npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

Article in Press

https://doi.org/10.1038/s41746-026-03413-8

Visual-affective decoding enhances neurofunctional objective and generalizable signatures in unseen adolescent depression diagnosis

Check for updates

Jingxu Liu[1,6], Ke Liu[2,3,6], Mei Tang[1], Zhuo Xi[4], Hanggai Taorigen[2,3], Yichen Liu[1], Le Xu[1], Dehua Liu[1], Kun Yi[5], Liang Hu[1], Lei Zhu[1], Duoqian Miao[1] ✉, Qi Zhang[1] ✉ & Zhi Yang[2,3] ✉

Adolescent major depressive disorder is a highly prevalent and debilitating condition, yet objective and generalizable neurofunctional signatures remain elusive. Existing neuroimaging approaches largely rely on resting-state fMRI, or controlled tasks often lack ecological validity and generalizable signatures of depression-related neural dysfunctions. This study proposes a Visual–Affective Decoding (VAD) framework integrating cross-subject visual and affective decoding from whole-brain fMRI acquired during naturalistic movie viewing. A dual-pathway anomaly detection model trained exclusively on healthy controls establishes normative brain function and identifies MDD through deviations from this norm. In 200 adolescents (112 MDD, 88 HC), VAD achieved an AUC of 0.87, 95.56% sensitivity, 83.89% accuracy, and 72.22% specificity. Using the same anomaly threshold, it generalized to an independent external adolescent cohort (70 MDD, 68 HC), achieving an AUC of 0.81 with an accuracy of 75.36%, supporting its cross-site generalizability. VAD further revealed interpretable neurofunctional signatures, including blunted emotional reactivity, attenuated visual representation, and affective patterns consistent with emotional rumination. Region-level contribution analyses identified dissociable visual- and affective-sensitive cortical regions, consistent with dysfunctions in visual, default-mode, salience, and higher-order associative networks. These findings demonstrate that VAD provides objective, interpretable, and generalizable neurofunctional signatures for adolescent depression diagnosis and clinically relevant biomarker development.

Adolescent depression represents a critical global public health challenge, contributing profoundly to disability, academic impairment, and increased risk of suicide[1–4]. Its early and accurate identification is paramount for timely intervention and improved lifelong outcomes. However, current clinical diagnosis relies primarily on subjective self-report and behavioral observation, a process vulnerable to bias, stigma, and the inherent complexity of adolescent self-disclosure[5,6]. This fuels diagnostic delays, varied treatment responses, and an urgent need for objective, biologically grounded markers[7,8].

Neuroimaging, particularly functional magnetic resonance imaging (fMRI), has long promised to uncover such biomarkers[9–14]. Large-scale studies utilizing resting-state fMRI (rs-fMRI) have identified differences in brain connectivity between patients with major depressive disorder (MDD) or bipolar disorder (BD) and healthy controls[15–17]. Yet, the resulting functional connectivity signatures often show limited diagnostic specificity and reproducibility across cohorts. Traditional approaches using task-based fMRI (t-fMRI) typically use abstract and experimentally controlled stimuli[18,19], e.g., static images of emotional faces, lacking ecological validity[20].

[1]School of Computer Science and Technology, Tongji University, Shanghai, China. [2]Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders & National Center for Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing, China. [3]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China. [4]International & VIP Health Center, Peking University Shenzhen Hospital, Shenzhen, China. [5]State Information Center, Beijing, China. [6]These authors contributed equally: Jingxu Liu, Ke Liu. ✉e-mail: dqmiao@tongji.edu.cn; zhangqi_cs@tongji.edu.cn; yangz@mail.ccmu.edu.cn

These existing approaches exhibit significant limitations, failing to engage the real-world affective or cognitive processes central to depression. This impasse underscores the necessity for a new paradigm that captures the dynamic, naturalistic essence of affective processing while yielding signatures stable and generalizable enough for clinical application.

Bridging the gap between ecological validity and experimental control, naturalistic paradigm fMRI has emerged as a powerful alternative that exposes participants to dynamic, context-rich stimuli (e.g., visual scenes, narrative events, emotional arcs), improving participant engagement and compliance[21–23]. It engages the brain in continuous, multimodal processing, while reducing emotion artifacts and anxiety often associated with repetitive, abstract tasks, which is particularly beneficial for adolescent and clinical populations[24]. Crucially, naturalistic viewing evokes reliable, time-locked brain activity patterns across subjects, providing a robust shared neural timeline for group-level analysis and cross-cohort validation[25,26]. By leveraging the neural response differences elicited by the same naturalistic stimuli between healthy controls and patients with depression, we capably identify robust, stimulus-driven neurofunctional signatures that reflect depression-specific dysfunctions in cognitive and affective systems[27]. This raises a compelling question: can neurofunctional signatures, derived from brain responses to naturalistic stimuli using fMRI, serve as a generalizable and objective biomarker to accurately distinguish patients with MDD from healthy controls across independent, unseen cohorts?

Affective decoding uses computational models to map patterns of brain activity onto specific emotional features or states elicited by stimuli, naturally contributing to interpretable and effective neurofunctional signatures for depression diagnosis. It primarily focuses on classification tasks utilizing discrete emotion labels[28,29] and regression tasks based on valence (the positivity or negativity of emotion) and arousal (the intensity of emotional activation) scores[30,31]. For valence, researchers have used machine learning algorithms to identify specific neural signatures for emotional responses to aversive stimuli[32] and threat anticipation[33]. For arousal, recent studies leverage shallow computational models to map the neural correlates and representations of t-fMRI to arousal levels[34,35]. Further work has analyzed neural patterns or created emotion atlases in patients viewing emotional stimuli for depression classification, highlighting the complex interplay between emotion regulation and depression. However, affective decoding is fundamentally limited by its reliance on subjective self-reporting, which is vulnerable to recall bias, conscious manipulation, and an individual's inability to accurately articulate their emotional state. This makes it less scalable and objective for early detection compared to methods analyzing observable emotional biomarkers. We aim to introduce a visual-affective decoding framework, bridging this gap by applying advanced computational decoding to extract complementary stimulus-driven functional signatures that are inherently more robust and objective for depression diagnosis.

