $\mathcal{M}_s = \{\mathbf{M}_t\}_{t=1}^{L_s}$. For each frame, we estimate a similarity transform $G_t \in \mathrm{Sim}(2)$ that aligns its mask with a shared template $\mathbf{T}_s$ (Gower, 1975),

$$\min_{\{G_t\},\, \mathbf{T}_s} \sum_{t=1}^{L_s} \mathcal{L}_{\mathrm{Dice}}\left(\mathcal{W}(\mathbf{M}_t, G_t), \mathbf{T}_s\right). \tag{5}$$

This optimization maps all masks in a shot into a common coordinate system, allowing the shared template to aggregate subject shape that remains stable over time. Let a_t indicate whether a mask is valid. The template is obtained by averaging all valid aligned observations,

$$\mathbf{T}_s = \mathrm{Avg}_{t:a_t=1}\, \mathcal{W}(\mathbf{M}_t, G_t). \tag{6}$$

We further define the reliability of each pose observation by its soft Dice agreement (Tilborghs et al., 2022) with the template,

$$q_t = a_t\ \mathrm{Dice}\left(\mathcal{W}(\mathbf{M}_t, G_t), \mathbf{T}_s\right). \tag{7}$$

Empty masks therefore do not contribute to template construction, while observations that disagree with the shared shape receive lower weights during trajectory estimation.

Groupwise registration has a gauge ambiguity, because jointly transforming all masks and the template leaves the objective unchanged. We fix this degree of freedom by removing the mean Lie-algebra transform within each shot, placing $\mathbf{T}_s$ in the mean coordinate system predicted for the shot. We then define the observed pose of the template at frame t as $\mathbf{o}_t = -\log G_t$. This decomposition separates shape variation from positional variation in the original frame-wise masks. The template aggregates recurring subject structure within an event, while the pose sequence records the geometric change of each frame relative to that template.

Continuous pose trajectory. Pose observations from independently decoded clips remain noisy and may introduce implausible jumps at clip boundaries. To preserve the underlying motion trend while establishing cross-clip continuity, we jointly estimate a pose trajectory $\boldsymbol{\pi}_t$ over each shot. Its four components represent log-scale, rotation, and two translation coordinates in the Lie algebra. Because these channels have different units, we standardize each channel by the standard deviation of its within-shot observations before optimization. To simplify notation, we continue to use $\mathbf{o}_t$ and $\boldsymbol{\pi}_t$ for the standardized observations and trajectory below, and define $\bar{q}_t = \max(q_t, \epsilon)$. The trajectory is estimated by

$$\boldsymbol{\pi}^\star = \arg\min_{\{\boldsymbol{\pi}_t\}} \sum_{t=1}^{L_s} \bar{q}_t \|\boldsymbol{\pi}_t - \mathbf{o}_t\|_2^2 + \lambda \sum_{t=2}^{L_s-1} \|\boldsymbol{\pi}_{t-1} - 2\boldsymbol{\pi}_t + \boldsymbol{\pi}_{t+1}\|_2^2. \tag{8}$$

After optimization, we restore the trajectory to its original pose scale and continue to use $\boldsymbol{\pi}_t$ for the restored trajectory during rendering. The objective balances reliable registration observations against a smooth motion trend. Its second-order temporal regularizer (Eilers, 2003) preserves constant-velocity motion while suppressing abnormal jumps caused by independent decoding. Because the trajectory is estimated over the complete timeline of a shot, adjacent clips share a coherent motion state, while the trajectory naturally resets at true shot boundaries.

Continuous geometric rendering. The shared template and continuous trajectory jointly define the geometric anchor at frame t. We place the template at its smoothed pose and, to preserve frame-specific shape evidence, align the original mask with G_t before placing it back at the same pose. The final re-rendered mask is

$$\widetilde{\mathbf{M}}_t = \alpha \mathcal{W}(\mathbf{T}_s, \exp(\boldsymbol{\pi}_t)) + (1-\alpha)\mathcal{W}(\mathbf{M}_t, \exp(\boldsymbol{\pi}_t) G_t)\,. \tag{9}$$

Here, α balances event-level shape consensus against frame-specific detail. The resulting geometric anchor is generated by moving a shared subject template along a continuous pose trajectory. It therefore provides a sequence of time-varying spatial conditions that preserves positional changes unavailable to a static broadcast mask.

Dual-Anchor Training and Long-Video Inference. CortexLink follows a staged neural decoding framework (Shen et al., 2026). The first stage learns a shared neural representation aligned with visual and textual spaces, and the second recovers semantic descriptions, subject masks, categories, and blurry videos as conditions for generation. Building on these representations, the event-aligned MaskBank connects clips from the same shot in the spatial branch, allowing subject segmentation to exploit shared spatial evidence within the event. Semantic consensus and continuous trajectories further organize the clip-level predictions at the shot level.

Under the continuous testing protocol of CC2017 (Wen et al., 2018), NeuroClips (Gong et al., 2024) pairs fMRI responses from fixed temporal windows with corresponding video clips and decodes them independently. CortexLink preserves this correspondence but reorganizes adjacent clips according to shot boundaries, transforming fixed sampling units into visual events with shared content and continuous motion. Each shot interval therefore provides a temporal unit for aggregating semantic evidence and modeling subject trajectories.

At inference time, the clip-level decoder first recovers text, keyframes, blurry videos, and subject masks from an ordered fMRI sequence. The semantic branch uses MBR to distill a shot-level description $\hat{h}_s$, providing consistent content guidance throughout the event. The geometric branch performs groupwise registration and trajectory estimation to produce re-rendered masks $\widetilde{\mathbf{M}}_t$. These masks constrain the blurry video stream frame by frame and are aggregated into a spatial condition for keyframe reconstruction. The semantic and geometric anchors then jointly condition the original video generator, coordinating subject content, shape, and location along the shot timeline and connecting independent short clips into a semantically coherent and spatially continuous long video.

