ENTINEURO: MULTI-ENTITY VISUAL DECODING AND VIDEO RECONSTRUCTION FROM EEG SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural visual scenes often contain multiple semantically distinct entities whose information must be represented simultaneously during perception. However, existing EEG-based visual decoding studies largely focus on a single dominant semantic target or overall scene content, and available EEG–video datasets do not explicitly control multi-entity compositions. Consequently, whether multiple entity semantics can be decoded from EEG and further support video reconstruction remains largely unexplored. To address this gap, we introduce **EntiNeuro**, a multi-entity visual decoding dataset with 624 dynamic videos across eight predefined cross-category compositions and synchronized EEG and eye-tracking recordings from 20 participants over three sessions. Based on EntiNeuro, we establish **EntiBench** for evaluating entity recognition and video reconstruction in multi-entity scenes. We further propose **EntiRecon**, a hierarchical EEG-based reconstruction framework that combines coarse entity semantics with fine-grained EEG conditioning using eye tracking as auxiliary attentional guidance. Experiments consistently show that multi-entity semantic information can be decoded from EEG. EntiRecon further improves semantic and multi-entity fidelity over existing EEG-to-video reconstruction methods. These results demonstrate the potential of EEG representations for reconstructing dynamic scenes containing multiple semantic entities.

1 INTRODUCTION

Visual decoding from non-invasive brain activity provides a promising avenue for understanding how visual information is represented in the human brain and for developing brain–computer interfaces (Lu et al., 2025; Liu et al., 2025; Huang et al., 2026b; Li et al., 2026). With recent advances in representation learning and generative models (Radford et al., 2021; Rombach et al., 2022), EEG-based visual decoding has progressed from category recognition and retrieval (Song et al., 2024; Li et al., 2024b) to image and video reconstruction (Dai et al., 2026; Zheng et al., 2026; Liu et al., 2024; Huang et al., 2026a). However, existing EEG-based video decoding and reconstruction studies largely focus on a single dominant semantic target (Liu et al., 2024; Huang et al., 2025; Zhou et al., 2026a), whereas natural dynamic scenes often contain multiple semantically distinct entities. Decoding such scenes requires recovering both individual entities and their composition from the same EEG response, a problem that remains largely unexplored. In this work, we use *multi-entity* to refer to scenes containing two or more salient entities from distinct semantic categories, rather than multiple instances of the same category.

These observations motivate two key research questions: **RQ1:** whether multi-entity semantic information can be decoded from EEG responses; and **RQ2:** whether such neural representations can further support the reconstruction of dynamic visual scenes. Answering these questions requires controlled multi-entity stimuli with explicitly defined cross-category compositions. Existing EEG–video datasets such as SEED-DV (Liu et al., 2024) and CineBrain (Gao et al., 2025) are not designed for this purpose: SEED-DV primarily consists of short videos centered on a single semantic entity or visual concept, whereas CineBrain uses long-form naturalistic audiovisual narratives without explicit control over multi-entity compositions. Consequently, neither provides a controlled setting for systematically studying EEG representations of predefined multi-entity compositions.

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.

To address this gap, we introduce **EntiNeuro**, a multi-entity visual decoding dataset comprising 624 dynamic videos organized into eight predefined compositions across six entity categories. Twenty participants viewed the complete stimulus set across three repeated sessions, with EEG and eye-tracking signals recorded synchronously. Building on EntiNeuro, we establish **EntiBench** with two complementary tasks: entity recognition evaluates whether multi-entity semantic information can be decoded from EEG alone, while video reconstruction examines whether such neural representations can support the recovery of dynamic multi-entity scenes. These two tasks directly address RQ1 and RQ2, respectively.

We further introduce **EntiRecon**, a hierarchical EEG-based framework for multi-entity video reconstruction. Directly mapping noisy and limited EEG data to the high-dimensional text-conditioning space of a video diffusion model is challenging (Huang et al., 2025; Zhou et al., 2026a). EntiRecon instead first decodes entity semantics from EEG to construct a compact coarse semantic anchor, and then learns a fine-grained EEG representation with synchronized eye tracking as auxiliary attentional guidance. The resulting coarse and fine representations jointly condition a pretrained video diffusion model. Experiments on EntiBench show that multi-entity semantic information can be decoded from EEG under both session-average and cross-session settings, while EntiRecon improves the preservation of entity composition and semantic consistency in reconstructed videos.

Our main contributions are threefold: (1) we introduce **EntiNeuro**, a multi-entity visual decoding dataset with synchronized EEG and eye-tracking recordings; (2) we establish **EntiBench**, a benchmark for multi-entity visual decoding from EEG through entity recognition and video reconstruction; and (3) we propose **EntiRecon**, a hierarchical EEG-based reconstruction framework that combines coarse entity semantics with fine-grained neural conditioning.

2 EntiNeuro: Multi-Entity Visual Decoding Dataset

2.1 Stimulus Construction and Validation

We constructed an initial pool of 800 candidate videos spanning eight predefined entity compositions across six categories: *person, dog, bird, flower, car, and ball*. The dataset includes six two-entity and two three-entity compositions. The main benchmark focuses on two-entity scenes, while the three-entity stimuli are reserved for generalization experiments. Two-entity videos last 4 s, whereas three-entity videos last 6 s to accommodate the greater scene complexity and provide participants sufficient time to observe all entities and their interactions.

Before data collection, three independent raters evaluated all candidate videos for entity suitability and interaction naturalness using a 5-point Likert scale. The top 78 videos from each composition were retained, yielding 624 final stimuli with mean ratings of 4.33 ± 0.30 and 3.93 ± 0.47. Participants also rated the final stimulus set after the experiment, with scores of 4.07 ± 0.95 for overall realism and 3.65 ± 1.16 for interaction naturalness. These evaluations support the overall quality of the selected stimuli. Additional details and stimulus examples are provided in Appendix B.

2.2 Subjects and Experimental Protocol

Twenty healthy students participated in the experiment (mean age: 22.8 years), all with normal or corrected-to-normal vision. All participants provided written informed consent, and the study protocol was approved by the institutional ethics committee. Before recording, participants were briefed on the procedure and instructed to attend to the videos while minimizing unnecessary head and body movements.

Data collection was conducted in a controlled laboratory environment to minimize environmental noise and recording artifacts. EEG signals were recorded using an ESI NeuroScan System with a 62-channel active AgCl cap arranged according to the international 10–10 system at 1000 Hz. Eye movements were simultaneously recorded using a Tobii Pro Fusion eye tracker at 250 Hz. Video stimuli were presented full-screen at a resolution of 1920×1080 (16:9) on a 25-inch display. Each participant completed three recording sessions separated by at least one day. In each session, all 624 videos were presented once, with the presentation order independently randomized for that session. As illustrated in Fig. 1, each session comprised 24 mini-blocks of 26 videos, with a 15-s rest period between consecutive mini-blocks to reduce fatigue. To maintain attention, 63 multiple-choice

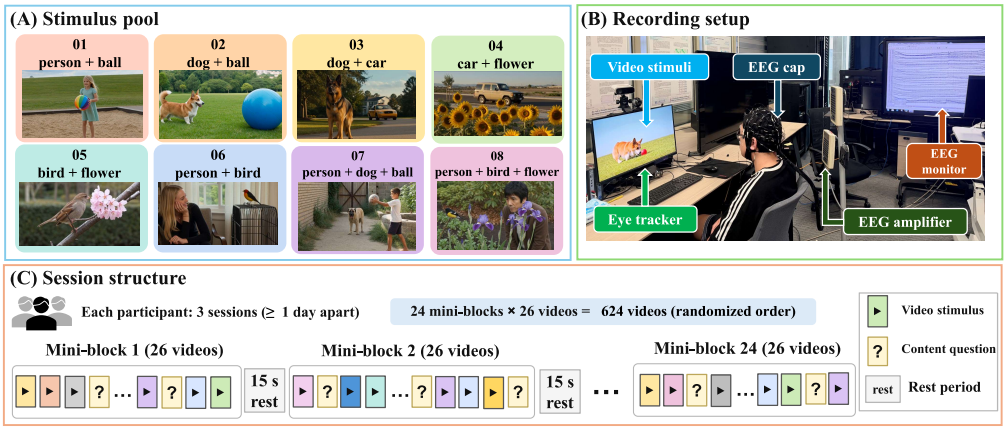


Figure 1: **Overview of the experimental setup and protocol.** (A) Stimulus pool with eight predefined entity compositions. (B) EEG and eye-tracking recording setup. (C) Session structure with randomized video presentation, interspersed content questions, and rest periods.

questions were approximately evenly distributed throughout each session, each referring to the immediately preceding video. Across all participants and sessions, the mean response accuracy was $92.25\% \pm 2.71\%$, indicating that participants remained generally attentive throughout the experiment.

2.3 Data Preprocessing

EEG signals were band-pass filtered at 0.1–100 Hz, downsampled from 1000 to 200 Hz, and cleaned using independent component analysis (ICA) to remove ocular artifacts. The continuous recordings were segmented by stimulus triggers into 4-s two-entity and 6-s three-entity trials. For spectral analyses, we additionally extracted PSD and DE features over the δ, θ, α, β, and γ bands. Eye-tracking recordings were aligned to individual video trials using stimulus timestamps and resampled from 250 to 200 Hz, with 49 variables retained for subsequent modeling. Further preprocessing and feature extraction details are provided in Appendix B.5.

3 Method

3.1 Problem Formulation and Overview

Given a stimulus video $\mathbf{V}_i$ presented across S repeated sessions, we denote the corresponding EEG and eye-tracking recordings as $\mathbf{X}_i^{\text{eeg}} = \{\mathbf{X}_{i,s}^{\text{eeg}}\}_{s=1}^{S}$ and $\mathbf{X}_i^{\text{eye}} = \{\mathbf{X}_{i,s}^{\text{eye}}\}_{s=1}^{S}$, where $\mathbf{X}_{i,s}^{\text{eeg}} \in \mathbb{R}^{C \times T}$ and $\mathbf{X}_{i,s}^{\text{eye}} \in \mathbb{R}^{C_e \times T}$. Our goal is to reconstruct $\mathbf{V}_i$ from EEG while using eye tracking as auxiliary attentional guidance.

As illustrated in Fig. 2, **EntiRecon** employs hierarchical conditioning for multi-entity video reconstruction. The coarse branch decodes entity semantics from repeated EEG responses to construct a semantic anchor. The fine branch uses eye-tracking representations to guide EEG fusion and refines the resulting neural tokens with the coarse semantic context. The coarse and fine conditions then jointly condition a frozen video diffusion model.

3.2 Coarse Entity Semantic Bottleneck

Directly mapping limited and noisy EEG data to the high-dimensional text-conditioning space of a video diffusion model is challenging. We therefore introduce a coarse semantic bottleneck that decodes entity composition and individual entity presence from EEG, providing a compact semantic anchor without reconstructing the full text-conditioning representation.

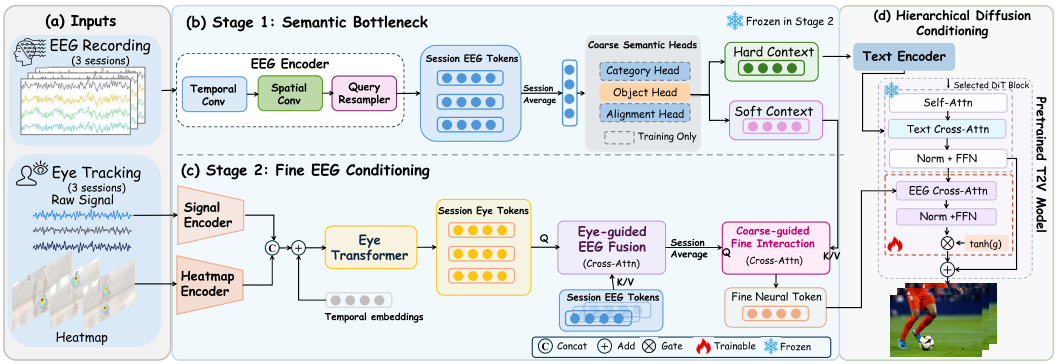


Figure 2: **Overview of EntiRecon.** The coarse branch extracts entity-level semantics from repeated EEG responses, while the fine branch captures fine-grained semantic information through eye-guided EEG fusion. The hierarchical representations jointly condition video reconstruction.

Repeated-session EEG encoding. For each session, a shared EEG encoder independently maps the input signal to latent tokens:

$$\mathbf{H}_{i,s} = E_{\text{eeg}}\left(\mathbf{X}_{i,s}^{\text{eeg}}\right) \in \mathbb{R}^{M\times d}. \tag{1}$$

The encoder applies temporal and spatial convolutions followed by a learnable query resampler, producing M tokens of dimension d. To reduce session-specific variability, we average the representations across repeated sessions and then pool over tokens:

$$\overline{\mathbf{H}}_i = \frac{1}{S}\sum_{s=1}^{S}\mathbf{H}_{i,s}, \qquad \mathbf{h}_i = \frac{1}{M}\sum_{m=1}^{M}\overline{\mathbf{H}}_{i,m}. \tag{2}$$

The pooled representation $\mathbf{h}_i$ is used for coarse semantic prediction, while the session-specific tokens $\mathbf{H}_{i,s}$ are retained for the fine-grained conditioning branch.

Entity prediction and auxiliary supervision. The pooled EEG representation $\mathbf{h}_i$ is mapped to three branches:

$$\mathbf{z}_i^{\text{cat}} = f_{\text{cat}}(\mathbf{h}_i), \qquad \mathbf{z}_i^{\text{obj}} = f_{\text{obj}}(\mathbf{h}_i), \qquad \mathbf{b}_i = \text{Norm}(f_{\text{sem}}(\mathbf{h}_i)). \tag{3}$$

Here, f_{cat} predicts the entity composition, f_{obj} performs multi-label entity prediction, and f_{sem} aligns EEG with frozen text and video representations. The coarse branch is pretrained using

$$\mathcal{L}_{\text{coarse}} = \mathcal{L}_{\text{cat}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{align}}, \tag{4}$$

combining composition classification, multi-label entity prediction, and cross-modal contrastive alignment. The composition and semantic alignment branches serve only as auxiliary supervision during pretraining and are discarded afterward, while f_{obj} is retained for reconstruction.