One major challenge in linking neurofunctional signatures to naturalistic stimuli in adolescent depression diagnosis is the complexity and uncontrolled nature of the stimuli, hindering traditional, hypothesis-driven models from capturing effective correlations. A promising alternative has been the development of advanced computational decoding techniques, which use deep learning techniques to model the intricate relationship between the rich feature space of the naturalistic stimulus and the evolving brain response. Such solution decoding of continuous brain states helps to identify how the brain's functional architecture dynamically configures to process real-world emotional content. Especially, the recent advancements in visual decoding based on t-fMRI have enabled the investigation of complex neural mechanisms underlying visual processing, object recognition, and scene understanding[36,37], offering superior objectivity and early detection potential by analyzing subtle, informative visual cues. Existing visual decoding studies[37–42], e.g., fMRI-to-video reconstruction[43,44], identify state-specific brain activation patterns in response to emotionally salient or cognitively demanding stimuli, promisingly yielding enhanced biomarkers that are more mechanistically linked to core depressive symptoms. Coupling the decoding techniques, naturalistic paradigms provide a unique window into the systems-level dysfunction in psychiatric disorders, capturing not just where the brain is atypical, but how it processes affectively salient, real-world information—a critical step towards objective biomarkers.

Critically, most existing depression diagnosis models are developed and validated on relatively homogeneous subjects, suffering a drastic performance decline when applied to unseen populations, a failure of generalizability that is highly critical for real-world clinical feasibility. The limitation also fundamentally limits the cross-subject generalizability of t-fMRI-based brain decoding[45], lying in the high individual variability of brain activity across subjects[46]. Fortunately, with the open releases of deep learning models such as CLIP[47] and Stable Diffusion[48], beginning with MindEye2[49] in 2024, cross-subject brain decoding models such as MindTuner[50], BrainGuard[51] and MindAligner[52] have demonstrated the ability to reconstruct visual stimulus with minimal data from new subjects. Notably, for video stimuli reconstruction, MinD-Video[43] conditions a diffusion model on visual fMRI features, while NeuroClips[44] achieves remarkable performance improvements through a fine-grained dual design for semantics and perception reconstruction. More recently, a groundbreaking study achieved fully zero-shot brain-to-image decoding[53], marking a significant step toward generalizable neural decoding. Building upon these advances, we capably apply cross-subject visual decoding frameworks to detect depression and seek to leverage the decoding process of t-fMRI to explore a wider range of downstream tasks related to cognition and emotion. This would benefit inferring neural biomarkers associated with depression and investigating model-inferred neurofunctional associations through visual-based assessments.

In this context, as illustrated in Fig. 1, we propose a framework for diagnosing adolescent depression that integrates coupled Visual and Affective Decoding (VAD) within a naturalistic paradigm using t-fMRI. The VAD framework is trained exclusively on a healthy cohort and effectively distinguishes unseen subjects with MDD from healthy controls. The visual decoding pipeline aligns upsampled fMRI data with video stimuli. We employ residual multilayer perceptron (MLP) networks to embed the fMRI data, utilizing contrastive learning to synchronize these embeddings with a visual latent space shaped by CLIP. Concurrently, the affective decoding pipeline downsamples the fMRI embeddings through average pooling to match the temporal resolution of the video clips. It predicts the affective features of these clips—specifically valence and arousal—using a single-layer MLP. By combining the video reconstruction embeddings with the predicted emotional embeddings, our framework learns complementary visual and affective information, leading to richer neurofunctional signatures for distinguishing healthy controls from subjects with MDD. During inference, for an unseen cohort watching the same video clips, the fMRI data are preprocessed identically and passed through the trained framework. Following an anomaly detection paradigm, we compute the deviations of the fMRI embeddings (projected into the CLIP visual space via the shared MLP backbone) and assess anomalies in the predicted valence and arousal for each clip. A fixed diagnostic threshold is then applied to the subject-level overall anomaly score to identify subjects with MDD versus healthy controls. In the present study, this threshold was determined in the internal development cohort by maximizing the Youden[54] index and was subsequently kept fixed for external validation.

Furthermore, to explore the differences between subjects with MDD and healthy controls in terms of visual perception reconstruction, we train a diffusion prior that remaps fMRI embeddings into the CLIP image space, following MindEye2. This enables us to reconstruct high-quality, high-fidelity video frames, facilitating downstream analyzes that provide interpretable, multi-perspective insights into the model-inferred neurofunctional alterations associated with depression and may inform clinical interventions. To further assess the robustness and cross-site generalizability of the proposed framework, we additionally evaluate its diagnostic performance in an independent external adolescent cohort acquired at a separate site. This study also emphasizes the identification of neurofunctional signatures for MDD. We conduct ablation studies to assess the contribution of individual brain regions to the anomaly scores in MDD patients, thereby