4 Experiments

Dataset and Protocol. We evaluate CortexLink on the public CC2017 dataset (Wen et al., 2018), following the standard data split and preprocessing protocol used by prior fMRI-to-video studies (Gong et al., 2024; Shen et al., 2026). CC2017 contains fMRI responses from three subjects viewing natural videos, acquired with a 3-T MRI scanner at a temporal resolution of 2 seconds. The training and test stimuli comprise 18×8-minute and 5×8-minute videos, respectively. Each subject viewed the training stimuli twice and the test stimuli ten times; repeated test responses are averaged to improve the signal-to-noise ratio. Pairing each 2-second fMRI sample with six video frames yields 8,640 training samples and 1,200 test samples. For long-video evaluation, we retain the original temporal order of the test stimuli and partition the 1,200 clips into 597 non-overlapping shots using stimulus-side shot detection. All event-level operations are restricted to frames within the same shot. The original 2-second clips remain the input unit for all methods, ensuring that CortexLink and the clip-level baselines use identical fMRI–video correspondences.

Evaluation Metrics. We evaluate reconstructed videos from semantic, pixel, and spatiotemporal perspectives. Following prior fMRI-to-video studies (Gong et al., 2024; Wang et al., 2025), semantic



Figure 3: **Qualitative comparison on shots.** Within each four-row group, the rows show the ground truth, NEURONS, CortexVideo, and CortexLink from top to bottom, where CortexVideo is the strongest baseline. Four frames are sampled from each consecutive clip. CortexLink produces more stable subject appearance, scale, and placement across clip boundaries.

fidelity is measured by 2-way and 50-way identification accuracy using features from a pretrained VideoMAE classifier (Tong et al., 2022). Pixel-level fidelity is evaluated with structural similarity (SSIM) (Wang et al., 2004) and peak signal-to-noise ratio (PSNR) (Huynh-Thu & Ghanbari, 2008). For temporal dynamics, ΔDD measures the absolute difference in Dynamic Degree (Huang et al., 2024) between a reconstruction and its ground-truth video. Cross-clip continuity is measured by Boundary Seam (cos), the ratio between the mean CLIP-feature distance (Radford et al., 2021) across clip boundaries and that between adjacent frames within clips. Only boundaries inside the same ground-truth shot are evaluated, thereby excluding natural discontinuities caused by true scene cuts. Lower values of ΔDD and Boundary Seam indicate better dynamic agreement and temporal continuity, respectively.

Qualitative Results. Figure 3 compares complete-shot reconstructions. In the livestock example, NEURONS and CortexVideo change categories and scenes across clips, whereas CortexLink consistently preserves a herd in grassland with similar counts, scales, and layouts. In the cat close-up, the baselines vary identity, appearance, orientation, and framing, while CortexLink maintains the subject and composition with gradual within-shot changes. These examples show that semantic consensus suppresses content drift and geometric tracing stabilizes subject scale and placement across clip boundaries.

Quantitative Results. Table 1 reports duration-stratified averages over three subjects. On 6–8 s shots, CortexLink improves 50-way identification over CortexVideo from 0.284 to 0.312, while reducing ΔDD from 0.101 to 0.073 and Boundary Seam from 6.577 to 2.075. On shots of at least 10 s, it ranks first on five of six metrics: 50-way identification increases from 0.322 to 0.372, ΔDD falls from 0.082 to 0.040, and Boundary Seam from 5.933 to 1.939, with concurrent SSIM and PSNR gains. The larger semantic gain on longer shots and the consistent seam reduction support event-level evidence aggregation and continuous geometric modeling.

Component Ablation. We ablate the two core event-level components to examine their respective contributions. The semantic comparison evaluates whether event-level consensus improves semantic

Table 1: **Quantitative comparison of long-video reconstruction performance on CC2017.** Results are averaged across three subjects and reported by shot duration. The best and second-best results are shown in **bold** and underlined, respectively.

	Semantic-level		Spatiotemporal-level		Pixel-level	
METHODS	2-way ↑	50-way ↑	ΔDD ↓	seam (cos) ↓	SSIM ↑	PSNR ↑
6–8 s shots (3–4 clips)						
NeuroClips	0.846±0.004	0.223±0.020	0.550±0.094	8.947±0.938	**0.395**±0.004	9.473±0.049
NEURONS	0.869±0.015	0.275±0.015	0.201±0.052	7.346±0.628	0.375±0.008	**9.647**±0.065
CortexVideo	0.876±0.002	0.284±0.017	0.101±0.040	6.577±1.079	0.363±0.013	9.504±0.135
CortexLink	**0.881**±0.013	**0.312**±0.016	**0.073**±0.036	**2.075**±0.103	0.381±0.016	9.534±0.041
10 s+ shots (5+ clips)						
NeuroClips	0.869±0.013	0.267±0.017	0.505±0.066	8.411±0.159	0.419±0.015	9.359±0.289
NEURONS	0.887±0.012	0.319±0.026	0.120±0.032	6.648±0.658	0.400±0.019	9.498±0.280
CortexVideo	0.894±0.013	0.322±0.019	0.082±0.075	5.933±0.278	0.399±0.022	9.399±0.302
CortexLink	**0.902**±0.009	**0.372**±0.039	**0.040**±0.051	**1.939**±0.045	**0.420**±0.014	**9.671**±0.074

Table 2: **Component ablation on Subject 1.** Semantic scores are reported on long shots where event-level consensus is applied, while the Trace columns compare mask predictions before and after continuous trajectory modeling.

	Semantic caption				Trace mask	
METHOD	BLEU-1 ↑	BLEU-4 ↑	ROUGE-L ↑	CIDEr-D ↑	IoU ↑	Mask Seam ↓
Baseline	0.2530	0.0476	0.2460	0.7738	0.2369	1.67
CortexLink	**0.2576**	**0.0586**	**0.2607**	**0.9713**	**0.2419**	**1.03**

recovery, while the Trace comparison evaluates how continuous geometric modeling affects subject localization and motion continuity across clip boundaries.