Semantic anchor construction. At reconstruction time, entity probabilities are obtained from the multi-label predictions:

$$\mathbf{p}_i = \sigma\left(\mathbf{z}_i^{\text{obj}}\right). \tag{5}$$

Let $\mathbf{A} \in \{0,1\}^{K\times O}$ denote the predefined composition matrix, where $A_{k,o} = 1$ indicates that entity o belongs to composition k. Each valid composition is scored according to the predicted entity probabilities:

$$r_{i,k} = \sum_{o=1}^{O}\left[A_{k,o}\log p_{i,o} + (1-A_{k,o})\log(1-p_{i,o})\right], \qquad \pi_{i,k} = \frac{\exp(r_{i,k}/\tau_c)}{\sum_{j=1}^{K}\exp(r_{i,j}/\tau_c)}. \tag{6}$$

Each composition is associated with an entity-level prompt q_k, with frozen text representation $\mathbf{T}_k = E_{\text{text}}(q_k)$. The soft context and hard prompt are constructed as

$$\mathbf{C}_i^{\text{soft}} = \sum_{k=1}^{K}\pi_{i,k}\mathbf{T}_k, \qquad q_i^{\text{hard}} = q_{\text{argmax}_k\,\pi_{i,k}}. \tag{7}$$

The soft context guides fine-grained neural conditioning, whereas the hard prompt is passed through the native text-conditioning pathway of the video diffusion model.

3.3 Eye-Guided Fine-Grained Neural Conditioning

Eye context encoding. To provide auxiliary attentional cues, both eye-tracking signals and gaze locations are encoded for each stimulus and session. Let $\mathbf{G}_{i,s} \in \mathbb{R}^{F\times 1\times H\times W}$ denote the gaze heatmaps constructed by placing Gaussian kernels at the recorded gaze locations. Temporal and spatial encoders extract

$$\mathbf{U}_{i,s} = f_{\text{sig}}(\mathbf{X}^{\text{eye}}_{i,s}), \qquad \mathbf{V}_{i,s} = f_{\text{map}}(\mathbf{G}_{i,s}), \tag{8}$$

where $\mathbf{U}_{i,s}, \mathbf{V}_{i,s} \in \mathbb{R}^{F\times d/2}$. The two representations are concatenated, after which learnable temporal embeddings are added before Transformer encoding:

$$\mathbf{Q}^{e}_{i,s} = \text{Transformer}_e\left((\mathbf{U}_{i,s} \,\|\, \mathbf{V}_{i,s}) + \mathbf{P}^{e}\right). \tag{9}$$

Here, $\|$ denotes feature concatenation and $\mathbf{P}^{e}$ denotes learnable temporal embeddings. Each session is encoded independently with shared parameters to preserve correspondence with its session-specific EEG representation.

Eye-guided EEG fusion. The resulting eye tokens $\mathbf{Q}^{e}_{i,s} \in \mathbb{R}^{M\times d}$ and the corresponding session-specific EEG tokens $\mathbf{H}_{i,s} \in \mathbb{R}^{M\times d}$ are fused through multi-head cross-attention, with the eye tokens serving as queries and the EEG tokens as keys and values:

$$\mathbf{H}^{f}_{i,s} = \text{MHA}\left(\text{LN}(\mathbf{Q}^{e}_{i,s}), \text{LN}(\mathbf{H}_{i,s}), \text{LN}(\mathbf{H}_{i,s})\right). \tag{10}$$

The resulting representations are then averaged across repeated sessions:

$$\mathbf{H}^{f}_{i} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{H}^{f}_{i,s}. \tag{11}$$

The aggregated EEG tokens $\mathbf{H}^{f}_{i}$ are subsequently passed to the coarse-guided fine interaction module.

Coarse-guided fine semantic interaction. The eye-guided EEG representation is further refined using the soft semantic context. Specifically, $\mathbf{M}^{c}_{i} = f_c(\mathbf{C}^{\text{soft}}_{i})$ projects the coarse context into the neural feature space, and $\mathbf{Z}^{(0)}_{i} = \mathbf{H}^{f}_{i} + \mathbf{P}^{f}$ initializes the fine queries with learnable embeddings $\mathbf{P}^{f}$. The queries are refined through N_f cross-attention blocks:

$$\widehat{\mathbf{Z}}^{(\ell)}_{i} = \mathbf{Z}^{(\ell)}_{i} + \text{MHA}\left(\text{LN}(\mathbf{Z}^{(\ell)}_{i}), \text{LN}(\mathbf{M}^{c}_{i}), \text{LN}(\mathbf{M}^{c}_{i})\right), \tag{12}$$

$$\mathbf{Z}^{(\ell+1)}_{i} = \widehat{\mathbf{Z}}^{(\ell)}_{i} + \text{FFN}\left(\text{LN}(\widehat{\mathbf{Z}}^{(\ell)}_{i})\right). \tag{13}$$

The final condition is $\mathbf{Z}^{\text{fine}}_{i} = \text{LN}(\mathbf{Z}^{(N_f)}_{i}) \in \mathbb{R}^{M\times d}$.

3.4 Hierarchical Conditioning of Video Diffusion

The coarse and fine representations condition the pretrained video diffusion model at complementary levels: the coarse representation provides entity semantics through the native text-conditioning pathway, while the fine representation modulates intermediate video features.

Coarse text conditioning. The hard semantic prompt q^{hard}_{i} is encoded by the frozen text encoder and passed through the native text-conditioning pathway of the video diffusion model, providing an explicit entity-level semantic anchor.

Gated fine conditioning. The fine condition $\mathbf{Z}_i^{\text{fine}} \in \mathbb{R}^{M\times d}$ is injected into selected Transformer layers through lightweight cross-attention. Let $\mathbf{X}_{i,\ell}^v \in \mathbb{R}^{N_v\times d_v}$ denote the video tokens at the ℓ-th selected layer. Using the video tokens as queries and the fine tokens as keys and values,

$$\begin{aligned}\Delta\mathbf{X}_{i,\ell}^v &= \mathrm{FFN}\big(\mathrm{Norm}\big(\mathrm{CrossAttn}\left(\mathbf{X}_{i,\ell}^v, \mathbf{Z}_i^{\text{fine}}\right)\big)\big),\\ \mathbf{X}_{i,\ell}^v &\leftarrow \mathbf{X}_{i,\ell}^v + \tanh(g_\ell)\Delta\mathbf{X}_{i,\ell}^v.\end{aligned} \tag{14}$$

The layer-specific gate g_ℓ controls the contribution of fine neural conditioning while retaining the pretrained generative prior.

3.5 Training Strategy

EntiRecon is trained in two stages. Stage 1 pretrains the coarse semantic bottleneck with $\mathcal{L}_{\text{coarse}}$, after which the entire module is frozen. Stage 2 jointly uses the coarse semantic condition and fine neural condition for reconstruction, optimizing the eye encoder, eye-guided EEG fusion, fine semantic interaction, and gated conditioning modules while keeping the pretrained video diffusion model frozen.

The Stage 2 objective combines the diffusion loss with an auxiliary semantic loss:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}, \tag{15}$$

where $\mathcal{L}_{\text{sem}}$ aligns the pooled fine representation with the frozen text embedding of the ground-truth caption. The caption is used only for auxiliary semantic supervision rather than as a generation condition. Context dropout is further applied during training to reduce reliance on the coarse semantic prior. Additional implementation and optimization details are provided in Appendix C.

4 Experiments

4.1 EntiBench: Multi-Entity Visual Decoding Benchmark

EntiBench evaluates multi-entity visual decoding through two tasks: *entity recognition* and *video reconstruction*. The main benchmark focuses on two-entity scenes, while three-entity stimuli are reserved for generalization experiments reported in Appendix G.

Entity recognition. Entity recognition is formulated as a six-class multi-label classification task under two evaluation settings. In the *session-average* setting, EEG responses to the same video are averaged across the three sessions at the signal level before decoding. In the *cross-session* setting, one session is held out for testing in turn, with performance averaged over the three splits. All models are evaluated under the same protocol for a fair comparison. Results are averaged across all 20 participants and reported using Macro/Micro AP, Macro/Micro AUROC, Top-2 F1, Top-2 Set Accuracy, and Top-2 Jaccard.

Video reconstruction. Video reconstruction is evaluated from four aspects: semantic fidelity, frame fidelity, video fidelity, and multi-entity fidelity. Semantic fidelity is measured using 2-way-I, 2-way-V, and ViFi-Score; frame fidelity uses SSIM and PSNR; video fidelity uses FVD and Warp-LPIPS; and multi-entity fidelity uses Entity F1 and Entity Count Error (ECE). We additionally conduct a five-rater human evaluation of entity consistency, temporal coherence, and overall similarity. Further benchmark details are provided in Appendix D.

4.2 Entity Recognition

4.2.1 Evaluation

We first investigate whether multi-entity semantic information can be reliably decoded from EEG signals. Representative EEG encoders are evaluated using preprocessed raw EEG under both session-average and cross-session settings. Additional feature-based and individual-session analyses are provided in Appendix E, with the main results summarized in Table 1.

Session-average evaluation. Repeated EEG responses to the same video are averaged across the three sessions at the signal level to reduce trial-specific noise. Under this setting, recognition performance exceeds chance levels across all evaluated architectures and metrics. Similar results across

Table 1: Entity recognition under session-average and cross-session settings. Results are mean ± standard deviation across 20 subjects; best results are in bold.

EEG Encoder	Macro AP ↑	Micro AP ↑	Macro AUC ↑	Micro AUC ↑	Top-2 F1 ↑	Set Acc. ↑	Jaccard ↑
Chance Level	0.3333	0.3333	0.5000	0.5000	0.3333	0.0667	0.2444
Session-Average Evaluation							
ShallowNet	0.398 ± 0.027	0.374 ± 0.026	0.545 ± 0.031	0.544 ± 0.031	0.369 ± 0.028	0.162 ± 0.028	0.300 ± 0.027
DeepNet	0.395 ± 0.029	0.371 ± 0.027	0.540 ± 0.035	0.539 ± 0.034	0.366 ± 0.032	0.161 ± 0.033	0.298 ± 0.032
EEGNet	0.386 ± 0.024	0.362 ± 0.024	0.531 ± 0.029	0.529 ± 0.029	0.356 ± 0.025	0.151 ± 0.031	0.288 ± 0.026
Conformer	**0.405 ± 0.029**	0.370 ± 0.028	0.538 ± 0.036	0.537 ± 0.036	0.364 ± 0.031	**0.175 ± 0.028**	0.301 ± 0.030
TSConv	0.394 ± 0.028	0.371 ± 0.026	0.541 ± 0.033	0.541 ± 0.032	0.369 ± 0.027	0.161 ± 0.027	0.299 ± 0.026
GLMNet	0.398 ± 0.030	**0.377 ± 0.028**	**0.546 ± 0.035**	0.547 ± 0.035	0.372 ± 0.028	0.164 ± 0.031	0.303 ± 0.029
Our Encoder	0.400 ± 0.026	0.370 ± 0.027	0.539 ± 0.032	**0.550 ± 0.030**	**0.375 ± 0.029**	0.173 ± 0.037	**0.311 ± 0.031**
Cross-Session Evaluation							
ShallowNet	0.353 ± 0.011	0.346 ± 0.011	0.516 ± 0.015	0.515 ± 0.015	0.349 ± 0.013	0.131 ± 0.022	0.276 ± 0.014
DeepNet	0.363 ± 0.022	0.353 ± 0.019	0.526 ± 0.030	0.523 ± 0.026	0.355 ± 0.023	0.156 ± 0.028	0.288 ± 0.024
EEGNet	0.352 ± 0.015	0.345 ± 0.014	0.515 ± 0.020	0.513 ± 0.018	0.345 ± 0.016	0.128 ± 0.031	0.273 ± 0.020
Conformer	0.358 ± 0.014	0.351 ± 0.014	0.522 ± 0.019	0.521 ± 0.018	0.352 ± 0.016	0.154 ± 0.020	0.286 ± 0.016
TSConv	0.356 ± 0.013	0.348 ± 0.012	0.520 ± 0.018	0.518 ± 0.017	0.350 ± 0.015	0.137 ± 0.024	0.279 ± 0.017
GLMNet	0.353 ± 0.013	0.347 ± 0.013	0.516 ± 0.018	0.516 ± 0.018	0.349 ± 0.016	0.142 ± 0.017	0.280 ± 0.015
Our Encoder	**0.372 ± 0.021**	**0.368 ± 0.021**	**0.537 ± 0.025**	**0.533 ± 0.023**	**0.362 ± 0.014**	**0.166 ± 0.025**	**0.300 ± 0.021**

substantially different EEG encoders suggest that the observed decodability is not specific to a particular model design or evaluation criterion. These findings support the presence of multi-entity semantic information in EEG responses.

Cross-session evaluation. To examine whether entity-related information remains decodable across recording sessions, we further evaluate all models under the cross-session setting. Performance is generally lower than in the session-average setting, reflecting the increased difficulty of generalizing to an unseen recording session. Nevertheless, all evaluated encoders remain above chance, suggesting that multi-entity semantic information is not limited to session-specific response patterns and can partially generalize across sessions. Together, these results show that multi-entity semantic information can be decoded from EEG and remains partially robust to session-level variation.

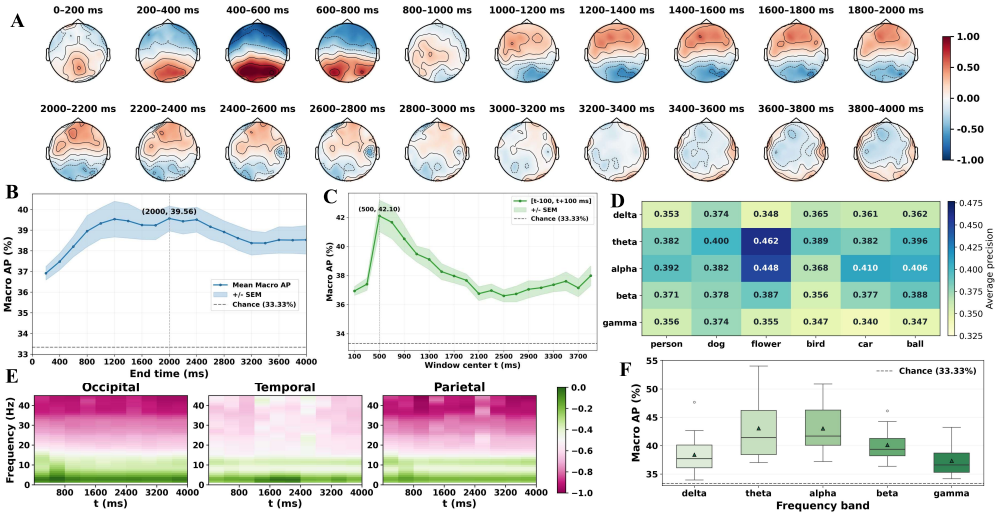


Figure 3: **Temporal, spatial, and spectral EEG analysis.** (A) Scalp topographies over time; (B–C) cumulative and local-window Macro AP; (D) class-wise frequency-band AP; (E) regional time–frequency representations; (F) subject-wise band decoding. Group averages over 20 participants are shown in (A–D); (E) shows a representative participant and (F) the across-participant distribution.