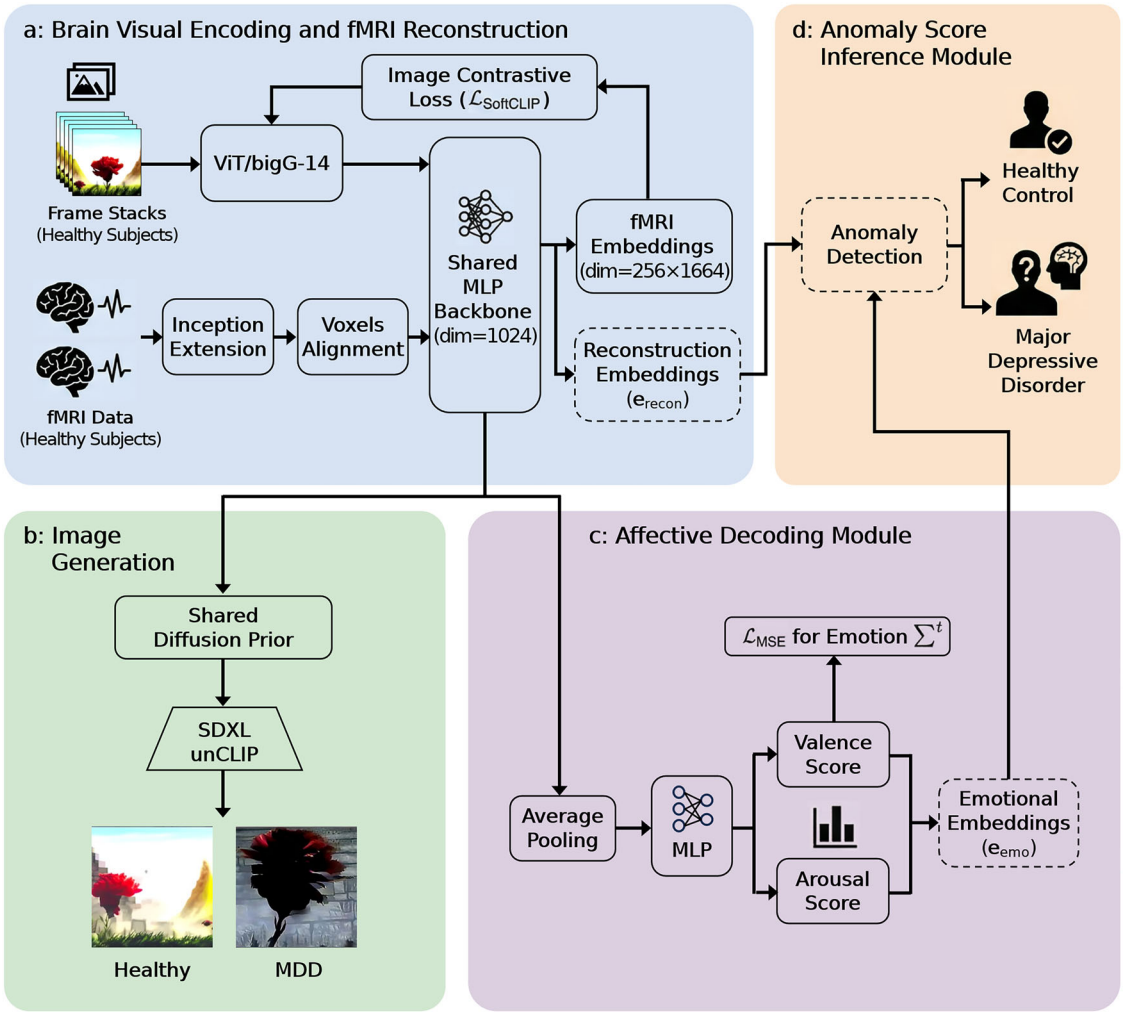


Fig. 1 | Schematic overview of the VAD framework. a Brain visual encoding and fMRI reconstruction module. Video clips and fMRI data are resampled to a consistent temporal dimension through the Inception Extension module. Voxel-level alignment is then achieved via 1D convolution combined with positional encoding. The aligned fMRI representations are passed through a shared MLP backbone containing four residual blocks, aligning fMRI-derived features with visual frame embeddings extracted by the ViT-bigG/14 encoder within a shared latent space. **b** Image generation module. The processed fMRI embeddings are mapped through a diffusion prior and reconstructed via the SDXL unCLIP model, enabling high-fidelity video frame reconstruction. This provides an interpretable foundation for visual reconstruction comparisons between MDD and HC groups. **c** Affective decoding module. After temporal alignment, a single-layer MLP network predicts the valence and arousal scores for each video clip, yielding emotional embeddings consistent with true emotional scores. **d** Anomaly score inference module. For unseen subjects, the pretrained model estimates the visual anomaly score (e_{recon}) based on contrastive loss deviation and the emotional anomaly score (e_{emo}) via mean-squared error. The clip-level abnormalities are averaged and weighted to compute a subject-level overall anomaly score (e), which serves as the classification criterion distinguishing HC from MDD patients. The dashed module in the figure denotes an anomaly detection component, which calculates anomaly scores arising during inference on unseen subjects.

pinpointing candidate brain regions for future hypothesis-driven validation. These findings are further validated by comparing them against large-scale meta-analytic depression brain maps from prior research[55–57], reinforcing the neurobiological plausibility of our proposed model.

Results

Participant characteristics

For the internal training and validation cohort, the study included 220 adolescent recruited participants, 124 diagnosed with MDD and 96 healthy controls (HC) (Table 1). Among them, 12 MDD and 8 HC were excluded due to incomplete clinical scale data or missing task-fMRI scans during the acquisition session. Ultimately, a final cohort of 200 adolescent participants, 112 MDD and 88 HC, was included in the final analyzes. In the external validation cohort, 70 MDD and 68 HC adolescent participants who viewed the same naturalistic movie stimulus with complete clinical data and acceptable head-motion metrics were included.