Event-level semantic consensus improves all four caption metrics within its operating range. The largest gains occur in BLEU-4, which increases from 0.0476 to 0.0586, and CIDEr-D, which increases from 0.7738 to 0.9713, corresponding to relative improvements of 23.1% and 25.5%, respectively. The simultaneous gains in n-gram overlap and TF–IDF-weighted semantic similarity indicate that aggregating repeated within-event evidence selects descriptions that better preserve both linguistic structure and discriminative content.

The Trace module raises mask IoU from 0.2369 to 0.2419 and reduces Mask Seam from 1.67 to 1.03, corresponding to a 38.3% relative reduction in cross-clip displacement discontinuity. The simultaneous improvement in overlap and boundary continuity indicates that the gain does not arise from merely smoothing the trajectory at the expense of spatial localization. Instead, groupwise registration consolidates repeated shape evidence, while the continuous pose model regularizes its evolution across clip boundaries.

5 CONCLUSION

We presented CortexLink, an event-aligned framework that extends fMRI-to-video reconstruction from independent clips to continuous visual events. CortexLink treats clip-level descriptions within a shot as noisy observations and aggregates them into a semantic consensus, while groupwise registration separates a shared subject template from its evolving pose trajectory. By modeling stable content and dynamic spatial states separately, CortexLink improves semantic consistency and suppresses discontinuities in subject identity, scale, and location across clip boundaries. Rather than introducing another clip-level branch, CortexLink exploits cross-clip structure in continuous stimuli to organize distributed evidence into event-level semantic and geometric representations, providing a framework for coherent and controllable long-video brain decoding.

AI USE STATEMENT

We used generative AI tools, including large language models accessed through our development and writing environment, to aid and polish writing, draft and revise portions of the manuscript and figure captions, retrieve and organize relevant literature, and support research ideation and execution. The latter included assistance with code implementation and debugging, experiment and evaluation workflow checking, interpretation of experimental results, figure preparation, and Chinese-to-English translation. We did not use generative AI to generate synthetic datasets or establish mathematical claims. All AI-assisted outputs were treated as suggestions: the authors manually verified literature references, equations, quantitative claims, code, experimental commands, figures, and translations against the source materials and observed results. The authors reviewed and edited the final manuscript in full and take responsibility for all text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Geoffrey M Boynton, Stephen A Engel, Gary H Glover, and David J Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *The journal of neuroscience*, 16(13): 4207–4221, 1996.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21938–21948, 2025.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36:24841–24858, 2023.

Zijian Dong, Jianxiong Zhou, Kwun Kei Ng, Jan Paolo Macapinlac Balagtas, Zhizhou Li, Zijiao Chen, and Juan Helen Zhou. Neuroworld: A latent brain world model for stimulus-conditioned human brain dynamics. *arXiv preprint arXiv:2608.01773*, 2026.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Bryan Eikema and Wilker Aziz. Is map decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, 2020.

Paul HC Eilers. A perfect smoother. *Analytical chemistry*, 75(14):3631–3636, 2003.

Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33:13038–13048, 2020.

Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37:51655–51683, 2024.

John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *The Journal of neuroscience*, 28(10):2539–2550, 2008.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):1, 2008.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Sungnyun Kim, Gihun Lee, Sangmin Bae, and Se-Young Yun. Mixco: Mix-up contrastive learning for visual representation. *arXiv preprint arXiv:2010.06300*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5):5, 2017.

Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, Xujin Li, and Huiguang He. Animate your thoughts: Decoupled reconstruction of dynamic natural vision from slow brain activity. *arXiv preprint arXiv:2405.03280*, 2024.

Joseph P Magliano and Jeffrey M Zacks. The impact of continuity editing in narrative film on event segmentation. *Cognitive science*, 35(8):1489–1517, 2011.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current biology*, 21(19):1641–1646, 2011.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Xiaoquan Shen, Jiajun Li, Jiaxuan Chen, and Gang Pan. Cortexvideo: A semantic-spatial dual-anchor framework for high-fidelity fmri-to-video reconstruction. In *European Conference on Computer Vision*, pp. 676–693. Springer, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pp. 369–386. SPIE, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Sofie Tilborghs, Jeroen Bertels, David Robben, Dirk Vandermeulen, and Frederik Maes. The dice loss in the context of missing or empty labels: introducing ϕ and ϵ. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 527–537. Springer, 2022.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. *arXiv preprint arXiv:2503.11167*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yujie Wei, Chenglong Ma, Jianxiong Gao, Chenhui Wang, Shiwei Zhang, Biao Gong, Shuai Tan, Hangjie Yuan, and Hongming Shan. Bridging brain and semantics: A hierarchical framework for semantically enhanced fmri-to-video reconstruction. *arXiv preprint arXiv:2605.14569*, 2026.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12): 4136–4160, 2018.

Rafael Yuste, Sara Goering, Blaise Agüera Y Arcas, Guoqiang Bi, Jose M Carmena, Adrian Carter, Joseph J Fins, Phoebe Friesen, Jack Gallant, Jane E Huggins, et al. Four ethical priorities for neurotechnologies and ai. *Nature*, 551(7679):159–163, 2017.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, and Corey J Maley. The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in human neuroscience*, 4:168, 2010.

Wenwen Zeng, Yonghuang Wu, Yifan Chen, Xuan Xie, Chengqian Zhao, Feiyu Yin, Guoqing Wu, and Jinhua Yu. Mindshot: Multi-shot video reconstruction from fmri with llm decoding. *arXiv preprint arXiv:2508.02480*, 2025.

Appendix overview. Appendix A makes the inherited CortexVideo training objective explicit. Appendix B details the event partition and the semantic and geometric procedures introduced by CortexLink. Appendix C describes CC2017 preprocessing, Appendix D provides the complete optimization and inference settings, and Appendix F concludes with limitations and ethical considerations.