4.2.2 Temporal, Spatial, and Spectral Analysis

Temporal and spatial dynamics. The scalp topographies in Fig. 3(A) show pronounced posterior responses during the early viewing period, particularly around 400–800 ms, followed by weaker and more diffuse patterns at later stages. Consistently, cumulative decoding performance

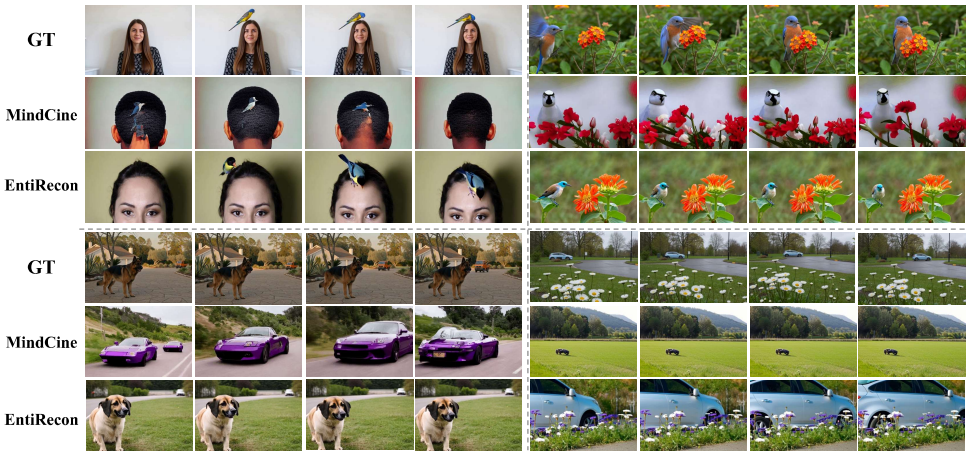


Figure 4: **Qualitative comparison of video reconstruction.** Ground-truth videos and reconstructions from MindCine and EntiRecon are shown for representative multi-entity scenes.

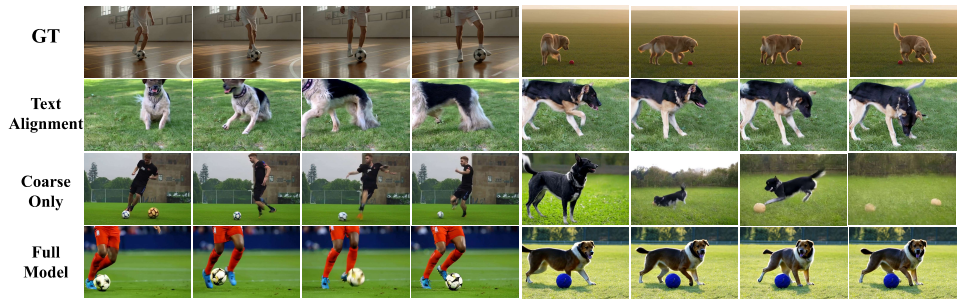


Figure 5: **Qualitative ablation results.** Comparison of Text Alignment, Coarse Only, and the full EntiRecon model on representative multi-entity scenes.

increases rapidly within the first second and gradually saturates, reaching 39.56% Macro AP at 2.0 s (Fig. 3(B)). Using fixed 200-ms local windows, performance peaks at 42.10% for the 400–600 ms interval (Fig. 3(C)). These results suggest that the early viewing period is particularly informative for entity decoding.

Spectral characteristics. The class-wise results in Fig. 3(D) show that theta or alpha achieves the highest AP across all six entity categories. The time–frequency representations in Fig. 3(E) further show stronger low-frequency activity across occipital, temporal, and parietal regions. Consistently, the subject-wise analysis in Fig. 3(F) shows relatively higher Macro AP in the theta and alpha bands than in the other frequency bands. These results indicate a consistent advantage of theta and alpha bands for entity recognition.

4.3 Video Reconstruction

4.3.1 Reconstruction Performance

As shown in Table 2, EntiRecon achieves the best performance on most semantic, video, and multi-entity metrics, while maintaining competitive frame fidelity. It obtains the highest 2-way retrieval accuracy, ViFi-Score, SSIM, Entity F1, and human ratings, together with the lowest FVD, Warp-LPIPS, and ECE. Human evaluation consistently favors EntiRecon in entity consistency, temporal naturalness and coherence, and overall similarity. GT Caption is included only as a semantic reference. Moreover, the qualitative results in Fig. 4 show a similar trend. MindCine often captures part of the target semantics but may omit or alter one of the entities, whereas EntiRecon more consistently preserves the target multi-entity composition across frames. Additional reconstruction results and failure cases are provided in Appendix F.

Table 2: **Video reconstruction results.** Best reconstruction results are in bold; GT Caption serves as a semantic reference.

(a) Automatic Evaluation

Method	Semantic Fidelity			Frame Fidelity		Video Fidelity		Multi-Entity Fidelity	
	2-way-I ↑	2-way-V ↑	ViFi ↑	SSIM ↑	PSNR ↑	FVD ↓	Warp-LPIPS ↓	Entity F1 ↑	ECE ↓
GT Caption	0.957	0.982	0.495	–	–	–	–	0.895	0.190
EEG2Video	0.757	0.832	0.258	0.184	9.78	2452.89	0.408	0.332	0.807
MindCine	0.779	0.844	0.261	0.190	10.16	2595.78	0.411	0.322	0.835
DynaMind	0.747	0.831	0.154	0.148	**11.72**	2956.18	0.458	0.221	1.069
EntiRecon	**0.830**	**0.915**	**0.338**	**0.222**	10.17	**2005.60**	**0.107**	**0.468**	**0.530**

(b) Human Evaluation

Method	# Raters	Entity Consistency ↑	Temporal Naturalness & Coherence ↑	Overall Similarity ↑
EEG2Video	5	2.16 ± 0.30	2.23 ± 0.26	2.45 ± 0.30
MindCine	5	2.28 ± 0.15	2.42 ± 0.25	2.72 ± 0.25
DynaMind	5	1.77 ± 0.31	1.91 ± 0.15	1.98 ± 0.20
EntiRecon	5	$\mathbf{3.49} \pm \mathbf{0.20}$	$\mathbf{3.00} \pm \mathbf{0.25}$	$\mathbf{3.75} \pm \mathbf{0.22}$

4.3.2 Ablation Study

Table 3 shows that direct Text Alignment performs substantially worse than coarse semantic conditioning, highlighting the challenge of directly aligning EEG with the high-dimensional text-conditioning space. Fine Only also yields limited semantic and multi-entity fidelity, emphasizing the importance of a stable coarse semantic anchor. Combining coarse and fine conditioning further improves reconstruction performance, supporting their complementary roles. Shuffling the EEG–stimulus correspondence markedly degrades semantic and multi-entity fidelity, indicating that the model benefits from stimulus-specific neural information rather than relying solely on the pretrained generative prior. The single-session variant also underperforms the full model, showing the benefit of repeated-session feature aggregation. Adding eye guidance yields further gains in semantic, video, and multi-entity fidelity, with the full model achieving the strongest overall performance. The qualitative examples in Fig. 5 show a similar trend, with the full model more consistently preserving multi-entity composition across frames.

Table 3: **Ablation study of EntiRecon.** Best results are highlighted in bold.

Method	Semantic Fidelity			Frame Fidelity	Video Fidelity	Multi-Entity Fidelity
	2-way-I ↑	2-way-V ↑	ViFi ↑	SSIM ↑	FVD ↓	ECE ↓
Text Alignment	0.744	0.820	0.184	0.174	2910.88	0.855
Coarse Only	0.815	0.907	0.317	0.211	2774.27	0.608
Fine Only	0.768	0.842	0.148	0.193	2165.77	0.814
Coarse + Fine	0.824	0.910	0.328	**0.240**	2219.83	0.561
Coarse + Fine (Shuffled EEG)	0.749	0.839	0.223	0.231	2173.75	0.815
Full (Single Session)	0.799	0.902	0.275	0.200	2286.79	0.690
Coarse + Fine + Eye (Full)	**0.830**	**0.915**	**0.338**	0.222	**2005.60**	**0.530**

5 Conclusion

We introduced EntiNeuro and EntiBench to investigate multi-entity visual decoding from EEG. Results under both session-average and cross-session settings show that multi-entity semantic information is decodable from EEG. EntiRecon further employs hierarchical coarse-to-fine conditioning to improve semantic consistency and multi-entity fidelity in video reconstruction. Overall, our work establishes a unified setting for studying EEG visual decoding beyond a single dominant semantic target in dynamic multi-entity scenes.

Limitations and future directions. Several limitations remain. The current framework does not explicitly model stimulus-specific motion, which may lead to discrepancies in reconstructed dynamics. In addition, the main benchmark centers on two-entity scenes, while generalization to three-entity compositions remains challenging. Future work could incorporate explicit motion guidance, extend to more complex multi-entity interactions, and explore richer integration of EEG and eye tracking.

AI USE STATEMENT

We used generative AI tools at several stages of this work. Seedance 2.0 was used to generate candidate video stimuli from manually designed prompts. All generated videos were subsequently evaluated by independent human raters, and only the selected videos were used for EEG and eye-tracking data collection. Qwen3-VL-32B-Instruct was used to assist in generating attention-check questions from sampled video frames, followed by automatic filtering and manual verification. We also used large language models to assist with brainstorming and providing feedback on experimental and evaluation design, discussing the interpretation of experimental results, assisting with code development and debugging, and improving the organization, clarity, and language of the manuscript. AI-assisted code was manually reviewed, executed, and tested by the authors before being used in the experiments. The final methodology, experimental protocols, analyses, and scientific claims were determined and verified by the authors. All AI-assisted text and research suggestions were manually reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study protocol was reviewed and approved by the institutional ethics committee prior to data collection. All participants volunteered to take part in the study, provided written informed consent, and received monetary compensation upon completion of the experiment. Given that this work involves human neurophysiological recordings, participant privacy and data protection were carefully considered throughout data collection, processing, and release. All EEG, eye-tracking, and associated experimental data were anonymized before analysis, and only anonymized data containing no personally identifiable information will be released. The proposed benchmark is constructed and evaluated exclusively using these anonymized data. All procedures were conducted in accordance with the approved protocol and relevant ethical requirements for research involving human participants.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the EntiNeuro dataset, data acquisition, preprocessing procedures, and experimental protocol in the main paper and Appendix. The training and evaluation splits, benchmark settings, and evaluation metrics for both entity recognition and video reconstruction are specified in the experimental sections. Architectural details, training objectives, hyperparameters, optimization settings, and hardware configurations for EntiRecon are documented in Appendix C. Additional reconstruction examples, including failure cases, are also provided in Appendix F. We will release the dataset and implementation code to facilitate reproducibility and future research.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Roman Beliy, Amit Zalcher, Jonathan Kogman, et al. Brain-it: Image reconstruction from fmri via brain-interaction transformer. In *International Conference on Learning Representations*, volume 2026, pp. 124487–124521, 2026.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. In *International Conference on Learning Representations*, volume 2024, pp. 7846–7858, 2024.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4724–4733. IEEE, 2017.

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.

Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from eeg signals. *arXiv preprint arXiv:2602.22555*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jianxiong Gao, Yichang Liu, Baofeng Yang, Jianfeng Feng, and Yanwei Fu. Cinebrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing. *arXiv preprint arXiv:2503.06940*, 2025.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Alessandro T Gifford, Pablo Oyarzo, Anne W Zonneveld, Christina Sartzetaki, Iris IA Groen, and Radoslaw M Cichy. A large dataset of human eeg responses to short naturalistic videos for studying dynamic visual event processing. *arXiv preprint arXiv:2608.28768*, 2026.

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.

Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022.

Runduo Han and Hongchen Tan. Cross-subject eeg-to-video reconstruction and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23294–23303, 2026.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Shuai Huang, Yongxiong Wang, Huan Luo, Haodong Jing, Chendong Qin, and Jingqun Tang. Mindev: Multi-modal integrated diffusion framework for video reconstruction from eeg signals. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3350–3359, 2025.

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, et al. Need: Cross-subject and cross-task generalization for video and image reconstruction from eeg signals. *Advances in Neural Information Processing Systems*, 38:173134–173173, 2026a.

Zheng Huang, Enpei Zhang, Weikang Qiu, Yinghao Cai, Carl Yang, Elynn Chen, Xiang Zhang, Rex Ying, Dawei Zhou, and Yujun Yan. Seeing through the brain: New insights from decoding visual stimuli with fmri. In *International Conference on Learning Representations*, volume 2026, pp. 88145–88169, 2026b.

Haodong Jing, Panqi Yang, Dongyao Jiang, Zhipeng Liu, Nanning Zheng, and Yongqiang Ma. Evoke: Efficient and high-fidelity eeg-to-video reconstruction via decoupling implicit neural representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5539–5547, 2026.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1):6241, 2024.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Chong Li, Xuelin Qian, Yun Wang, Jingyang Huo, Xiangyang Xue, Yanwei Fu, and Jianfeng Feng. Enhancing cross-subject fmri-to-video decoding with global-local functional alignment. In *European Conference on Computer Vision*, pp. 353–369. Springer, 2024a.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024b.

Dongyang Li, Kunpeng Xie, Mingyang Wu, Yiwei Kong, Jiahua Tang, Haoyang Qin, Chen Wei, and Quanying Liu. Mindpilot: Closed-loop visual stimulation optimization for brain modulation with eeg-guided diffusion. *arXiv preprint arXiv:2602.10552*, 2026.

Junxiang Liu, Junming Lin, Jie Zhou, Wei Xiong, Jiangtong Li, Jie Li, Jie Zhuang, and Hongfei Ji. Dynamind: Reconstructing dynamic visual scenes from eeg by aligning temporal dynamics and multimodal semantics to guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5378–5387, 2026a.

Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024.

Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. Eegmirror: Leveraging eeg data in the wild via montage-agnostic self-supervision for eeg to video decoding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18273–18283. IEEE, 2025.

Xuan-Hao Liu, Yan-Kai Liu, Tianyi Zhou, Bao-Liang Lu, and Wei-Long Zheng. Mindcross: Fast new subject adaptation with limited data for cross-subject video reconstruction from brain signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17589–17597, 2026b.

Yizhuo Lu, Changde Du, Chong Wang, Xuanliu Zhu, Liuyun Jiang, Xujin Li, and Huiguang He. Animate your thoughts: Reconstruction of dynamic natural vision from human brain activity. In *International Conference on Learning Representations*, volume 2025, pp. 67529–67567, 2025.

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.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6545–6554. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

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.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *International conference on learning representations*, volume 2024, pp. 47648–47665, 2024.

Jingyuan Sun, Mingxiao Li, and Marie-Francine Moens. Neuralflix: A simple while effective framework for semantic decoding of videos from non-invasive brain recordings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7096–7104, 2025.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

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.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

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.

Minghan Yang, Lan Yang, Ke Li, Honggang Zhang, Kaiyue Pang, and Yi-Zhe Song. Semvideo: Reconstructs what you watch from brain activity via hierarchical semantic guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13658–13669, 2026.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2063–2073. IEEE, 2025b.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. *arXiv preprint arXiv:2602.07495*, 2026.