The demographic and clinical characteristics of the internal training and validation cohort are summarized in Table 1. Participants in both the MDD and HC groups were aged between 12 and 14 years, representing a critical developmental period marked by significant biological, psychological, and social changes. This age range was specifically chosen to focus on this vulnerable stage of adolescence, during which depressive symptoms may manifest differently and more severely, potentially leading to long-term mental health consequences if left untreated. The proportion of females was 82.14% and 45.98% in the MDD and HC groups, respectively, with significant differences observed between groups ($\chi^2 = 28.24$, $p < 0.001$). The potential effects of demographic differences, such as age and sex, on the experimental results were further examined through correlation analyzes and will be discussed in Supplementary Table 1. Compared with the HC group (mean = 1.13, SD = 2.06), the MDD group exhibited significantly higher HAMD-17 scores (mean = 23.24, SD = 5.46, $p < 0.001$) as well as higher HAMA scores (HC group: mean = 1.43, SD = 2.20; MDD group:

Table 1 | Demographic and clinical information of internal cohort

	HC (88)	MDD (112)	*p* value (two-tailed)
Age (median, range)	13 (12–14)	14 (12–14)	0.493
Sex (male/female)	48/40	20/92	<0.001*, $\chi^2 = 28.24$
HAMD-17	1.13 ± 2.06	23.34 ± 5.46	<0.001*
HAMA	1.43 ± 2.20	25.71 ± 8.00	<0.001*

*Results indicate statistical significance at an error rate of 5%. *HAMD-17* 17-item Hamilton Depression Rating Scale, *HAMA* Hamilton Anxiety Scale.

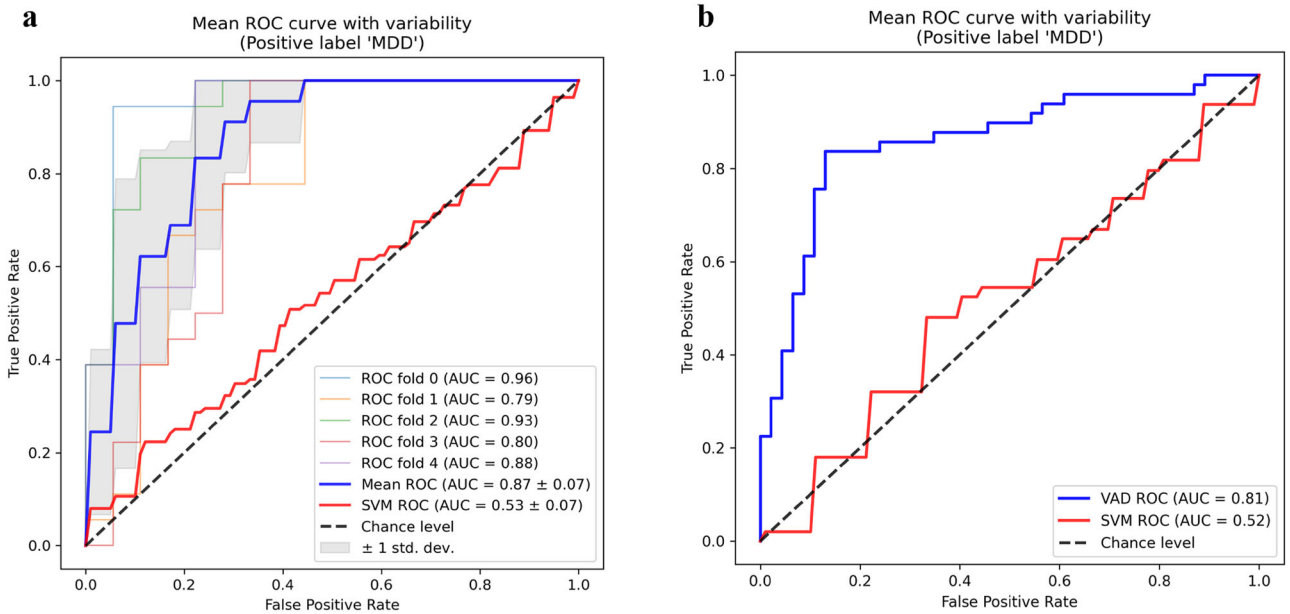


Fig. 2 | Performance evaluation of the VAD. a The ROC (Receiver Operating Characteristic) of the 5-fold cross-validation. The AUC is 0.87, *accuracy* is 83.89%, sensitivity is 95.56%, and specificity is 72.22%. For comparison, the ROC curve of a linear kernel SVM classifier is shown (thick red line), with an AUC of 0.53, highlighting how our advanced model significantly improves classification accuracy compared to a simple linear model for distinguishing HC and MDD. **b** The ROC on the external validation cohort from a separate center. The AUC is 0.81, accuracy is 75.36%, sensitivity is 81.43%, and specificity is 69.12% using the same fixed threshold ($t = 13.135$) determined in the internal cohort. The AUC of the same linear kernel SVM classifier is 0.52 for comparison.

mean = 25.71, SD = 8.00; $p < 0.001$). Details of demographic and clinical characteristics for the 68 HC and 70 MDD patients in the external validation cohort are provided in Supplementary Table 2.

Model performance

Before model evaluation, we first characterized the naturalistic fMRI responses by a whole-brain ROI-level leave-one-out inter-subject correlation (ISC) analysis to further examine whether this inter-individual similarity differed between HC and MDD groups (see Supplementary Table 3).

In the internal training and validation cohort, using a random fivefold cross-validation strategy, the model was trained on the HC group and evaluated on the held-out HC and MDD groups in each fold. During inference, the trained VAD framework generated visual anomaly scores, emotional anomaly scores, and combined anomaly scores for each held-out participant. A binary diagnostic label was then assigned by applying the internally determined Youden threshold ($t = 13.135$) to the combined anomaly score, with participants classified as MDD when $e \geq 13.135$ and as HC otherwise. The area under the curve (AUC) of the prediction model was 0.87, with a sensitivity of 95.56%, specificity of 72.22%, and accuracy of 83.89% (Fig. 2a). To further characterize the dependence of accuracy, sensitivity, and specificity on the anomaly-score threshold, a threshold sensitivity analysis for the internal cohort is shown in Supplementary Fig. 1a, and the corresponding numerical values are provided in Supplementary Table 4. The distributions of the combined anomaly score, visual anomaly score, and emotional anomaly score across folds are shown separately for HC and MDD in Supplementary Fig. 2.