A Model and Training Details

CortexLink builds on the clip-level neural decoder of CortexVideo (Shen et al., 2026) and organizes its outputs through event-level semantic and geometric modeling. This section presents the objectives of the semantic encoder, backbone, and auxiliary decoding heads. Event-level semantic consensus, groupwise registration, and trajectory fitting subsequently operate on the clip predictions, while MaskBank supplies a shot-conditioned prior to the segmentation decoder during second-stage training.

A.1 Semantic Encoder

The semantic encoder E_f^{ASD} maps an fMRI sample X to a dense representation y. Following CortexVideo (Shen et al., 2026; Chen et al., 2025), the representation is aligned with a sparse set of stimulus patches selected through differentiable attention. If x_i is the CLIP (Radford et al., 2021) feature of patch i and f_{MLP} projects it into the fMRI embedding space, the Gumbel-perturbed patch distribution (Jang et al., 2016) is

$$p_i = \frac{\exp\big(f_{\mathrm{MLP}}(x_i)^\top y + \gamma_i\big)}{\sum_{j=1}^{N} \exp(f_{\mathrm{MLP}}(x_j)^\top y + \gamma_j)}, \qquad \gamma_i \sim \mathrm{Gumbel}(0,1). \tag{10}$$

We apply straight-through top-K selection (Bengio et al., 2013) to p to obtain $p^\star$. The adaptive alignment loss is

$$\mathcal{L}_{\mathrm{adaptive}} = \left\| y - \frac{1}{K} \sum_{i=1}^{N} p_i^\star f_{\mathrm{MLP}}(x_i) \right\|_2^2. \tag{11}$$

A global pre-alignment term prevents the representation from collapsing onto a few local patches,

$$\mathcal{L}_{\mathrm{prealign}} = \left\| E_f^{\mathrm{ASD}}(X_{\mathrm{sem}}) - x_{[\mathrm{CLS}]} \right\|_2^2, \tag{12}$$

where $x_{[\mathrm{CLS}]}$ is the global CLIP visual feature. A pretrained autoregressive decoder D_{sem}, connected to y through trainable cross-attention, is optimized with token negative log likelihood,

$$\mathcal{L}_{\mathrm{sem}} = -\mathbb{E}_{(X,\mathcal{T})\sim\mathcal{D}} \sum_{m=1}^{M} \log p(t_m \mid t_{<m}, F_{\mathrm{CA}}(y)). \tag{13}$$

The complete semantic objective is

$$\mathcal{L}_{\mathrm{ASD}} = w_{\mathrm{adapt}}\mathcal{L}_{\mathrm{adaptive}} + w_{\mathrm{prealign}}\mathcal{L}_{\mathrm{prealign}} + w_{\mathrm{sem}}\mathcal{L}_{\mathrm{sem}}. \tag{14}$$

The resulting h_i is a clip-level semantic description decoded directly from fMRI and becomes one candidate in the event-level consensus set.

A.2 Backbone and Auxiliary Objectives

The main encoder E_f aligns fMRI with visual and textual CLIP spaces using BiMixCo (Scotti et al., 2023). For a mini-batch sample x_i, MixCo (Kim et al., 2020) constructs

$$x_i^\star = \lambda_i x_i + (1-\lambda_i) x_{m_i}, \qquad \lambda_i \sim \mathrm{Beta}(\beta,\beta), \tag{15}$$

where m_i is a randomly permuted batch index. Let $z_i = E_f(x_i^\star)$, v_i be the corresponding visual target, and

$$s(a,b) = \frac{\cos(a,b)}{\tau}. \tag{16}$$

BiMixCo includes both fMRI-to-visual and visual-to-fMRI retrieval:

$$\mathcal{L}_{\mathrm{CLIPv}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\lambda_i \log\frac{\exp(s(z_i,v_i))}{\sum_{k=1}^{B}\exp(s(z_i,v_k))} + (1-\lambda_i)\log\frac{\exp(s(z_i,v_{m_i}))}{\sum_{k=1}^{B}\exp(s(z_i,v_k))}\right] - \frac{1}{2B}\sum_{j=1}^{B}\left[\lambda_j \log\frac{\exp(s(z_j,v_j))}{\sum_{k=1}^{B}\exp(s(z_k,v_j))} + \sum_{l:m_l=j}(1-\lambda_l)\log\frac{\exp(s(z_l,v_j))}{\sum_{k=1}^{B}\exp(s(z_k,v_j))}\right]. \tag{17}$$

The first line treats each mixed fMRI embedding as a query with two weighted visual positives. The second line reverses retrieval: each visual target receives its original sample and every mixed sample containing it as weighted positives. Let z_i^t denote the projected fMRI textual representation and t_i the CLIP text target. The text alignment objective is the same bidirectional loss,

$$\mathcal{L}_{\mathrm{CLIPt}} = \mathcal{L}_{\mathrm{CLIPv}}|_{z_i \leftarrow z_i^t,\ v_i \leftarrow t_i}. \tag{18}$$

The aligned representation supports the same reconstruction heads as CortexVideo (Shen et al., 2026). A diffusion prior predicts (Ramesh et al., 2022; Scotti et al., 2023) the CLIP image embedding F_{GT} required by the keyframe decoder,

$$\mathcal{L}_{\mathrm{prior}} = \mathbb{E}_{\epsilon}\left\|\mathrm{PriorNet}(F_{\mathrm{vis}},\epsilon) - F_{\mathrm{GT}}\right\|_2^2. \tag{19}$$