Tian-Yi Zhou, Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. Mindcine: Multimodal eeg-to-video reconstruction with large-scale pretrained models. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7427–7431, 2026a. doi: 10.1109/ICASSP55912.2026.11460466.

Tian-Yi Zhou, Zhenghao Xiao, Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. Multi-granularity eeg decoding: Bridging object-background semantics and temporal motion for video reconstruction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 48, 2026b.

EntiNeuro: Multi-Entity Visual Decoding and Video Reconstruction from EEG Signals

————Appendix————

CONTENTS

A Related Work **15**

A.1 Brain Visual Reconstruction 15

A.2 Brain-Stimulus Datasets 15

B Additional Dataset Details **15**

B.1 Stimulus Construction 15

B.2 Pre-experiment Evaluation 16

B.3 Attention Check Question Design 16

B.4 Synchronized EEG and eye-tracking recording 17

B.5 Data Preprocessing 17

B.6 Post-experiment Subjective Assessment 18

C Implementation Details **18**

C.1 Architecture Details 18

C.2 Training Details 19

D Additional Details of EntiBench **21**

D.1 Compared Methods 21

D.2 Evaluation Metrics 22

D.3 Evaluation Protocol Details 24

E Additional Entity Recognition Results **25**

E.1 Frequency-Domain Feature Evaluation 25

E.2 Effect of Session Averaging 25

F Additional Reconstruction Results **26**

F.1 Detailed Quantitative Results 26

F.2 Ablation Settings 26

F.3 Additional Qualitative Results 27

F.4 Failure Cases 27

G Extension to Three-Entity Scenes **27**

G.1 Three-Entity Recognition 27

G.2 Three-Entity Video Reconstruction 28

A Related Work

A.1 Brain Visual Reconstruction

Brain visual reconstruction typically maps neural activity into pretrained visual or generative representation spaces to recover perceived stimuli. For static images, fMRI-based methods such as MindEye and MindEye2 combine semantic and perceptual representations with diffusion models for high-fidelity reconstruction (Scotti et al., 2023; 2024; Beliy et al., 2026), while related studies extend visual reconstruction to MEG and EEG (Benchetrit et al., 2024; Zhang et al., 2025a; Dai et al., 2026). Extending brain decoding from images to videos further requires modeling temporal dynamics in addition to visual semantics. Recent fMRI-based methods address this challenge through spatiotemporal representation learning, semantic guidance, and motion-aware conditioning (Chen et al., 2023; Lu et al., 2025; Gong et al., 2024; Sun et al., 2025; Yang et al., 2026).

More recently, EEG-based video reconstruction has emerged by aligning neural representations with pretrained semantic or generative spaces. EEG2Video combines semantic prediction with visual and dynamic guidance for EEG-conditioned video reconstruction (Liu et al., 2024). Subsequent methods further explore multi-granularity semantic conditioning, temporal modeling, motion-aware representations, and cross-subject generalization (Zhou et al., 2026a;b; Liu et al., 2026a; Huang et al., 2025; Jing et al., 2026; Huang et al., 2026a; Liu et al., 2025; 2026b; Han & Tan, 2026). Despite this progress, existing EEG-based video reconstruction largely focuses on a dominant semantic concept or overall scene representation, while controlled reconstruction of multiple semantically distinct entities remains underexplored. Our work addresses this gap by explicitly studying multi-entity video reconstruction from EEG, with a focus on preserving the semantic composition of multiple entities in dynamic scenes.

A.2 Brain-Stimulus Datasets

Brain–stimulus datasets have progressively expanded from static object and scene perception to naturalistic dynamic vision. NSD provides densely sampled fMRI responses to natural-scene images from COCO (Allen et al., 2022), while THINGS-EEG, THINGS-EEG2, and THINGS-data support large-scale investigation of object representations across EEG, fMRI, and MEG (Grootswagers et al., 2022; Gifford et al., 2022; Hebart et al., 2023). For dynamic visual perception, CC2017, fMRI-Video, and BOLD Moments provide fMRI responses to natural video stimuli (Wen et al., 2018; Li et al., 2024a; Lahner et al., 2024). EEG-based and multimodal video datasets have also emerged: SEED-DV records EEG responses to a large collection of video clips (Liu et al., 2024), CineBrain provides simultaneous EEG–fMRI recordings during audiovisual viewing (Gao et al., 2025), and EEG Moments combines EEG and eye-tracking recordings during natural video perception (Gifford et al., 2026).

Despite these advances, existing brain–stimulus datasets are either organized around individual visual concepts or use video stimuli without explicitly controlling the number, category composition, and interactions of multiple entities. To our knowledge, our dataset is the first visual decoding benchmark specifically designed for multi-entity perception with synchronized EEG and eye-tracking recordings. Its stimuli are organized around predefined cross-category entity compositions and interactions, supporting systematic evaluation of multi-entity decoding within dynamic scenes. Beyond data collection, the benchmark explicitly considers both entity recognition and video reconstruction, with particular emphasis on multi-entity preservation and temporal fidelity.

B Additional Dataset Details

B.1 Stimulus Construction

As illustrated in Fig. 1(A), we defined eight entity compositions based on six entity categories: person, dog, bird, flower, car, and ball. These compositions consist of six two-entity combinations and two three-entity combinations.

For each composition, we manually constructed prompts consisting of a scene-specific description and a shared set of generation constraints. The scene descriptions specified the target entities, their

interactions, and the surrounding context, while varying entity appearance, interaction patterns, camera viewpoints, and scene settings to increase stimulus diversity. The common constraints required all target entities to be present, prohibited additional benchmark entity categories outside the predefined composition, and encouraged realistic and natural scenes without watermarks or overlaid text. Using these prompts, we generated 100 candidate videos for each composition with Seedance 2.0 (Seedance et al., 2026), resulting in an initial pool of 800 videos. Two-entity videos were generated with a duration of 4 s, whereas three-entity videos lasted 6 s.

Table 4: Representative scene descriptions and common constraints used for stimulus generation.

Entity Composition	Scene Description
Person + Ball	A young man in a yellow basketball jersey repeatedly dribbles a basketball on an outdoor court.
Dog + Ball	A golden retriever puppy walks across a grassy field, approaches a tennis ball, sniffs it, and pauses beside it.
Bird + Flower	A small green hummingbird hovers beside a red hibiscus flower and repeatedly feeds from the blossom with its beak.
Person + Dog + Ball	A woman walks along a vineyard row while gently kicking a ball forward, as a French bulldog trots beside her and follows the rolling ball.
Common generation constraints	
All target entities should remain clearly visible during the video; no additional entities from the benchmark categories should appear; the scene and interaction should be realistic and natural; no watermark or overlaid text should be present.	

B.2 Pre-experiment Evaluation

Before data collection, three independent raters, who did not participate in the subsequent EEG and eye-tracking experiments, evaluated all 800 candidate videos individually. Each video was assessed along two dimensions: *entity suitability* and *interaction naturalness*. Entity suitability evaluates whether the target entities are clearly represented and consistent with the predefined entity composition, while interaction naturalness evaluates whether the spatial and dynamic interactions among the target entities appear plausible and natural.

Both criteria were rated independently using a 5-point Likert scale, as summarized in Table 5. For each entity composition, the ratings on the two criteria were considered jointly when selecting the final stimuli, rather than being collapsed into a single predefined weighted score. Based on this evaluation, 78 videos were retained from the 100 candidates in each composition, resulting in a balanced final set of 624 videos.

Table 5: Criteria used for pre-experiment stimulus evaluation. Each criterion was rated on a 5-point Likert scale, where 1 and 5 indicate the lowest and highest ratings, respectively.

Criterion	Description
Entity Suitability	Whether the target entities are clearly represented and consistent with the predefined entity composition.
Interaction Naturalness	Whether the spatial and dynamic interactions among the target entities appear plausible and natural.

B.3 Attention Check Question Design

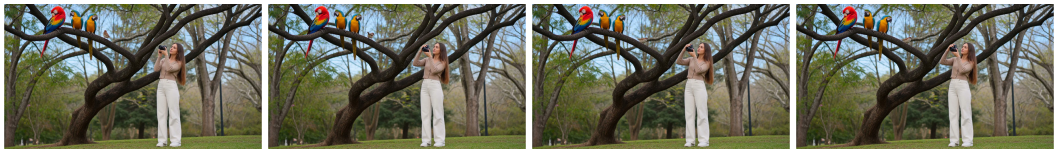
Before data collection, we created one three-option multiple-choice question for each stimulus video. These questions were designed to verify that participants attended to the visual content while avoiding unnecessary memory or reasoning demands. For each video, four candidate frames were uniformly sampled at 20%, 40%, 60%, and 80% of the video duration and provided to Qwen3-VL-32B-Instruct. The model first selected the frame in which the target entities were most clearly visible and then generated a question that could be answered from that frame alone. Each question was required to refer to at least one entity in the corresponding composition.

Question generation. We considered four types of objective visual questions: color (35%), count (25%), presence (25%), and visual attribute (15%). Question types were randomly assigned within each entity composition while preserving these overall proportions. We excluded questions about actions, relative spatial positions, background details, subjective judgments, and future events, since such questions may require temporal inference or admit ambiguous answers. The model returned the selected frame, question, three candidate answers, correct answer label, target entities, and a confidence score in a structured format.

Quality control. We automatically rejected questions with missing fields, fewer than three answer options, an invalid answer label, confidence below 0.75, or no reference to a target entity. The remaining questions were exported to a review table for final verification. Figure 6 presents representative examples covering all six entity categories.



Q: What color is the car?
A. Red **B.** Blue **C.** White



Q: What is the woman using to observe the birds?
A. Binoculars **B.** A camera **C.** A book



Q: How many dogs are visible?
A. One **B.** Two **C.** Three

Figure 6: Representative multiple-choice questions associated with the stimulus videos. Each row shows four frames in temporal order, followed by the corresponding question and three answer options. The examples are drawn from the car–flower, person–bird, and person–dog–ball compositions and collectively cover all six entity categories.

B.4 Synchronized EEG and Eye-Tracking Recording

Figure 7 illustrates an example of synchronized EEG and eye-tracking recording during video viewing. Both modalities were recorded continuously and temporally aligned with the corresponding video stimuli, including 4-s two-entity and 6-s three-entity videos. The eye-tracking trajectories overlaid on the video frames illustrate participants' visual exploration during viewing, while the timeline also shows the interspersed attention questions and inter-block rest periods.

B.5 Data Preprocessing

Raw EEG recordings were stored in continuous `.cnt` format at a sampling rate of 1000 Hz. We applied a 0.1–100 Hz band-pass filter and downsampled the signals to 200 Hz. Independent component analysis (ICA) was then used to identify and remove components dominated by ocular artifacts. The continuous EEG recordings were segmented according to stimulus triggers, resulting in 4-s trials for two-entity videos and 6-s trials for three-entity videos.

For frequency-domain analyses, the preprocessed 62-channel EEG signals were divided into non-overlapping 1-s windows. For each window, we computed log-relative power spectral density (PSD)

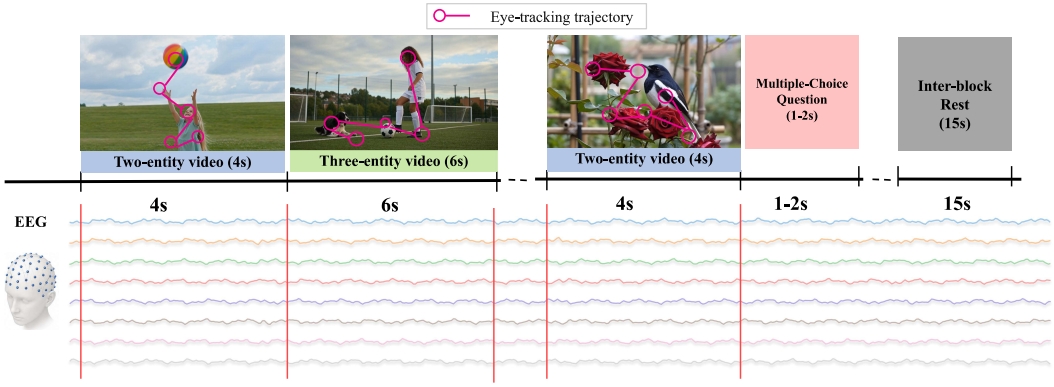


Figure 7: Example of synchronized EEG and eye-tracking recording during video viewing.

and Gaussian differential entropy (DE) over five frequency bands: δ (1–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), and γ (30–45 Hz).

Eye-tracking recordings were aligned with individual video trials according to the corresponding stimulus timestamps and resampled from 250 to 200 Hz to match the temporal resolution of the EEG signals. We retained 49 eye-tracking variables for subsequent modeling, yielding temporally aligned EEG and eye-tracking sequences for each video trial.

B.6 Post-experiment Subjective Assessment

After completing the experiment, participants provided an overall subjective assessment of the stimulus set using a 5-point Likert scale. Unlike the pre-experiment evaluation, which was conducted on individual candidate videos for stimulus selection, this assessment reflected participants' overall perceptions of the videos encountered during the experiment. Participants rated the stimulus set in terms of *overall realism* and *interaction naturalness*. The resulting mean ratings were 4.07 ± 0.95 and 3.65 ± 1.16, respectively, indicating generally positive subjective evaluations of the selected stimuli.

C Implementation Details

C.1 Architecture Details

EEG encoder. For raw EEG, we apply a temporal convolution with kernel size 1×25, followed by a depthwise spatial convolution across the 62 EEG channels. The resulting features are projected to 256 dimensions, processed by residual temporal blocks, and adaptively pooled to 32 temporal tokens. A learnable query resampler further compresses them into eight 256-dimensional EEG tokens for each session. For PSD and DE features, the four 1-s window features are flattened over channel and frequency dimensions and projected to 128 dimensions with learnable window embeddings. A two-layer query resampler with eight queries and eight attention heads, followed by a linear projection, produces the same eight 256-dimensional EEG tokens.

Eye encoder. For each session, the temporal branch encodes the 49-channel eye signal using a temporal convolution and residual temporal blocks, followed by adaptive pooling to 32 temporal features with 128 dimensions. In parallel, 32 Gaussian gaze heatmaps are encoded by a lightweight convolutional network into 128-dimensional spatial features. The temporal and spatial features are concatenated into 256-dimensional representations, after which learnable temporal embeddings are added. The resulting features are processed by a two-layer Transformer encoder and adaptively pooled to eight tokens, yielding eight 256-dimensional eye tokens for each session.

Eye-guided EEG fusion. For each session, the eight 256-dimensional eye tokens serve as queries, while the corresponding eight 256-dimensional EEG tokens serve as keys and values in multi-head cross-attention. The resulting eye-guided EEG representations are averaged across the three sessions, yielding eight 256-dimensional fused tokens.