In contrast, we employed a linear kernel Support Vector Machine (SVM) model to directly classify HC and MDD groups, using the same fivefold cross-validation setup. As shown by the thick red line in Fig. 2a, the AUC of this model was only 0.53, indicating that a simple linear classification model is unable to effectively distinguish between HC and MDD. In comparison, our model, leveraging a meticulously designed pipeline and more advanced and complex neural network architecture, significantly improved classification performance.

To further evaluate the generalizability of our method, we tested the model on an independent external validation cohort from a separate clinical site. Using the same fixed threshold ($t = 13.135$) determined in the internal cohort, the VAD model achieved an AUC of 0.81, with an accuracy of 75.36%, sensitivity of 81.43%, and specificity of 69.12% (Fig. 2b, SVM AUC = 0.52). Despite differences in acquisition sites and MRI acquisition parameters between the internal and external cohorts, the model maintained strong and comparable predictive performance, providing further evidence that it can generalize across sites and capture clinically relevant neurofunctional patterns. (The threshold sensitivity analysis for the external cohort is shown in Supplementary Fig. 1b, and Supplementary Table 4.)

Abnormal affective patterns in MDD

Given prior research suggesting the Emotional Context Insensitivity (ECI) hypothesis[57], which posits that depression is associated with a generalized

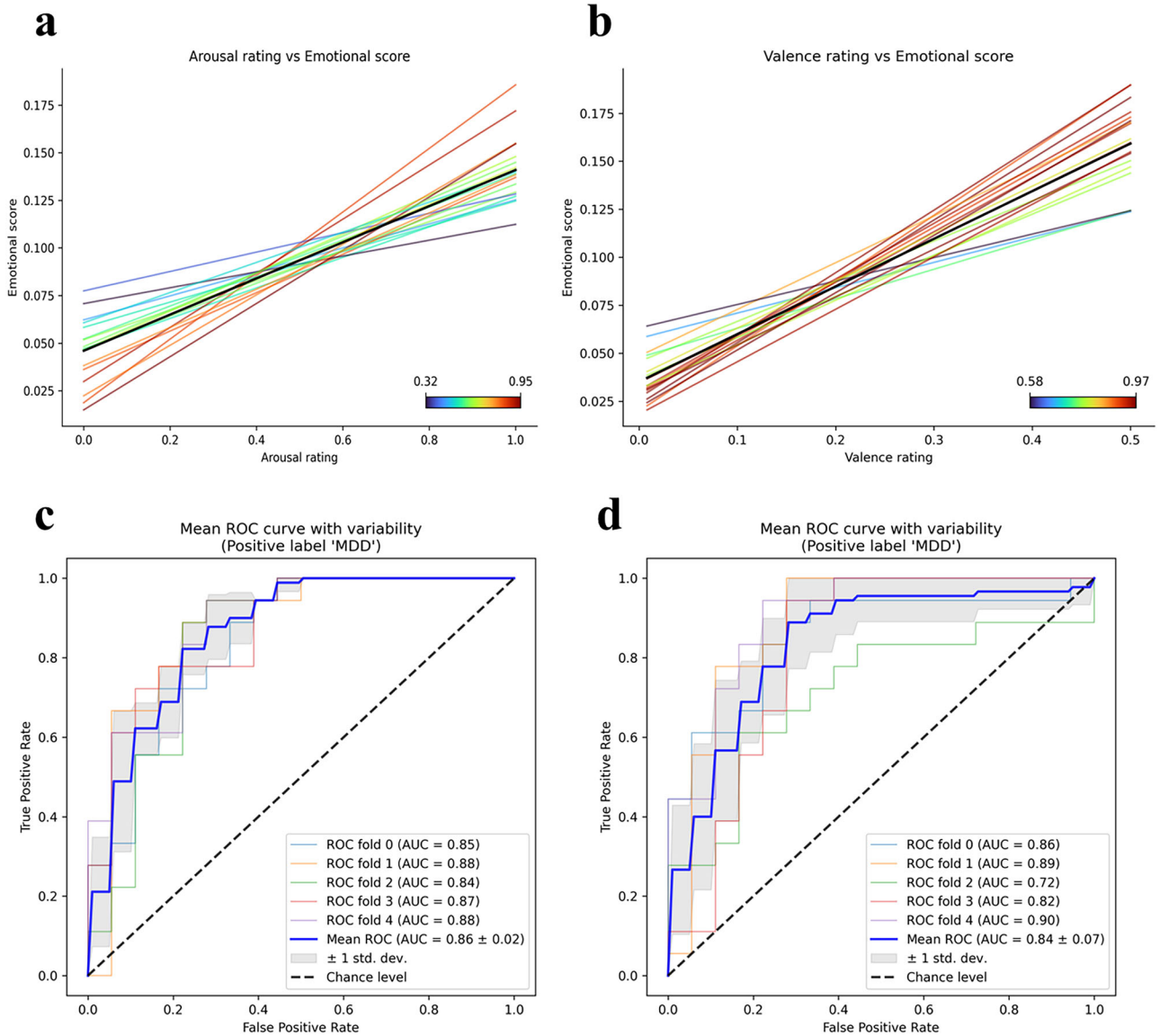


Fig. 3 | Analysis of abnormal affective patterns in MDD. a, **b** Illustrate the fitting results between predicted emotional anomaly scores and actual valence and arousal scores in MDD patients, respectively. Colored lines represent subject-level linear regression fits, while the black line denotes the group-level fit across the entire MDD cohort. **c**, **d** Depict model performance using the standard deviation of valence and arousal scores from each subject as classification features. Model performance was also evaluated with fivefold cross-validation, and the ROC curves were used to quantify classification accuracy. Using valence standard deviation as the feature achieved an AUC of 0.86, accuracy of 81.67%, sensitivity of 83.37%, and specificity of 80.00%. Using arousal standard deviation as the feature achieved an AUC of 0.84, accuracy of 81.31%, sensitivity of 86.87%, and specificity of 76.77%. Both correlations remained significant after BH-FDR correction for two parallel tests.

blunting of emotional reactivity across all affective stimuli, our study aimed to empirically validate this perspective. Specifically, we fitted the ground-truth valence scores (ranging from 0 to 1, with 0.5 considered neutral; the deviation from 0.5 was defined as the *x*-axis Valence score) against emotional anomaly scores predicted by our model. We observed a significant within-subject correlation between the two measures (Pearson's $r = 0.724 \pm 0.013$ standard error (SE), uncorrected $p = 0.009$, FDR-corrected $q = 0.018$; Fig. 3a), indicating that when MDD patients viewed video clips with highly positive or highly negative valence, their emotional responses were more attenuated, shifting closer to neutrality—consistent with the ECI hypothesis.