The subject segmentation decoder is trained with soft Dice loss (Tilborghs et al., 2022),

$$\mathcal{L}_{\mathrm{seg}} = 1 - \frac{2\sum_i \hat{y}_i y_i + \epsilon}{\sum_i \hat{y}_i + \sum_i y_i + \epsilon}, \tag{20}$$

where $\hat{y}_i = \sigma(a_i)$ for segmentation logit a_i, and the sum covers all pixels of the predicted and target masks. The remaining auxiliary objectives are

$$\mathcal{L}_{\mathrm{cat}} = -\frac{1}{BC}\sum_{b=1}^{B}\sum_{c=1}^{C}\left[y_{bc}^{\mathrm{cat}}\log\sigma(r_{bc}) + (1-y_{bc}^{\mathrm{cat}})\log(1-\sigma(r_{bc}))\right], \tag{21}$$

$$\mathcal{L}_{\mathrm{txt}} = \mathrm{CE}_{\varepsilon=0.1}\left(\{\ell_m\}_{m=1}^{|S|-1}, \{s_{m+1}\}_{m=1}^{|S|-1}\right), \tag{22}$$

$$\mathcal{L}_{\mathrm{mot}} = \frac{1}{F}\sum_{f=1}^{F}\left\|\widehat{Y}_{\mathrm{blur},f} - Y'_f\right\|_1. \tag{23}$$

Here r_{bc} is the logit for concept c in sample b, ℓ_m is the next-token logit at position m, padding tokens are ignored, and $\varepsilon = 0.1$ is the label-smoothing coefficient (Szegedy et al., 2016). The three losses supervise multi-label concept recognition, scene-description decoding, and reconstruction of the blurry video in the VAE latent space, respectively.

The spatial perception head predicts a continuous subject heatmap rather than only a binary mask (Shen et al., 2026). Given DINO patch features (Siméoni et al., 2025) $\Phi^{(j)}$ and a grounded subject mask M (Ren et al., 2024), the subject feature and heatmap target are

$$F_{\mathrm{subj}} = \frac{\sum_j M_j \Phi^{(j)}}{\sum_j M_j}, \qquad H_{\mathrm{GT}}^{(k)} = \frac{F_{\mathrm{subj}}\cdot\Phi^{(k)}}{\|F_{\mathrm{subj}}\|\,\|\Phi^{(k)}\|}. \tag{24}$$

The corresponding spatial loss is

$$\mathcal{L}_{\mathrm{spat}} = \mathbb{E}_{(X,I)\sim\mathcal{D}}\left\|\sigma(H_{\mathrm{pred}}) - H_{\mathrm{GT}}\right\|_2^2. \tag{25}$$

A.3 Integrated Training Objective

Training follows the two-stage schedule of CortexVideo (Shen et al., 2026). Pre-alignment establishes the visual, textual, and spatial representation:

$$\mathcal{L}_{\mathrm{stage1}} = w_{\mathrm{CLIPv}}\mathcal{L}_{\mathrm{CLIPv}} + w_{\mathrm{CLIPt}}\mathcal{L}_{\mathrm{CLIPt}} + w_{\mathrm{spat}}\mathcal{L}_{\mathrm{spat}}. \tag{26}$$

The second stage jointly trains the reconstruction heads,

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{stage1}} + w_{\text{prior}}\mathcal{L}_{\text{prior}} + w_{\text{seg}}\mathcal{L}_{\text{seg}} + w_{\text{cat}}\mathcal{L}_{\text{cat}} + w_{\text{mot}}\mathcal{L}_{\text{mot}} + w_{\text{txt}}\mathcal{L}_{\text{txt}}. \tag{27}$$

Following CortexVideo, the auxiliary tasks use staggered boost windows. Let A_q be the first optimization step and T_q the duration of the window for task q. Its weight is

$$w_q(c) = \begin{cases} 1 + 9\left|\sin\left(\dfrac{c - A_q}{T_q}\pi\right)\right|, & A_q \le c < A_q + T_q, \\ 1, & \text{otherwise.} \end{cases} \tag{28}$$

The windows for concept recognition, text decoding, and blurry-video reconstruction start half a window apart. In the reported MaskBank training, $w_{\text{seg}} = 1$ throughout, while the other three auxiliary terms retain their staggered schedules. The alignment, spatial, and prior objectives use their fixed coefficients. During this stage, MaskBank stores the detached mean mask most recently predicted for every training clip. For clip i, the other clips from the same shot form a leave-one-out neighborhood; their masks are centroid-aligned and averaged to construct $\mathbf{P}_i$, as defined in Sec. **??**. The encoded prior is supplied to the segmentation decoder. Bank entries are updated with detached predictions, and single-clip shots or cold-start entries use a zero prior.

B Event Construction and Long-Video Inference

B.1 Stimulus-Side Shot Partition

CortexLink uses editing shots as a reproducible approximation to visual events and estimates their boundaries directly from the stimulus frames presented to the subjects. For adjacent frames, we compute an HSV content difference s_t and normalize it by the median of a local neighborhood excluding the current pair,

$$r_t = \frac{s_t}{\text{median}_{j \in \mathcal{W}_t \setminus t}\, s_j + \epsilon}. \tag{29}$$

A frame at which both s_t and r_t exceed their thresholds is declared a cut. We use an absolute threshold of 22.0, an adaptive threshold of 2.5, a six-frame half-window, and a minimum separation of three frames. Detection is performed independently inside each acquisition run, with run boundaries included as shot boundaries. A clip containing a cut is assigned by majority duration, and the resulting frame-level intervals are converted into 597 non-overlapping shot intervals over the 1,200 test clips. The complete partition is derived from stimulus pixels.

For a shot $s = [a_s, b_s)$, the semantic and geometric observations are

$$\mathcal{H}_s = \{h_i \mid a_s \le i < b_s\}, \qquad \mathcal{M}_s = \{\mathbf{M}_{i,f} \mid a_s \le i < b_s,\, 1 \le f \le F\}. \tag{30}$$

The partition defines event-level neighborhoods for evidence sharing while retaining the standard fMRI–video pairing and evaluation protocol.

B.2 Semantic Consensus

The MBR objective in Sec. **??** performs an event-level medoid decision over $\mathcal{H}_s$. Captions are lowercased and tokenized, and TF–IDF vectors are constructed over the candidate set. Pairwise cosine similarity provides the utility matrix. Candidate reliability is obtained by a softmax over similarity to the shot centroid with temperature $\tau = 0.1$. We apply consensus to shots containing at least three clips; shorter shots retain their original clip-wise descriptions. Ties are resolved deterministically by the earliest candidate index.