Fine semantic interaction. The soft coarse semantic context is first projected from the text-encoder space to the 256-dimensional neural feature space. The eye-guided EEG tokens are then used as queries in two cross-attention blocks, with the projected coarse context serving as keys and values. This produces eight fine-grained neural tokens that combine session-aggregated EEG information, eye-guided selection, and the coarse entity-level semantic prior.

Tora conditioning. We use Tora (Zhang et al., 2025b) as a frozen video diffusion backbone. Fine-grained neural tokens are injected into selected Transformer blocks through gated cross-attention, with 3072-dimensional video states projected to 256-dimensional queries and the fine tokens used as keys and values. The attention output is projected back to 3072 dimensions and added through independent learnable scalar gates at blocks 10, 20, 30, and 40.

C.2 Training Details

Stage 1: coarse semantic pretraining. We jointly train the EEG encoder and coarse semantic heads using composition classification, multi-label entity prediction, and contrastive alignment with frozen CLIP text and video representations.

For the six predefined entity compositions, we apply cross-entropy loss to the composition logits $\mathbf{z}_i^{\mathrm{cat}}$:

$$\mathcal{L}_{\mathrm{cat}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(z_{i,y_i}^{\mathrm{cat}})}{\sum_{k=1}^{6}\exp(z_{i,k}^{\mathrm{cat}})}, \tag{16}$$

where y_i denotes the composition label of sample i.

For multi-label entity prediction, we apply a weighted binary cross-entropy loss to $\mathbf{z}_i^{\mathrm{obj}}$:

$$\mathcal{L}_{\mathrm{obj}} = -\frac{1}{6N}\sum_{i=1}^{N}\sum_{o=1}^{6}\left[w_+ y_{i,o}\log\sigma(z_{i,o}^{\mathrm{obj}}) + (1-y_{i,o})\log\left(1-\sigma(z_{i,o}^{\mathrm{obj}})\right)\right], \tag{17}$$

where $y_{i,o}\in\{0,1\}$ indicates whether entity o is present. We set the positive-class weight to $w_+ = 2.0$.

For semantic alignment, the normalized EEG embedding $\mathbf{b}_i$ is aligned with the corresponding normalized CLIP text and video embeddings, denoted by $\mathbf{t}_i$ and $\mathbf{v}_i$, respectively. For a target modality $a\in\{t,v\}$, we define

$$s_{ij}^{a} = \frac{\mathbf{b}_i^{\top}\mathbf{a}_j}{\tau}, \qquad \tau = 0.1, \tag{18}$$

and use a symmetric InfoNCE objective:

$$\mathcal{L}_{\mathrm{NCE}}(\mathbf{b},\mathbf{a}) = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(s_{ii}^{a})}{\sum_{j=1}^{N}\exp(s_{ij}^{a})} + \log\frac{\exp(s_{ii}^{a})}{\sum_{j=1}^{N}\exp(s_{ji}^{a})}\right]. \tag{19}$$

The text and video alignment losses are

$$\mathcal{L}_{\mathrm{txt}} = \mathcal{L}_{\mathrm{NCE}}(\mathbf{b},\mathbf{t}), \qquad \mathcal{L}_{\mathrm{vid}} = \mathcal{L}_{\mathrm{NCE}}(\mathbf{b},\mathbf{v}). \tag{20}$$

We combine them into the semantic alignment objective:

$$\mathcal{L}_{\mathrm{align}} = \lambda_{\mathrm{txt}}\mathcal{L}_{\mathrm{txt}} + \lambda_{\mathrm{vid}}\mathcal{L}_{\mathrm{vid}}. \tag{21}$$

The complete coarse pretraining objective is

$$\mathcal{L}_{\mathrm{coarse}} = \mathcal{L}_{\mathrm{cat}} + \lambda_{\mathrm{obj}}\mathcal{L}_{\mathrm{obj}} + \mathcal{L}_{\mathrm{align}}, \tag{22}$$

where $\lambda_{\mathrm{obj}} = 1.0$ and $\lambda_{\mathrm{txt}} = \lambda_{\mathrm{vid}} = 0.1$. After pretraining, the EEG encoder is frozen for reconstruction. The composition head and semantic alignment branch are discarded, while the entity prediction head is retained to construct the coarse semantic anchor.

Stage 2: reconstruction training. After coarse semantic pretraining, we freeze the EEG encoder, coarse semantic branch, T5 encoder, video autoencoder, and pretrained Tora backbone. We optimize only the eye encoder, eye-guided EEG fusion, fine semantic interaction, semantic projection, and gated conditioning modules.

During reconstruction training, the coarse and fine conditions are used jointly. The soft context $\mathbf{C}_i^{\text{soft}}$ first refines the eye-guided EEG representation to produce the fine tokens $\mathbf{Z}_i^{\text{fine}}$. The hard prompt q_i^{hard} is encoded by the frozen text encoder as $\mathbf{C}_i^{\text{hard}} = E_{\text{text}}(q_i^{\text{hard}})$ and provided through Tora's native text-conditioning pathway, while $\mathbf{Z}_i^{\text{fine}}$ is injected into intermediate Transformer blocks through gated cross-attention. The diffusion objective follows the original Tora training formulation (Zhang et al., 2025b), denoted as $\mathcal{L}_{\text{diff}}$.

To preserve semantic information in the fine representation, we introduce an auxiliary semantic objective. Given $\mathbf{Z}_i^{\text{fine}} \in \mathbb{R}^{M\times d}$, we mean-pool the fine tokens and project them into the T5 embedding space:

$$\widehat{\mathbf{c}}_i = \text{Norm}\left(W_{\text{sem}}\,\text{LN}\left(\frac{1}{M}\sum_{m=1}^{M}\mathbf{Z}_{i,m}^{\text{fine}}\right)\right). \tag{23}$$

The target $\mathbf{c}_i^{\text{GT}}$ is obtained by masked mean pooling over the valid tokens of the frozen T5 representation of the corresponding ground-truth caption, followed by normalization. We define

$$\mathcal{L}_{\text{sem}} = 1 - \cos\left(\widehat{\mathbf{c}}_i, \mathbf{c}_i^{\text{GT}}\right) + \lambda_{\text{smooth}}\,\text{SmoothL1}\left(\widehat{\mathbf{c}}_i, \mathbf{c}_i^{\text{GT}}\right), \tag{24}$$

where $\lambda_{\text{smooth}} = 0.1$. The ground-truth caption is used only as an auxiliary semantic target and is not used as a generation condition.

We first perform a 200-step semantic warmup using only $\mathcal{L}_{\text{sem}}$, followed by joint optimization with

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}}, \tag{25}$$

where $\lambda_{\text{sem}} = 0.1$.

To reduce reliance on the coarse semantic prior, we apply context dropout at two locations. During fine semantic interaction, the soft coarse context is dropped with probability $p_{\text{fine}} = 0.3$:

$$\widetilde{\mathbf{C}}_i^{\text{soft}} = \begin{cases} \mathbf{0}, & \text{with probability } 0.3, \\ \mathbf{C}_i^{\text{soft}}, & \text{otherwise.} \end{cases} \tag{26}$$

During diffusion training, the hard coarse text condition is replaced with Tora's empty-text embedding with probability $p_{\text{text}} = 0.2$:

$$\widetilde{\mathbf{C}}_i^{\text{hard}} = \begin{cases} \mathbf{C}^{\text{empty}}, & \text{with probability } 0.2, \\ \mathbf{C}_i^{\text{hard}}, & \text{otherwise,} \end{cases} \tag{27}$$

where $\mathbf{C}^{\text{empty}}$ denotes the frozen T5 representation of an empty prompt. The two dropout operations respectively discourage the fine branch from relying solely on the coarse semantic context and reduce over-reliance on Tora's native text-conditioning pathway.

Optimization settings. For Stage 1, we use AdamW with a learning rate of 3×10^{-4}, weight decay 0.03, and betas $(0.9, 0.999)$. The batch size is 24, training lasts for at most 100 epochs, and the learning rate follows cosine annealing. Gradient clipping is set to 1.0.

For Stage 2, we use AdamW with betas $(0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 10^{-4}, while no weight decay is applied to the injection gates. The eye encoder, eye-guided EEG fusion, fine semantic interaction, semantic projection, Tora attention projections, and injection gates use learning rates of 5×10^{-5}, 1×10^{-4}, 5×10^{-5}, 5×10^{-5}, 1×10^{-4}, and 5×10^{-4}, respectively. We train for 1000 optimization steps with a micro-batch size of 1 and gradient accumulation of 8, giving an effective batch size of 8. During the first 100 steps, SAT finetuning uses 5% of the target learning rates before switching to the full values. We use gradient clipping of 1.0 and BF16 precision. All experiments are conducted on two NVIDIA L40 GPUs with 48 GB memory each.

D Additional Details of EntiBench

D.1 Compared Methods

D.1.1 Entity Recognition Encoders

To assess whether entity-related information can be consistently decoded across different model architectures, we evaluate a diverse set of representative EEG encoders, including ShallowNet, DeepNet, EEGNet, EEG Conformer, TSConv, and GLMNet, together with our shared EEG encoder. These models cover convolutional, attention-based, and hybrid architectures commonly used for EEG representation learning. All encoders are evaluated using the same data splits, input preprocessing, and multi-label classification protocol to ensure a consistent comparison.

ShallowNet and DeepNet. We adopt ShallowNet and DeepNet (Schirrmeister et al., 2017) as representative convolutional EEG architectures. ShallowNet employs a compact temporal-spatial convolutional design, whereas DeepNet uses a deeper hierarchical convolutional architecture for learning EEG representations.

EEGNet. EEGNet (Lawhern et al., 2018) is a compact convolutional architecture based on temporal and depthwise spatial convolutions. Its lightweight design provides a commonly used baseline for EEG decoding.

EEG Conformer. EEG Conformer (Song et al., 2022) combines convolutional feature extraction with self-attention, enabling joint modeling of local temporal patterns and longer-range dependencies in EEG signals.

TSConv. TSConv (Song et al., 2024) employs temporal and spatial convolutional operations to capture complementary structure in multichannel EEG signals.

GLMNet. GLMNet (Liu et al., 2024), introduced in EEG2Video, combines global features extracted from all EEG channels with local features from vision-associated channels to emphasize visually relevant neural responses.

D.1.2 Video Reconstruction Baselines

We compare our method with three representative EEG-to-video reconstruction approaches with publicly available implementations: EEG2Video (Liu et al., 2024), MindCine (Zhou et al., 2026a), and DynaMind (Liu et al., 2026a). For all baselines, we follow the released implementations and adapt them to our dataset while using the same training and test splits for a consistent comparison.

EEG2Video. EEG2Video (Liu et al., 2024) reconstructs dynamic visual stimuli by decoding complementary perceptual, semantic, and motion-related information from EEG and using these representations to condition video generation. We retrain the EEG-dependent components on our dataset following the released implementation.

MindCine. MindCine (Zhou et al., 2026a) reconstructs perceived videos by learning semantic and perceptual representations from EEG and integrating them with pretrained generative priors. We adapt its publicly available implementation to our experimental setting and retrain the EEG-dependent components using our training data.

DynaMind. DynaMind (Liu et al., 2026a) jointly models semantic information and temporal dynamics for EEG-based video reconstruction. It employs a regional-aware semantic mapper to derive semantic conditioning from EEG and a temporal-aware dynamic aligner to construct a temporal blueprint, which jointly guide the subsequent video generation process. We follow its released implementation and adapt the framework to our dataset.

Other recent EEG-based visual reconstruction methods, including NEED (Huang et al., 2026a), MINDEV (Huang et al., 2025), and EVOKE (Jing et al., 2026), are discussed in the main paper but are not included in the quantitative comparison because reproducible public implementations were unavailable at the time of our experiments.

GT Caption. As a semantic reference, GT Caption uses the same pretrained Tora backbone conditioned directly on the ground-truth caption through its native text-conditioning pathway, without

neural conditioning. Three videos are generated for each stimulus using different random seeds, and the reported results are aggregated across the three generations.

D.2 Evaluation Metrics

D.2.1 Entity Recognition Metrics

Entity recognition is formulated as a six-class multi-label classification problem. Each sample contains two positive entity labels in the primary two-entity benchmark and three positive labels in the three-entity extension. We evaluate recognition performance using average precision (AP), area under the receiver operating characteristic curve (AUROC), and Top-K set metrics, where K is set to the number of target entities ($K = 2$ for the main benchmark and $K = 3$ for the three-entity extension).

Average Precision. We report both Macro AP and Micro AP. Macro AP computes average precision independently for each entity category and then averages across the six categories, assigning equal weight to each class. Micro AP aggregates predictions and labels over all samples and categories before computing average precision.

AUROC. Macro AUROC is obtained by averaging the class-wise AUROC values, whereas Micro AUROC is computed after aggregating predictions across all samples and categories. Both AP and AUROC are independent of a fixed classification threshold.

Top-K metrics. For each sample i, we construct the predicted entity set $\hat{\mathcal{Y}}_i$ from the K highest-scoring classes, where $K = 2$ for the primary two-entity benchmark and $K = 3$ for the three-entity extension. Let $\mathcal{Y}_i$ denote the corresponding ground-truth entity set. We compute

$$\mathrm{F1}_i = \frac{2|\hat{\mathcal{Y}}_i \cap \mathcal{Y}_i|}{|\hat{\mathcal{Y}}_i| + |\mathcal{Y}_i|}, \tag{28}$$

$$\mathrm{SetAcc}_i = \mathbb{I}\left[\hat{\mathcal{Y}}_i = \mathcal{Y}_i\right], \tag{29}$$

$$\mathrm{Jaccard}_i = \frac{|\hat{\mathcal{Y}}_i \cap \mathcal{Y}_i|}{|\hat{\mathcal{Y}}_i \cup \mathcal{Y}_i|}. \tag{30}$$

The final Top-K F1, Set Accuracy, and Jaccard scores are obtained by averaging the corresponding sample-wise values over the test set.

D.2.2 Video Reconstruction Metrics

We evaluate video reconstruction using complementary automatic and human-centered criteria that cover semantic fidelity, frame and video fidelity and multi-entity preservation.

Semantic fidelity. Following prior EEG-to-video reconstruction protocols, we perform a 2-way Top-1 retrieval test at both the image and video levels. In each trial, the ground-truth sample corresponding to a reconstructed stimulus is paired with one randomly sampled distractor, and the retrieval is considered correct if the matched ground-truth sample achieves the higher similarity. For image-level evaluation, we use a Vision Transformer (ViT) pretrained on ImageNet (Dosovitskiy et al., 2020) to extract visual representations, whereas a pretrained VideoMAE (Tong et al., 2022) is used for video-level evaluation. The resulting image- and video-level metrics are denoted as 2-way-I and 2-way-V, respectively.