Furthermore, in the general affective space, arousal and valence are known to form a V-shaped relationship, where arousal increases with both highly positive and highly negative valence[58,59]. We therefore examined arousal scores as an alternative indicator and found that they were also significantly correlated with emotional anomaly scores (within-subject Pearson's $r = 0.545 \pm 0.024$ SE, uncorrected $p = 0.028$, FDR-corrected $q = 0.028$; Fig. 3b). This finding further supports that MDD patients exhibit emotional blunting and suppression particularly under high-arousal, emotionally intense conditions.

In addition, we identified the top five video clips with the highest mean emotional anomaly scores among MDD patients (Supplementary Fig. 3). Consistent with the aforementioned findings, all of these clips were characterized by either highly positive or highly negative valence.

To further validate the aberrant emotional patterns observed in MDD patients, we quantified each subject's emotional variability as the standard deviation of predicted valence or arousal across all 46 video clips. Compared with healthy controls, MDD patients exhibited significantly reduced variability in both valence and arousal. When using the valence standard deviation of each subject as a classification feature, the model achieved an AUC of 0.86, with a sensitivity of 83.37%, a specificity of 80.00%, and an accuracy of 81.67% (Fig. 3c). Similarly, when using the arousal standard

Table 2 | The comparative results of brightness and color features in the reconstructed images between HC and MDD

	HC	MDD	*t*	*p* value (two-tailed)
Brightness	0.261 ± 0.039	0.217 ± 0.027	4.189	3.108×10^{-4}*
R	0.278 ± 0.041	0.223 ± 0.030	4.228	2.662×10^{-4}*
G	0.260 ± 0.040	0.217 ± 0.028	4.082	4.214×10^{-4}*
B	0.244 ± 0.033	0.203 ± 0.020	4.574	1.208×10^{-4}*

deviation as the classification feature, the resulting AUC was 0.84, with a sensitivity of 86.87%, a specificity of 76.77%, and an accuracy of 81.31% (Fig. 3d).

Visual decoding feature results

To further investigate the visual decoding characteristics of MDD patients, we conducted a comparative analysis of image reconstruction quality between MDD and HC. Specifically, we examined the reconstructed visual outputs in terms of brightness (computed as the weighted sum of RGB values) and the individual color channels (R, G, and B). Independent-samples *t*-tests were performed to assess group differences.

As summarized in Table 2, MDD patients exhibited significantly reduced reconstructed image brightness (mean ± SD = 0.217 ± 0.027) compared with HC (0.261 ± 0.039; $t = 4.189$, uncorrected $p = 3.108 \times 10^{-4}$, FDR-corrected $q = 4.144 \times 10^{-4}$). Similarly, significant reductions were observed across all color channels—R ($t = 4.228$, uncorrected $p = 2.662 \times 10^{-4}$, FDR-corrected $q = 4.144 \times 10^{-4}$), G ($t = 4.082$, uncorrected $p = 4.214 \times 10^{-4}$, FDR-corrected $q = 4.214 \times 10^{-4}$), and B ($t = 4.574$, uncorrected $p = 1.208 \times 10^{-4}$, FDR-corrected $q = 4.144 \times 10^{-4}$), which indicates a generalized attenuation in luminance and chromatic representation during visual reconstruction in MDD (Fig. 4a).

To mitigate potential confounding effects of stimulus-level brightness variability, we further performed a stimulus-brightness-stratified control analysis. All extracted video frames were divided into low-, medium-, and high-brightness groups according to their original luminance, and HC–MDD comparisons of reconstructed brightness and RGB values were repeated within each stratum. As shown in Supplementary Table 5, MDD patients consistently showed lower reconstructed brightness and RGB channel values across all brightness strata, and all effects remained significant after BH-FDR correction. This suggests that the observed attenuation of reconstructed visual features was not solely driven by differences in stimulus brightness.

Subsequently, we randomly selected representative frames and compared the reconstructed images from a subset of HC and MDD in a qualitative manner (Fig. 4b). This visual comparison provided further validation of the aforementioned findings by revealing differences in reconstruction results between the two groups.