The decision $\hat{h}_s \in \mathcal{H}_s$ becomes the shared semantic anchor of the shot. Uniform weights yield the standard semantic medoid, while random within-shot broadcasting provides a zero-information control for the sharing mechanism. Together, these controls separate the contribution of candidate weighting from that of event-level sharing.

B.3 Geometric Optimization

MaskBank uses translation alignment during training because it is inexpensive and robust to early noisy predictions. At inference, all valid masks in a shot are registered jointly with four-degree-of-freedom $\mathrm{Sim}(2)$ transforms. Empty masks are excluded from template construction, and soft Dice agreement with the template determines the reliability q_t. Registration is initialized from identity with centroid translation, focusing subsequent iterations on scale, rotation, and residual translation.

The registration objective has a gauge ambiguity: applying a common transform to every observation and the template leaves their relative alignment unchanged. We place the template in the mean predicted coordinate system by enforcing

$$\frac{1}{L_s}\sum_{t=1}^{L_s}\log G_t = 0 \tag{31}$$

after each optimization update. This convention places the template in the mean predicted coordinate system and provides a common pose reference for all frames.

Let $\mathbf{o}$ stack the standardized pose observations, $W = \mathrm{diag}(\bar{q}_t)$, and D_2 be the second-difference operator (Eilers, 2003). The trajectory objective in the main paper is the quadratic system

$$\left(W + \lambda D_2^\top D_2\right)\boldsymbol{\pi} = W\mathbf{o}. \tag{32}$$

We solve this banded system independently for log-scale, rotation, and the two translation coordinates, then restore each channel to its original scale. The held configuration uses $\lambda = 300$. Re-rendering blends the shared template with registered frame-specific evidence using $\alpha = 0.85$. Templates and trajectories are estimated independently for each shot and reset at every detected cut.

B.4 Complete Inference Flow

Inference begins with the CortexVideo decoder(Shen et al., 2026), which maps each ordered fMRI sample to a description, keyframe condition, blurry video, category prediction, and subject-mask sequence. CortexLink groups these outputs by the precomputed shot partition. The semantic branch supplies the selected event prompt when the minimum-length criterion is satisfied. In parallel, the geometric branch registers all valid masks, solves one continuous trajectory over the shot, and re-renders a frame-wise spatial condition. The event prompt, re-rendered mask, decoded keyframe, and blurry stream jointly condition the pretrained video generator. Generation remains clip-based for computational compatibility, while adjacent clips inside one shot share semantic and geometric states.

C Data Processing Details

We use the publicly released CC2017 data (Wen et al., 2018) preprocessed with the minimal pipeline adopted by NeuroClips and CortexVideo (Gong et al., 2024; Shen et al., 2026). Preprocessing includes artifact removal, six-degree-of-freedom motion correction, registration to MNI space, and projection to a common cortical surface. Stimulus-responsive voxels are selected by measuring voxel-wise response correlations across repeated training movies, applying a Fisher z-transform, and evaluating the mean response over the 18 training segments with a one-sample test and Bonferroni correction at $p < 0.05$. This procedure yields 13,447, 14,828, and 9,114 selected visual-cortex voxels for Subjects 1–3, respectively. A 4-second temporal shift is applied to account for hemodynamic response delay.

The original videos are downsampled from 30 to 3 frames per second for neural-decoder training. Each 2-second fMRI sample is paired with six ordered frames. Test responses are averaged over ten repeated presentations before decoding. Shot detection is applied after restoring the original temporal ordering inside each run, with clip sampling and the train/test split following the standard CC2017 protocol.

D Implementation Details

Clip-level training. The main backbone is optimized with AdamW (Loshchilov et al., 2017) and a OneCycle learning-rate schedule (Smith & Topin, 2019) with peak learning rate 5×10^{-5}. Pre-alignment runs for 30 epochs with batch size 120; enhanced joint training runs for 50 epochs with batch size 10. The semantic encoder uses a 24-layer fMRI ViT (Dosovitskiy, 2020), a 256-dimensional projection MLP, frozen CLIP visual and GPT-2 (Radford et al., 2019) modules, and trainable reduced-dimensional cross-attention. It is optimized with Adam (Kingma & Ba, 2014) at learning rate 10^{-5} and weight decay 10^{-4}.

Event-level processing. The shot partition is shared across subjects because all participants viewed the same ordered stimuli, while caption candidates, masks, templates, and trajectories are computed independently for each subject. Mask predictions are thresholded at 0.5 when binary support is required; soft probabilities are retained for registration and blending. Groupwise registration runs for 300 optimization steps with learning rate 0.05, a transform bound of 1.0, and a minimum valid-mask area of four pixels. At most 2,048 frames are registered in one block. All semantic and geometric hyperparameters are fixed across subjects. Subject-level metrics are computed first and then aggregated with equal subject weights.

Video synthesis. Following CortexVideo (Shen et al., 2026), we use AnimateDiff (Guo et al., 2023) as the pretrained generator and perform 25 DDIM steps (Song et al., 2020). The decoded blurry stream is interpolated to 8 frames per second for final synthesis. CortexVideo and CortexLink use the same generation settings and fixed random seeds.

Computational resources. The inherited neural decoders are trained on an NVIDIA H100 GPU. Semantic consensus and trajectory fitting are deterministic computations, with groupwise registration accounting for most of the additional cost. Its complexity scales linearly with the number of masks and the fixed number of optimization iterations, and shots are processed independently.

E More Experimental Results

E.1 Additional Mask and Video Results

Additional mask visualizations follow the complete geometric pipeline. For each shot, we display decoded masks, registered observations, the shared template, raw pose measurements, the smoothed trajectory, and the re-rendered sequence. This ordering reveals whether a successful case arises from repeated shape evidence, corrected placement, or trajectory regularization. A fixed selection rule covers best, median, and difficult shots.