We further report ViFi-Score using the pretrained ViFiCLIP model (Rasheed et al., 2023). Given a reconstructed video $\hat{V}$ and its corresponding ground-truth video V, ViFi-Score is computed as the cosine similarity between their video-level representations:

$$\mathrm{ViFi}(\hat{V}, V) = \frac{f_{\mathrm{ViFi}}(\hat{V})^\top f_{\mathrm{ViFi}}(V)}{\|f_{\mathrm{ViFi}}(\hat{V})\|_2 \|f_{\mathrm{ViFi}}(V)\|_2}. \tag{31}$$

Higher values indicate stronger semantic agreement between the reconstructed and ground-truth videos.

Frame fidelity. We report the structural similarity index measure (SSIM) (Wang et al., 2004) and peak signal-to-noise ratio (PSNR) to quantify low-level visual correspondence between reconstructed and ground-truth videos. Both metrics are computed frame-wise and averaged over the full video sequence, with higher values indicating better visual reconstruction quality.

Video fidelity. We evaluate video fidelity using Fréchet Video Distance (FVD) (Unterthiner et al., 2018) and Warp-LPIPS. FVD measures the distributional discrepancy between reconstructed and ground-truth videos in a pretrained spatiotemporal feature space, whereas Warp-LPIPS measures motion-compensated temporal consistency within each reconstructed video. Lower values indicate better performance for both metrics.

Fréchet Video Distance. Following the common temporal alignment protocol described above, all 4-s videos are first normalized to 12 uniformly spaced frames. Since the I3D feature extractor used for FVD requires 16-frame inputs, the aligned sequences are further converted to 16 frames using identical linear temporal interpolation for all reconstructed and ground-truth videos. We then use a frozen I3D network (Carreira & Zisserman, 2017) pretrained on Kinetics-400 to extract a 400-dimensional feature for each video. For each subject, features from the 66 ground-truth videos and the corresponding 66 reconstructed videos form two distributions with means and covariance matrices (μ_r, Σ_r) and (μ_g, Σ_g), respectively. FVD is computed as

$$\mathrm{FVD} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right). \tag{32}$$

FVD is therefore computed once over the complete test set for each subject rather than independently for individual video pairs.

Warp-LPIPS. Warp-LPIPS evaluates temporal self-consistency within each reconstructed video without reference to the corresponding ground-truth video. Each reconstruction is temporally aligned to 12 frames and resized to 256×448. For each adjacent pair (I_t, I_{t+1}), we use a frozen RAFT-small model (Teed & Deng, 2020) to estimate backward optical flow,

$$B_t = \mathrm{RAFT}(I_{t+1}, I_t), \tag{33}$$

and warp the preceding frame into the coordinate system of the subsequent frame:

$$\hat{I}_{t+1}(x) = I_t\left(x + B_t(x)\right). \tag{34}$$

We then measure the perceptual discrepancy between the warped frame and the actual subsequent frame using LPIPS (Zhang et al., 2018) with an AlexNet backbone:

$$d_t = \mathrm{LPIPS}_{\mathrm{Alex}}\left(\hat{I}_{t+1}, I_{t+1}\right). \tag{35}$$

The video-level score is obtained by averaging over the 11 adjacent frame pairs:

$$\mathrm{WarpLPIPS}(V) = \frac{1}{11}\sum_{t=0}^{10} d_t. \tag{36}$$

The resulting scores are averaged over the 66 test videos for each subject. Lower Warp-LPIPS indicates better motion-compensated temporal coherence.

Multi-entity fidelity. We evaluate target-entity preservation using Entity F1, computed by comparing the set of entity categories detected in each reconstructed video with the corresponding ground-truth entity set. This metric penalizes both missing and incorrect entities, including cases where only a subset of the target entities is reconstructed. We further report Entity Count Error (ECE), defined as the mean frame-wise absolute difference between the numbers of entity categories detected in the reconstructed and ground-truth videos. Entity detection is performed using Grounding DINO over the benchmark entity vocabulary, with a box threshold of 0.35 and a text threshold of 0.25. Higher Entity F1 and lower ECE indicate better preservation of entity identity and entity cardinality, respectively.

Human evaluation. Automatic metrics provide complementary measurements of semantic, frame-level, video-level, and multi-entity fidelity, but may not fully reflect the perceptual quality of reconstructed videos. We therefore complement them with human evaluation along three dimensions. *Entity Consistency* assesses whether the target entities are correctly preserved throughout the reconstructed video. *Temporal Naturalness and Coherence* evaluates whether entity motion appears

natural and whether visual content evolves coherently over time with respect to the ground-truth video. *Overall Similarity* measures the overall perceptual similarity between the reconstructed and ground-truth videos, considering both visual content and temporal dynamics. Each criterion is rated on a 5-point Likert scale, with higher scores indicating better reconstruction quality.

D.3 Evaluation Protocol Details

Entity recognition protocol. Entity recognition is evaluated independently for each participant using the 468 two-entity videos. In the *session-average* setting, EEG responses corresponding to the same video are first aligned and averaged across the three recording sessions at the signal level, yielding one averaged EEG trial per video. We then perform five-fold cross-validation at the video level. In each fold, approximately 20% of the videos are held out for testing, while approximately 15% of the remaining videos are used for validation and the rest for training. The splits are grouped by video identity and stratified by entity composition, ensuring that the same video never appears across the training, validation, and test sets within a fold. Performance is averaged over the five test folds.

In the *cross-session* setting, EEG responses are retained separately for each recording session. Two sessions are used for training and validation, while the entire remaining session is held out for testing. We evaluate all three session combinations, i.e., Sessions 1–2 $\rightarrow$ Session 3, Sessions 1–3 $\rightarrow$ Session 2, and Sessions 2–3 $\rightarrow$ Session 1, and average performance across the three splits. Since the same set of stimuli is presented in all sessions, this setting specifically evaluates robustness to session-level variation for repeated stimuli rather than generalization to unseen videos. The same splitting and evaluation procedures are applied to all compared EEG encoders. Final results are reported as the mean and standard deviation across the 20 participants.

Video reconstruction protocol. Video reconstruction is evaluated on the 468 two-entity videos. We construct a fixed, composition-balanced training–test split at the video level. Specifically, 11 videos are randomly selected from each of the six two-entity compositions, yielding 66 test videos, while the remaining 402 videos are used for training. The same test video IDs are used for all 20 subjects and all reconstruction methods, ensuring that each method is evaluated on an identical set of stimuli. No video appears in both the training and test sets.

For reconstruction, the three repeated sessions are encoded independently and aggregated at the feature level before video generation. The same protocol is applied to all methods: each session is first processed by the corresponding neural encoder, after which the resulting representations are averaged across the three sessions to obtain a single neural condition for each stimulus. For EntiRecon, the EEG and eye-tracking representations are fused within each corresponding session before session-level averaging.

To ensure a fair comparison across methods with different native frame rates, all 4-s reconstructed and ground-truth videos are first temporally normalized to a shared grid of 12 uniformly spaced frames. Frame-level metrics, including SSIM and PSNR, are computed on these temporally aligned frames at a spatial resolution of 288×512. For Warp-LPIPS, the same 12-frame sequences are resized to 256×448, yielding 11 adjacent frame pairs for temporal consistency evaluation. For FVD, the aligned 12-frame sequences are uniformly interpolated to 16 frames to satisfy the input requirement of the I3D feature extractor. The same temporal interpolation and metric-specific spatial preprocessing are applied to all reconstruction methods and ground-truth videos.

For all automatic metrics except FVD, including 2-way-I, 2-way-V, ViFi-Score, SSIM, PSNR, Warp-LPIPS, Entity F1, and ECE, metric values are first computed for individual test videos and then averaged over the 66 test videos to obtain a subject-level score. The resulting scores are subsequently aggregated across the 20 subjects. FVD is treated separately because it measures the distributional discrepancy between two sets of videos rather than individual video pairs. For each subject, I3D features are extracted from the complete sets of 66 reconstructed and ground-truth test videos, and a single FVD score is computed from the corresponding feature distributions. The resulting subject-level FVD scores are then aggregated across subjects in the same manner.

For the GT Caption reference, reconstruction metrics are computed independently for each generation seed. The final GT Caption results are obtained by aggregating the scores across three generation seeds.

Human evaluation protocol. We recruited five participants for the human evaluation. For each trial, the reconstructed video and its corresponding ground-truth video were presented simultaneously for direct comparison. The order of reconstruction methods was randomized, and the method identities were hidden from the participants. Each participant rated the reconstruction using a 5-point Likert scale. Final scores are reported as mean ± standard deviation across the five raters.

E Additional Entity Recognition Results

E.1 Frequency-Domain Feature Evaluation

To complement the raw EEG results reported in the main paper, we further evaluate entity recognition using differential entropy (DE) and power spectral density (PSD) features. We use the same encoder architectures, data splits, and evaluation protocols as in the raw-signal experiments, changing only the EEG representation.

Differential entropy. As shown in Table 6, DE features yield consistently above-chance performance across all encoders and both evaluation settings. Compared with raw EEG, DE generally provides stronger recognition performance, particularly for AP, AUROC, and the set-based metrics. Notably, the gap between session-average and cross-session evaluation is substantially reduced, and several metrics remain comparable across the two settings. This suggests that DE captures discriminative spectral information that is less sensitive to session-level variation than the raw temporal waveform.

Table 6: Entity recognition performance using differential entropy (DE) features under session-average and cross-session evaluation settings. Results are reported as mean ± standard deviation across 20 subjects. The best result in each evaluation setting is highlighted in bold.

EEG Encoder	Macro AP ↑	Micro AP ↑	Macro AUC ↑	Micro AUC ↑	Top-2 F1 ↑	Set Acc. ↑	Jaccard ↑
Chance Level	0.3333	0.3333	0.5000	0.5000	0.3333	0.0667	0.2444
Session-Average Evaluation							
ShallowNet	0.4380 ± 0.0467	0.4118 ± 0.0466	0.5870 ± 0.0491	0.5822 ± 0.0468	0.4026 ± 0.0395	0.1952 ± 0.0439	0.3334 ± 0.0403
DeepNet	0.4060 ± 0.0308	0.3728 ± 0.0254	0.5513 ± 0.0359	0.5389 ± 0.0270	0.3680 ± 0.0276	0.1686 ± 0.0315	0.3015 ± 0.0276
EEGNet	0.4212 ± 0.0476	0.3991 ± 0.0500	0.5701 ± 0.0495	0.5701 ± 0.0497	0.3911 ± 0.0450	0.1942 ± 0.0411	0.3254 ± 0.0433
Conformer	0.4019 ± 0.0360	0.3751 ± 0.0314	0.5486 ± 0.0378	0.5425 ± 0.0311	0.3691 ± 0.0305	0.1812 ± 0.0288	0.3065 ± 0.0294
TSConv	0.4402 ± 0.0436	0.4163 ± 0.0430	0.5902 ± 0.0433	0.5885 ± 0.0426	**0.4095 ± 0.0404**	0.2031 ± 0.0393	0.3407 ± 0.0398
GLMNet	**0.4448 ± 0.0451**	**0.4210 ± 0.0481**	**0.5963 ± 0.0453**	**0.5923 ± 0.0462**	0.4093 ± 0.0415	**0.2097 ± 0.0360**	**0.3428 ± 0.0393**
Our Encoder	0.4400 ± 0.0352	0.4127 ± 0.0372	0.5931 ± 0.0382	0.5848 ± 0.0382	0.4028 ± 0.0366	0.2057 ± 0.0323	0.3371 ± 0.0346
Cross-Session Evaluation							
ShallowNet	0.4326 ± 0.0370	0.4115 ± 0.0340	0.6052 ± 0.0369	0.5872 ± 0.0338	0.4034 ± 0.0288	0.1977 ± 0.0264	0.3349 ± 0.0275
DeepNet	0.3901 ± 0.0262	0.3728 ± 0.0221	0.5601 ± 0.0289	0.5448 ± 0.0245	0.3657 ± 0.0214	0.1752 ± 0.0187	0.3022 ± 0.0200
EEGNet	0.4148 ± 0.0361	0.4011 ± 0.0305	0.5866 ± 0.0371	0.5774 ± 0.0309	0.3981 ± 0.0262	0.2020 ± 0.0264	0.3327 ± 0.0257
Conformer	0.3859 ± 0.0225	0.3696 ± 0.0160	0.5535 ± 0.0247	0.5432 ± 0.0195	0.3718 ± 0.0160	0.1850 ± 0.0137	0.3095 ± 0.0146
TSConv	0.4279 ± 0.0303	0.4110 ± 0.0284	0.6024 ± 0.0300	0.5891 ± 0.0267	0.4074 ± 0.0248	0.2049 ± 0.0239	0.3399 ± 0.0241
GLMNet	**0.4395 ± 0.0369**	**0.4181 ± 0.0377**	**0.6117 ± 0.0355**	**0.5929 ± 0.0343**	**0.4098 ± 0.0301**	**0.2119 ± 0.0284**	**0.3438 ± 0.0292**
Our Encoder	0.4307 ± 0.0334	0.4053 ± 0.0302	0.5957 ± 0.0351	0.5822 ± 0.0303	0.4017 ± 0.0258	0.2094 ± 0.0223	0.3376 ± 0.0244

Power spectral density. PSD features show a similar pattern, as summarized in Table 7. All architectures remain above chance, and most metrics improve over their raw-EEG counterparts. Although PSD exhibits a clearer decrease from session-average to cross-session evaluation than DE, the degradation remains smaller than that observed with raw EEG. The consistent gains obtained from both DE and PSD indicate that frequency-domain representations provide complementary and comparatively robust cues for decoding entity semantics.

Overall, DE and PSD provide stronger entity recognition than raw EEG across most architectures and metrics, with DE showing the strongest overall performance. Together with the raw-signal results, these findings indicate that multi-entity information is reflected in both temporal EEG responses and their spectral characteristics.

E.2 Effect of Session Averaging

We further examine the effect of averaging repeated EEG responses across recording sessions. For the *single-session* setting, each of the three sessions is evaluated independently, and the resulting metrics are averaged across sessions. For the *session-average* setting, EEG responses corresponding

Table 7: Entity recognition performance using power spectral density (PSD) features under session-average and cross-session evaluation settings. Results are reported as mean ± standard deviation across 20 subjects. The best result in each evaluation setting is highlighted in bold.