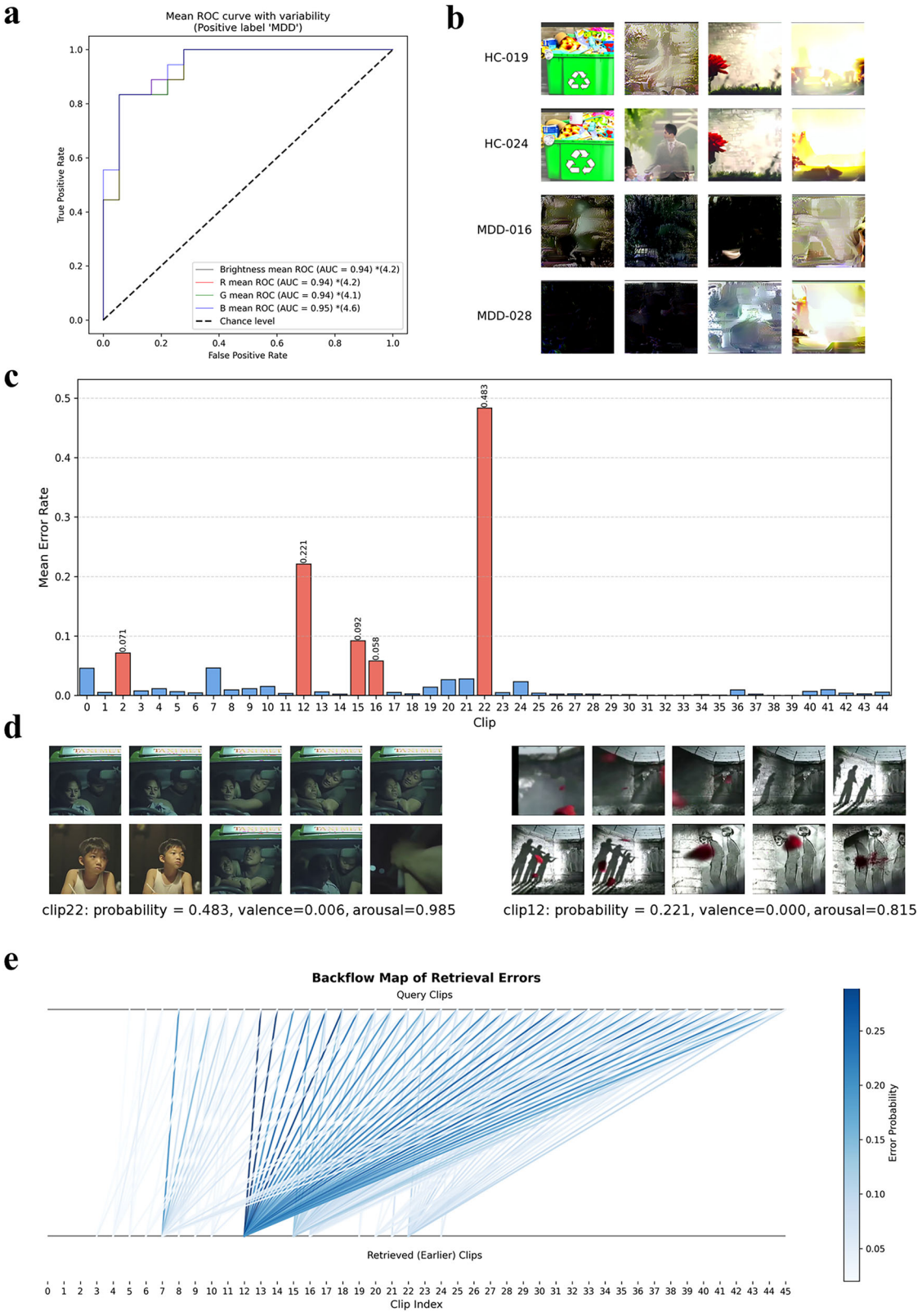
In addition, to examine the impaired visual representation and potential perseverative processing in MDD patients, we conducted an image retrieval analysis using reconstructed image features. The task involved reconstructing images corresponding to each second of a 460-s video (one middle frame per second) and comparing their CLIP feature embeddings with those of the original video frames using cosine similarity. For each time point, only the frames preceding the current one were included in the retrieval pool, allowing us to assess whether MDD patients exhibited signs of perseverative cognition or ruminative thinking, which means a tendency to remain influenced by prior emotional stimuli rather than the current perceptual input[60]. This retrieval-based measure was interpreted as an indirect index of retrospective visual-affective persistence, rather than a direct behavioral or subjective measure of rumination.

The average top-1 retrieval accuracy across all 460 reconstructed images was significantly lower in MDD patients (0.222 ± 0.048) than in HC (0.378 ± 0.086), indicating impaired alignment between neural representations and external visual inputs. To further examine whether retrieval errors showed a pattern consistent with affective perseveration, we analyzed the distribution of the retrieved frames when mismatches occurred. Given that our model incorporates affective features in 10-s clips, we examined which prior clips were most frequently retrieved during errors. Notably, clip22 (probability = 0.483, valence = 0.005, arousal = 0.985) and clip12 (probability = 0.221, valence = 0.001, arousal = 0.815) showed markedly higher retrieval probabilities than other clips (Fig. 4c). These two clips corresponded to the most negative emotional stimuli in the video. This pattern suggests that MDD patients may be more likely to remain influenced by prior negative affective content during visual reconstruction, a pattern consistent with perseverative cognition or emotional rumination. The ten representative frames of these clips within the retrieval pool are shown in Fig. 4d. To further quantify this effect, we compared clip-level retrospective retrieval error rates between HC and MDD groups for clips with an MDD error rate greater than 0.01. As shown in Supplementary Table 6, MDD group showed consistently higher clip-level retrospective retrieval error rates than HC group, and a Wilcoxon signed-rank test confirmed that the error rates were significantly lower in HC than in MDD across these clips ($p = 1.240 \times 10^{-3}$). In addition, across clips, the MDD retrieval error rate was negatively correlated with valence ($r = -0.365$, $p = 0.014$) and positively correlated with arousal ($r = 0.333$, $p = 0.025$), suggesting that more negative and more emotionally intense clips were more likely to be retrospectively retrieved during mismatch trials.

Furthermore, we quantitatively analyzed, across all MDD patients, the average misretrieval rate of frames in each clip with respect to the preceding frames to which they were incorrectly matched, and visualized these results as a backflow map of retrieval errors (Fig. 4e). This analysis was designed to characterize the temporal structure of retrieval errors and to identify whether specific prior emotional contexts exerted a disproportionate influence on later reconstruction errors. As shown in Fig. 4e, several subsequent clips are markedly affected by earlier clips, such as clips 12 and 22, which can be directly observed from the backflow patterns.