Figure 4 extends the qualitative comparison to human subjects, natural scenes, and animals. Across the four shots, both clip-level baselines frequently produce plausible individual clips but change category, identity, or scene composition at the boundaries. In the first shot, for example, NEURONS alternates between different people and indoor scenes, while CortexVideo changes clothing, viewpoint, and facial appearance. CortexLink instead maintains one female subject with a stable frontal composition throughout the shot. In the river sequence, it preserves a consistent water-dominated scene and coherent flow layout, whereas the baselines alternate among underwater imagery, nocturnal scenes, and visually distinct landscapes.

The turtle shot provides the clearest example of event-level semantic stabilization. The ground truth repeatedly depicts sea turtles, but the baselines drift to hikers, dogs, cattle, fish, and sharks across clips. CortexLink reconstructs turtles throughout the complete shot while preserving similar scale and underwater context. In the interview shot, this improvement is further reflected in subject consistency. CortexLink maintains one face, close-up scale, and viewing direction, whereas the baselines switch among multiple identities and compositions. The reconstructed identities and appearance do not always exactly match the ground truth, particularly in the human examples, but the markedly reduced cross-clip drift demonstrates that event-level consensus and continuous geometry turn independently generated clips into a more coherent long-video sequence.

Figure 5 isolates the effect of geometric trace refinement before video synthesis. In the bird sequence (top), the decoded masks are severely enlarged and vary markedly across clips. Trace refinement consolidates them into a stable support, reducing centroid dispersion from 6.5 to 4.4 pixels and

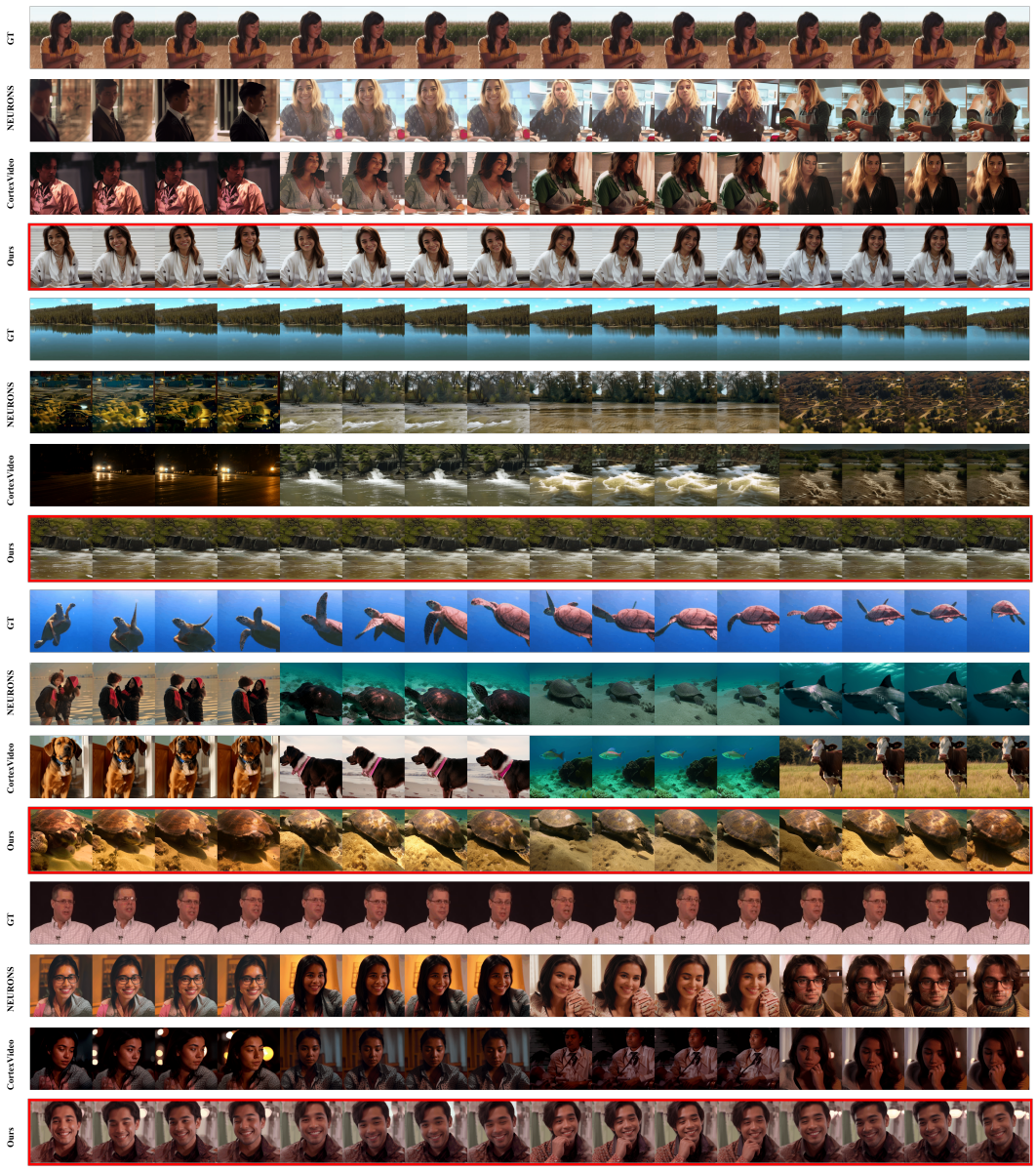

Figure 4: **Additional qualitative comparisons on four complete shots.** Each four-row group shows the ground truth, NEURONS, CortexVideo, and CortexLink from top to bottom. Four frames are sampled from each consecutive clip, and the red boxes mark CortexLink results.

slightly increasing IoU from 0.12 to 0.14. The centroid offset increases marginally from 15.9 to 16.3 pixels, however, showing that temporal regularization cannot correct a systematically misplaced or poorly shaped input. In the articulated human sequence (middle), refinement improves IoU from 0.35 to 0.41 and reduces centroid offset from 5.8 to 4.4 pixels. It removes large clip-dependent support changes, although the shared template smooths away limbs and other high-frequency articulation.