EEG Encoder	Macro AP ↑	Micro AP ↑	Macro AUC ↑	Micro AUC ↑	Top-2 F1 ↑	Set Acc. ↑	Jaccard ↑
Chance Level	0.3333	0.3333	0.5000	0.5000	0.3333	0.0667	0.2444
Session-Average Evaluation							
ShallowNet	0.4201 ± 0.0416	0.3956 ± 0.0426	0.5724 ± 0.0451	0.5700 ± 0.0469	0.3984 ± 0.0445	0.1934 ± 0.0463	0.3301 ± 0.0445
DeepNet	0.3991 ± 0.0283	0.3686 ± 0.0245	0.5487 ± 0.0340	0.5376 ± 0.0315	0.3680 ± 0.0299	0.1716 ± 0.0297	0.3026 ± 0.0293
EEGNet	0.4137 ± 0.0348	0.3893 ± 0.0354	0.5647 ± 0.0417	0.5637 ± 0.0421	0.3900 ± 0.0351	0.1966 ± 0.0297	0.3255 ± 0.0327
Conformer	0.4085 ± 0.0309	0.3827 ± 0.0285	0.5572 ± 0.0341	0.5540 ± 0.0335	0.3855 ± 0.0299	0.1972 ± 0.0287	0.3227 ± 0.0291
TSConv	**0.4245 ± 0.0358**	**0.4108 ± 0.0370**	0.5875 ± 0.0395	**0.5876 ± 0.0406**	**0.4116 ± 0.0400**	**0.2048 ± 0.0423**	0.3427 ± 0.0406
GLMNet	0.4237 ± 0.0432	0.4107 ± 0.0447	**0.5885 ± 0.0457**	0.5873 ± 0.0481	0.4082 ± 0.0432	0.2029 ± 0.0416	0.3398 ± 0.0422
Our Encoder	0.4167 ± 0.0406	0.4005 ± 0.0386	0.5760 ± 0.0448	0.5833 ± 0.0425	0.3874 ± 0.0346	0.1960 ± 0.0313	**0.3436 ± 0.0331**
Cross-Session Evaluation							
ShallowNet	0.4018 ± 0.0297	0.3855 ± 0.0286	0.5762 ± 0.0355	0.5677 ± 0.0342	0.3949 ± 0.0326	0.1941 ± 0.0361	0.3279 ± 0.0332
DeepNet	0.3831 ± 0.0187	0.3660 ± 0.0137	0.5542 ± 0.0217	0.5427 ± 0.0156	0.3728 ± 0.0160	0.1819 ± 0.0157	0.3092 ± 0.0153
EEGNet	0.3927 ± 0.0302	0.3794 ± 0.0274	0.5654 ± 0.0364	0.5588 ± 0.0343	0.3856 ± 0.0313	0.1817 ± 0.0420	0.3176 ± 0.0345
Conformer	0.3922 ± 0.0243	0.3796 ± 0.0245	0.5615 ± 0.0257	0.5543 ± 0.0248	0.3810 ± 0.0196	0.1839 ± 0.0278	0.3153 ± 0.0211
TSConv	**0.4107 ± 0.0288**	**0.3922 ± 0.0263**	**0.5870 ± 0.0320**	**0.5766 ± 0.0297**	**0.4025 ± 0.0289**	**0.2021 ± 0.0302**	**0.3357 ± 0.0291**
GLMNet	0.4080 ± 0.0273	0.3903 ± 0.0283	0.5843 ± 0.0323	0.5733 ± 0.0323	0.4003 ± 0.0314	0.1944 ± 0.0366	0.3316 ± 0.0327
Our Encoder	0.4029 ± 0.0265	0.3876 ± 0.0239	0.5743 ± 0.0283	0.5650 ± 0.0251	0.3879 ± 0.0237	0.2018 ± 0.0235	0.3259 ± 0.0233

to the same video are first averaged across the three sessions at the signal level before decoding. We conduct this comparison using ShallowNet, GLMNet, and our encoder as representative architectures.

As shown in Table 8, signal-level session averaging consistently improves entity recognition across all metrics and encoders. The gains are especially evident in Macro AP and set-based metrics, suggesting that aggregating repeated responses improves the recoverability of entity-related information. The consistent trend across architectures further indicates that this benefit is not specific to a particular encoder. These improvements are consistent with the expected reduction of trial- and session-specific variability through repeated-response averaging.

Table 8: Effect of signal-level session averaging on entity recognition. Single-session results are obtained by evaluating each recording session independently and averaging the resulting metrics across the three sessions.

EEG Encoder	Setting	Macro AP ↑	Micro AP ↑	Macro AUC ↑	Micro AUC ↑	Top-2 F1 ↑	Set Acc. ↑	Jaccard ↑
ShallowNet	Single-Session	0.3750 ± 0.0188	0.3626 ± 0.0181	0.5304 ± 0.0234	0.5211 ± 0.0228	0.3590 ± 0.0192	0.1458 ± 0.0185	0.2879 ± 0.0187
ShallowNet	Session-Average	**0.3980 ± 0.0267**	**0.3738 ± 0.0255**	**0.5447 ± 0.0307**	**0.5440 ± 0.0310**	**0.3693 ± 0.0279**	**0.1624 ± 0.0280**	**0.3003 ± 0.0269**
GLMNet	Single-Session	0.3782 ± 0.0194	0.3653 ± 0.0193	0.5332 ± 0.0219	0.5192 ± 0.0218	0.3511 ± 0.0194	0.1436 ± 0.0195	0.2819 ± 0.0191
GLMNet	Session-Average	**0.3982 ± 0.0297**	**0.3765 ± 0.0282**	**0.5456 ± 0.0348**	**0.5467 ± 0.0347**	**0.3721 ± 0.0284**	**0.1635 ± 0.0313**	**0.3026 ± 0.0288**
Our Encoder	Single-Session	0.3787 ± 0.0201	0.3645 ± 0.0196	0.5306 ± 0.0258	0.5282 ± 0.0243	0.3602 ± 0.0218	0.1501 ± 0.0252	0.2845 ± 0.0225
Our Encoder	Session-Average	**0.4004 ± 0.0263**	**0.3704 ± 0.0270**	**0.5388 ± 0.0316**	**0.5497 ± 0.0304**	**0.3748 ± 0.0288**	**0.1725 ± 0.0372**	**0.3107 ± 0.0306**

F Additional Reconstruction Results

F.1 Detailed Quantitative Results

Table 9 provides the detailed reconstruction results corresponding to Table 2 in the main paper, with variability reported in addition to the mean performance. For EEG-based methods, the standard deviation is computed across all evaluated participants. For GT Caption, the mean and standard deviation are computed across three generation seeds. The overall trends are consistent with those reported in the main paper, with EntiRecon achieving stronger semantic and multi-entity fidelity than the compared EEG-based reconstruction methods.

F.2 Ablation Settings

We consider six ablation variants to examine the contribution of different components in EntiRecon. **Text Alignment** directly maps the EEG representation to the pretrained text-conditioning space, without the coarse semantic bottleneck or fine EEG conditioning. **Coarse Only** retains only the coarse semantic branch. **Fine Only** removes both forms of coarse semantic conditioning: the

Table 9: **Detailed video reconstruction results.** Mean ± standard deviation is reported across participants for EEG-based methods and across three generation seeds for GT Caption.

	Semantic Fidelity			Frame Fidelity		Video Fidelity		Multi-Entity Fidelity	
Method	2-way-I ↑	2-way-V ↑	ViFi ↑	SSIM ↑	PSNR ↑	FVD ↓	Warp-LPIPS ↓	Entity F1 ↑	ECE ↓
GT Caption	0.957 ± 0.02	0.982 ± 0.01	0.495 ± 0.03	–	–	–	–	0.895 ± 0.03	0.190 ± 0.06
EEG2Video	0.757 ± 0.02	0.832 ± 0.02	0.258 ± 0.01	0.184 ± 0.03	9.78 ± 0.30	2452.89 ± 16.73	0.408 ± 0.03	0.332 ± 0.03	0.807 ± 0.06
MindCine	0.779 ± 0.01	0.844 ± 0.02	0.261 ± 0.01	0.190 ± 0.02	10.16 ± 0.17	2595.78 ± 103.21	0.411 ± 0.03	0.322 ± 0.05	0.835 ± 0.03
DynaMind	0.747 ± 0.02	0.831 ± 0.01	0.154 ± 0.02	0.148 ± 0.02	**11.72 ± 0.10**	2956.18 ± 71.21	0.458 ± 0.02	0.221 ± 0.03	1.069 ± 0.09
EntiRecon	**0.830** ± 0.03	**0.915** ± 0.02	**0.338** ± 0.03	**0.222** ± 0.03	10.17 ± 0.25	**2005.60** ± 50.68	**0.107** ± 0.02	**0.468** ± 0.04	**0.530** ± 0.03

coarse-guided fine semantic interaction module is disabled, and Tora's native coarse text condition is replaced with the empty-text embedding, leaving only the fine EEG conditioning pathway. **Coarse + Fine** combines both EEG conditioning branches without auxiliary eye guidance. **Coarse + Fine (Shuffled EEG)** uses the Coarse + Fine variant without eye guidance, while shuffling the EEG–stimulus correspondence to evaluate whether reconstruction relies on stimulus-specific neural information. **Full (Single Session)** uses the complete model with one recording session at a time, with performance averaged across the three sessions. Finally, **Coarse + Fine + Eye (Full)** denotes the full EntiRecon model with coarse and fine EEG conditioning, eye-guided fusion, and repeated-session feature aggregation.

F.3 Additional Qualitative Results

In Fig. 9, Fig. 10 and Fig. 11, we present additional reconstruction results across diverse multi-entity compositions. EntiRecon is able to reconstruct a range of dynamic scenes involving different entity combinations, while generally preserving the major entities and their semantic composition.

F.4 Failure Cases

For completeness, we also present representative reconstruction failure cases in Fig. 12. As shown in the examples, the main failures involve incorrect entity semantics and degraded object appearance, indicating that accurately preserving multiple entities remains challenging in some cases.

G Extension to Three-Entity Scenes

We further examine whether models trained exclusively on two-entity scenes can generalize to unseen three-entity compositions. All models are trained and validated on the 468 four-second two-entity videos and directly evaluated on 156 six-second three-entity videos, without using any three-entity data during training. At test time, the complete six-second EEG recordings are processed with adaptive temporal pooling to accommodate the longer input duration.

G.1 Three-Entity Recognition

We compare the **Original** multi-label classification objective used in the main benchmark with a **+ Composition** variant that additionally introduces auxiliary composition-level supervision. Both settings predict over the same six entity categories and use fixed Top-3 decoding. Evaluation is conducted on 20 participants under both session-average and cross-session protocols. As shown in Table 10, the Original objective achieves above-chance performance on unseen three-entity scenes under both evaluation settings, indicating that entity-level information learned from two-entity scenes can transfer to more complex compositions. Adding composition supervision provides little benefit and generally reduces the set-based metrics. In particular, Set Accuracy decreases from 0.1836 to 0.1509 under session-average evaluation and from 0.1313 to 0.1171 under cross-session evaluation, with similar trends in Top-3 F1 and Jaccard. These results suggest that direct multi-label entity supervision generalizes more effectively. Nevertheless, the relatively low set-level accuracy shows that robust generalization to unseen three-entity compositions remains difficult.

Table 10: Entity recognition on unseen three-entity scenes using complete six-second raw EEG. Results are mean ± sample standard deviation across 20 participants. The best mean in each setting is highlighted in bold.

Objective	Macro AP	Micro AP	Macro AUC	Micro AUC	Top-3 F1	Set Acc.	Jaccard
Chance Level	0.5000	0.5000	0.5000	0.5000	0.5000	0.0500	0.3650
Session-Average Evaluation							
Original	**0.5719 ± 0.0354**	**0.5651 ± 0.0408**	0.5648 ± 0.0409	**0.5767 ± 0.0482**	**0.5589 ± 0.0373**	**0.1836 ± 0.0614**	**0.4499 ± 0.0425**
+ Composition	0.5701 ± 0.0308	0.5632 ± 0.0434	**0.5649 ± 0.0370**	0.5739 ± 0.0434	0.5552 ± 0.0281	0.1509 ± 0.0408	0.4386 ± 0.0292
Cross-Session Evaluation							
Original	0.5282 ± 0.0181	**0.5250 ± 0.0238**	**0.5164 ± 0.0204**	**0.5298 ± 0.0240**	**0.5220 ± 0.0169**	**0.1313 ± 0.0282**	**0.4063 ± 0.0193**
+ Composition	**0.5284 ± 0.0144**	0.5191 ± 0.0247	0.5149 ± 0.0171	0.5178 ± 0.0295	0.5128 ± 0.0225	0.1171 ± 0.0307	0.3944 ± 0.0246

G.2 Three-Entity Video Reconstruction

We further evaluate whether a reconstruction model trained exclusively on two-entity scenes can generalize to unseen three-entity compositions. As shown in Table 11, EntiRecon consistently outperforms EEG2Video and MindCine on semantic and multi-entity metrics, achieving the highest 2-way retrieval accuracy, ViFi score, and Entity F1. These results indicate that the semantic representations learned from two-entity scenes can transfer to more complex three-entity compositions.

The qualitative results in Fig. 8 further show that EntiRecon can preserve the major entities and overall scene semantics in several unseen three-entity cases despite being trained only on two-entity videos. However, reconstruction becomes less consistent as scene complexity increases, and some examples still miss or incorrectly reconstruct one of the target entities, which is also reflected by the increased ECE. Moreover, semantic metrics such as 2-way retrieval, ViFi, and Entity F1 may remain relatively high when only part of the target composition is recovered. Overall, these results demonstrate promising generalization to unseen three-entity scenes while highlighting the remaining challenge of consistently reconstructing all target entities.

Table 11: Quantitative comparison on the three-entity OOD test set.

Method	Semantic Fidelity			Frame Fidelity		Video Fidelity		Multi-Entity Fidelity	
	2-way-I ↑	2-way-V ↑	ViFi ↑	SSIM ↑	PSNR ↑	FVD ↓	Warp-LPIPS ↓	Entity F1 ↑	ECE ↓
EEG2Video	0.746	0.827	0.218	0.160	9.78	2065.06	0.440	0.351	1.774
MindCine	0.748	0.823	0.191	0.149	10.17	1949.78	0.454	0.274	1.915
EntiRecon	0.848	0.930	0.363	0.182	9.98	2531.06	0.047	0.538	0.981



Figure 8: **Qualitative reconstruction results on unseen three-entity scenes.** EntiRecon is trained exclusively on two-entity videos and directly evaluated on three-entity compositions. While major scene semantics are often preserved, some target entities may be missing or incorrectly reconstructed.