The animal sequence (bottom) exhibits the clearest joint improvement. Its independently decoded masks vary substantially in scale and position, yielding 0.27 IoU, an 11.1-pixel offset, and a 15.0-pixel dispersion. The continuous trace raises IoU to 0.36 while reducing the offset and dispersion to 8.7 and 10.1 pixels, respectively. Across all three cases, the method is most effective at stabiliz-

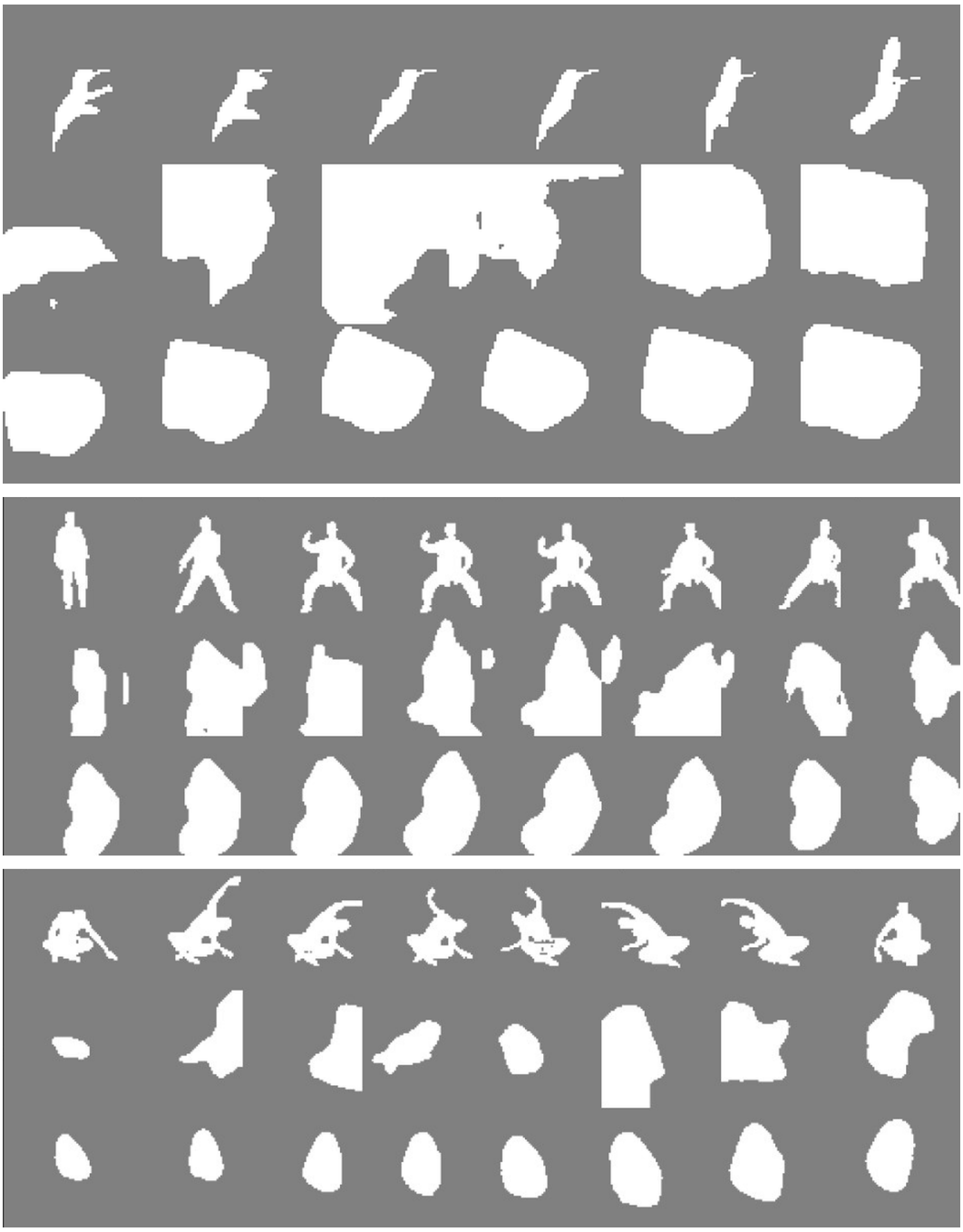

Figure 5: **Additional qualitative results for geometric trace refinement.** Each panel shows consecutive clips from one complete shot. Rows correspond to the ground-truth masks, independently decoded masks, and trace-refined masks, respectively; red boxes highlight the refined outputs. The annotations report mask IoU, centroid offset in pixels (*off*), and within-shot centroid RMS dispersion (*spr*).

ing foreground support and recovering low-frequency positional evolution. Fine contours that are absent from the decoded observations cannot be recreated by registration and trajectory smoothing, which explains the compact shapes in the refined rows and motivates retaining frame-specific mask evidence during re-rendering.

F LIMITATIONS AND ETHICS

Limitations. Editing shots are stable and reproducible but only approximate cognitive events. A semantic event may span several shots, while an uncut shot may contain a genuine change in action or subject (Magliano & Zacks, 2011). The geometric pipeline inherits the detail ceiling of its decoded masks: alignment and averaging stabilize observed contours, while fine structure depends on the input predictions. The constant-velocity preference of second-order smoothing may attenuate high-frequency motion, and the final effect on video remains dependent on how strongly the pretrained generator follows spatial conditions. Finally, the three-subject evaluation provides descriptive evidence; larger cohorts are required to establish population-level generalization.

Ethics. This study uses the publicly released, anonymized CC2017 dataset (Wen et al., 2018) under its intended research protocol. Current reconstruction relies on controlled scanning, subject-specific training, repeated measurements, and generative priors, and its outputs are model-conditioned visual reconstructions. Neural data remain sensitive; future systems require explicit consent, transparent data governance, and safeguards against involuntary decoding (Yuste et al., 2017). Generated videos may also inherit biases or hallucinations from pretrained vision–language and diffusion models, so interpretation should account jointly for neural evidence, generative priors, and model uncertainty.