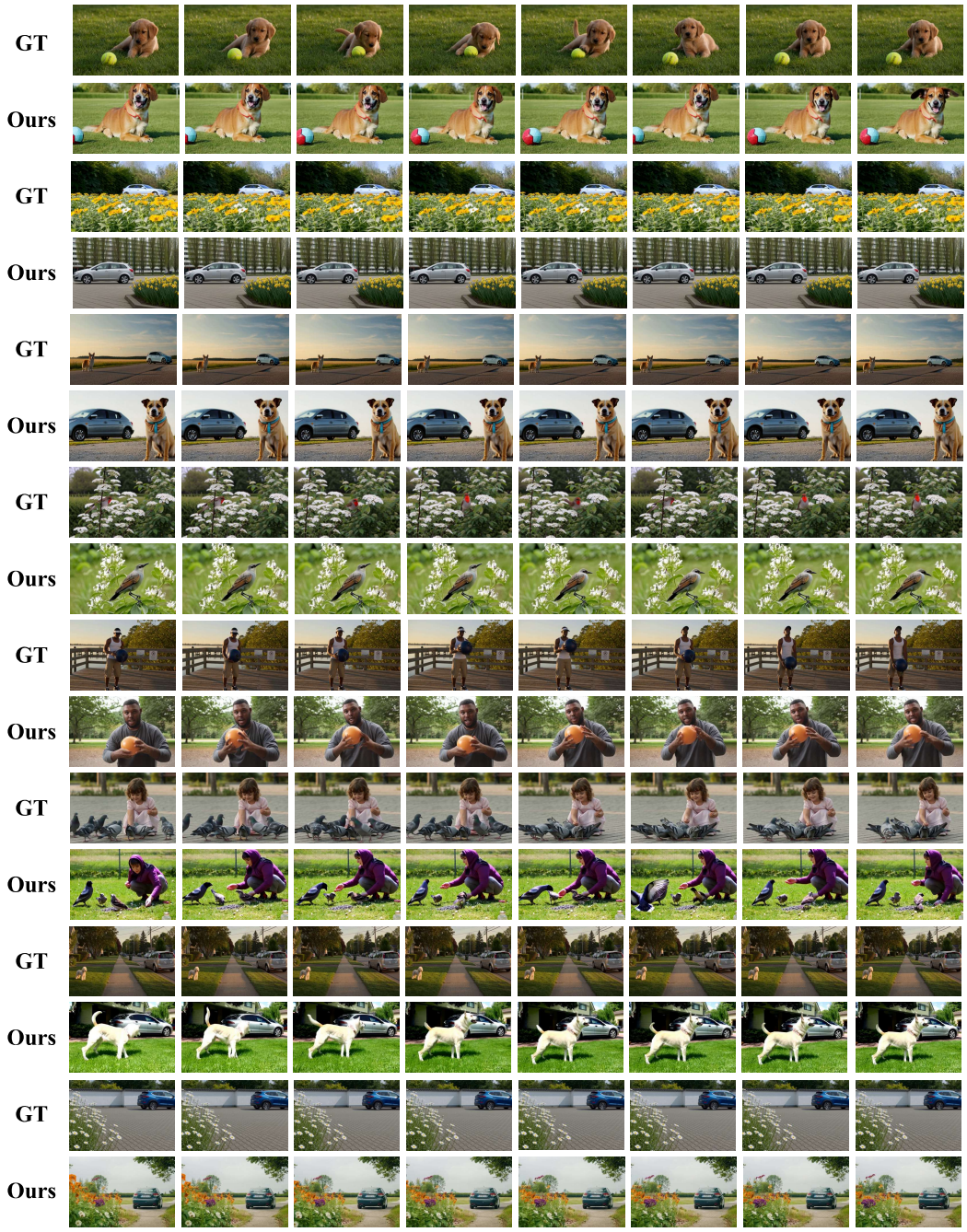


Figure 9: **Additional qualitative reconstruction results.** Comparison between ground-truth videos and reconstructions produced by EntiRecon across different multi-entity compositions.

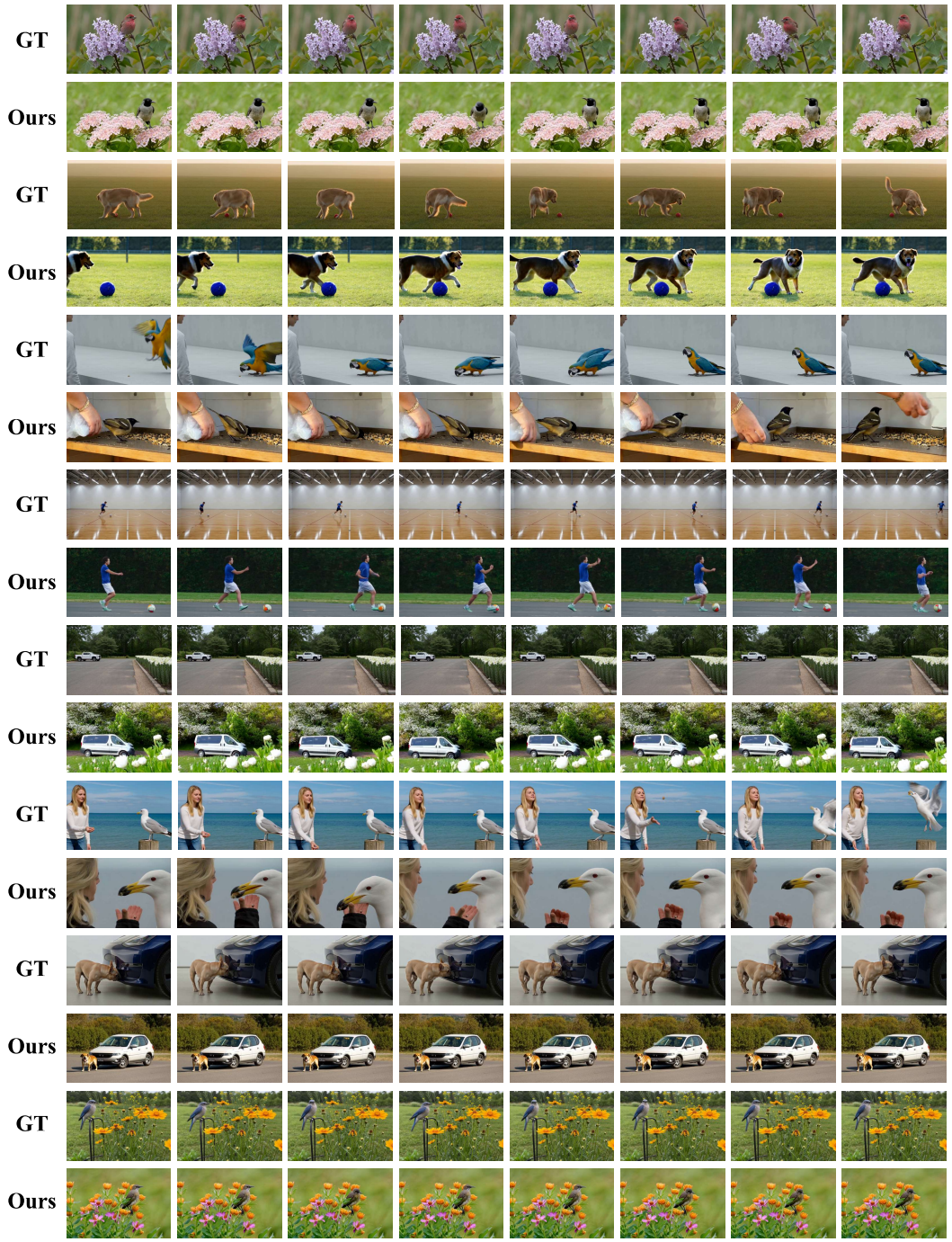


Figure 10: **Additional qualitative reconstruction results.** Comparison between ground-truth videos and reconstructions produced by EntiRecon across different multi-entity compositions.

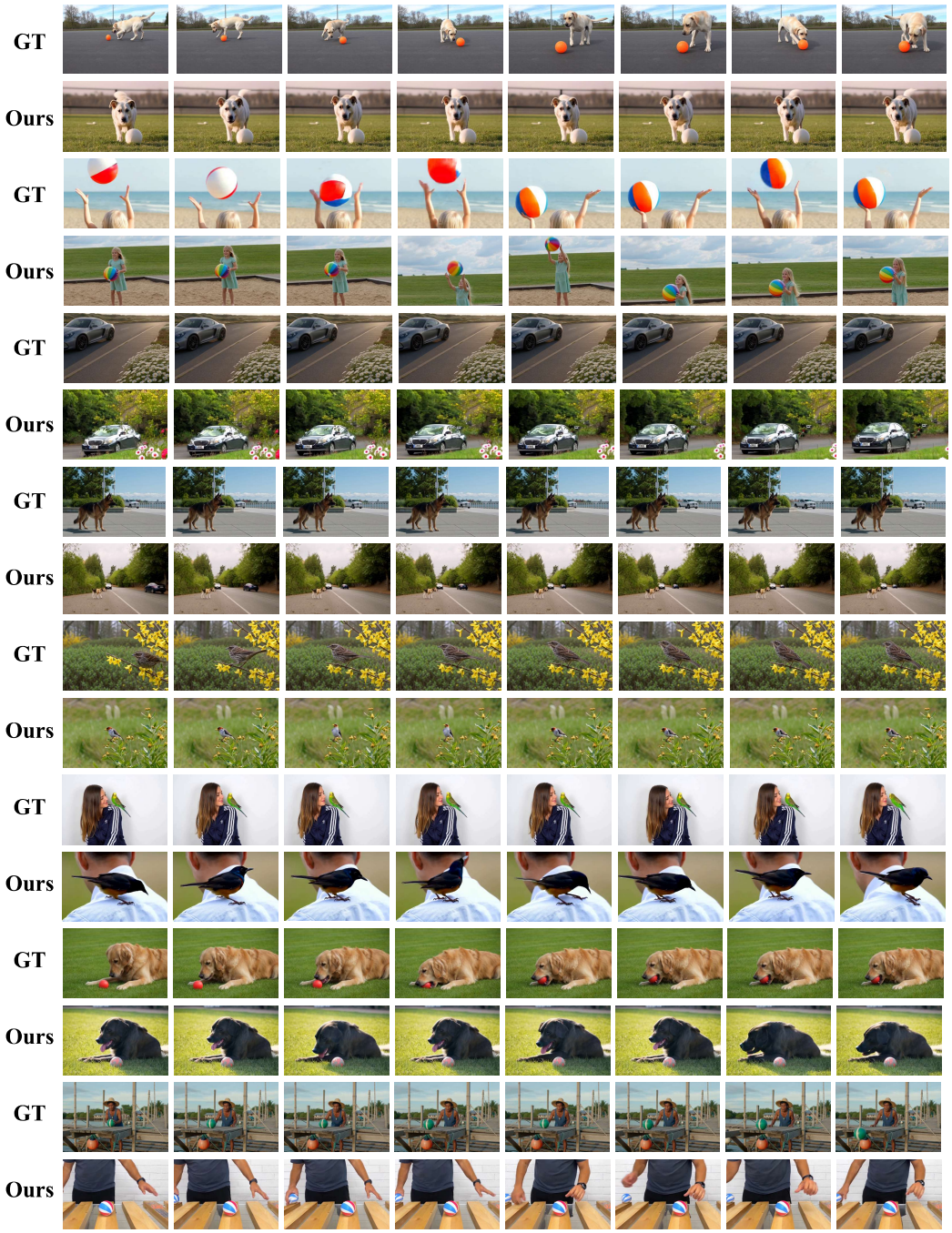


Figure 11: **Additional qualitative reconstruction results.** Comparison between ground-truth videos and reconstructions produced by EntiRecon across different multi-entity compositions.

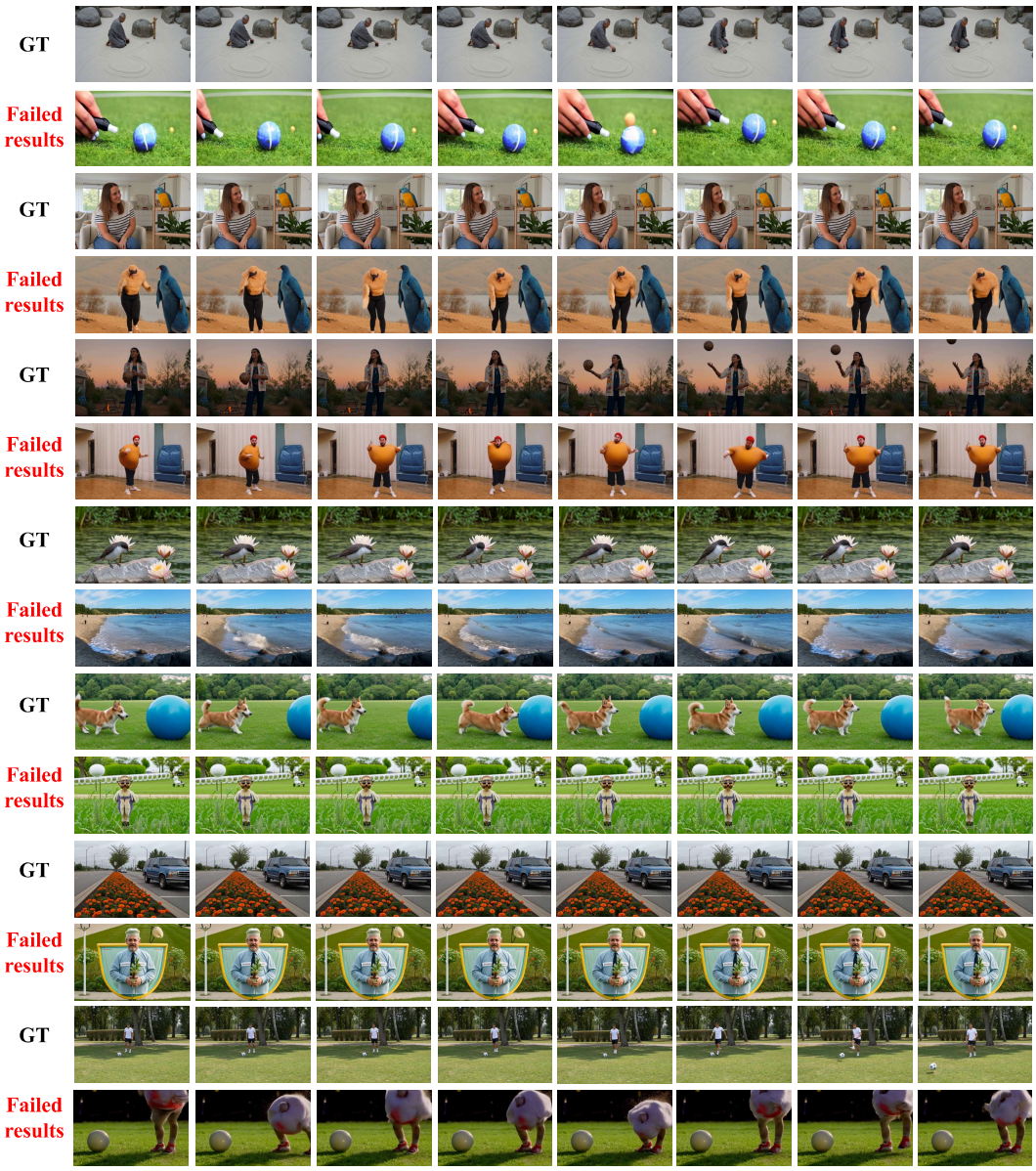


Figure 12: Failure cases of EntiRecon.