Taken together, these retrieval-based findings suggest a possible data-driven signature of affective perseveration in MDD during naturalistic viewing. However, standard experimental studies of rumination typically rely on dedicated rumination-induction or rumination-versus-distraction paradigms, in which participants are explicitly instructed to focus on their feelings, self-related thoughts, or symptom-related meanings, often together with subjective ratings or neural measures of the induced state[61,62]. Because the present study did not include a dedicated rumination induction paradigm, in-scanner thought probes, or post-clip subjective rumination ratings, these results should be interpreted as indirect evidence consistent with emotional rumination rather than as a direct measurement of rumination itself.

Given that emotional information is tightly embedded in visual representations during naturalistic perception[63,64], we further examined the relationship between visual and emotional anomaly scores in MDD group. Specifically, we performed clip-wise Pearson correlation analyzes between visual anomaly scores and emotional anomaly scores across MDD patients, followed by FDR correction across the 46 clips. As shown in Supplementary Table 7 and Supplementary Fig. 4, the two anomaly scores showed generally positive correlations across clips, with correlation coefficients ranging from $r = 0.1798$ to $r = 0.7343$ and an average correlation of $r = 0.4946$. After FDR correction, 44 of 46 clips remained significant. The remaining two clips, clip 9 and clip 39, did not reach the corrected significance threshold but still showed trend-level associations, with FDR-corrected q values of 0.0900 for both clips. These findings suggest that visual and emotional abnormalities are closely related during naturalistic viewing, consistent with prior evidence that emotional schemas are embedded in visual representations. However, the correlations were moderate rather than redundant, indicating that the two branches capture partially overlapping but non-identical dimensions of MDD-related abnormalities. This supports the use of a dual visual–affective decoding pipeline, which allows visual and emotional deviations to be estimated separately while preserving their complementary contribution to the overall anomaly score.



Correlation results between anomaly scores and HAMD-17 clinical measures

To further investigate the clinical relevance of the model-derived anomaly scores, we examined their relationship with the Hamilton Depression Rating Scale (HAMD-17) scores among MDD patients. A significant positive correlation was observed between the predicted anomaly scores and the total HAMD-17 scores ($r = 0.527$, uncorrected $p = 0.028$, FDR-corrected $q = 0.047$), indicating that participants with more severe depressive symptoms tended to exhibit higher anomaly scores.

Furthermore, based on the standard factorial structure of the HAMD-17, the scale was subdivided into four dimensions: anxiety, depression, insomnia, and somatic symptoms[65]. For each subject, the summed scores of

Fig. 4 | Visual decoding analysis in MDD. a Average model performance using mean luminance and RGB channel intensity (R, G, and B values) of reconstructed images as classification features. Since these features were derived from the visual decoding pipeline, all classification models achieved high discriminative power, with AUC values exceeding 0.90 and t-statistics greater than 4 across all metrics. These results indicate a global reduction in luminance and color representation during visual reconstruction in MDD patients. **b** Reconstruction results for four randomly selected frames from two HC participants (HC-019 and HC-024) and two MDD participants (MDD-016 and MDD-028). **c** Histogram of top-1 retrieval probabilities across video clips during erroneous retrieval trials. Red bars highlight the five clips with the highest retrieval probabilities, reflecting recurrent visual recall patterns during misclassification in MDD represents. **d** The ten frames (one frame per second) of the two clips (clip 12 and clip 22) with the highest error occurrence probability in the error retrieval task, along with their error probability, actual valence and arousal values. **e** The backflow map of retrieval errors. The upper axis represents the query clips (later in the sequence), while the lower axis denotes the preceding clips from which erroneous retrievals occurred. Darker colors indicate a higher Top-1 misrerieval rate, reflecting a stronger tendency for the current frame to be incorrectly matched with earlier visual clips. For each query clip, the top three preceding clips with the highest misretrieval rates are visualized as connecting lines.

items within each factor were computed and correlated with anomaly scores using Pearson correlation analysis. The results revealed correlation coefficients of $r = 0.390$ (uncorrected $p = 0.077$, FDR-corrected $q = 0.079$) for anxiety, $r = 0.718$ (uncorrected $p = 0.010$, FDR-corrected $q = 0.025$) for depression, $r = 0.726$ (uncorrected $p = 0.009$, FDR-corrected $q = 0.025$) for insomnia, and $r = 0.373$ (uncorrected $p = 0.079$, FDR-corrected $q = 0.079$) for somatic symptoms.

Collectively, after controlling for multiple comparisons across the five correlation tests using BH-FDR correction[66], the associations between anomaly scores and total HAMD-17, depression, and insomnia remained significant, whereas the associations with anxiety and somatic symptoms remained non-significant. These findings suggest that the depression and insomnia dimensions of the HAMD-17 exhibit the strongest associations with VAD-derived anomaly scores, implying that the model-derived neurofunctional abnormalities are more prominently related to core affective and sleep-related disturbances in MDD patients.

To further investigate whether the associations between anomaly scores and clinical severity differ across the visual and affective components of the VAD framework, we additionally examined Pearson correlations between HAMD-17 subscale scores and the visual anomaly score and emotional anomaly score separately. BH-FDR correction was applied within each anomaly-type family to control for the false discovery rate (FDR). The specific experimental results for each analysis are presented in Supplementary Tables 8 and 9.

For the visual anomaly score, the strongest associations were observed for the depression subscale ($r = 0.867$, uncorrected $p = 0.003$, FDR-corrected $q = 0.015$) and the insomnia subscale ($r = 0.681$, uncorrected $p = 0.018$, FDR-corrected $q = 0.045$).

For the emotional anomaly score, the insomnia subscale showed a significant association after FDR correction ($r = 0.730$, uncorrected $p = 0.009$, FDR-corrected $q = 0.045$). The moderate positive correlations with the depression subscale ($r = 0.474$, uncorrected $p = 0.039$, FDR-corrected $q = 0.065$) fell short of the corrected significance threshold, yet remained largely consistent in effect magnitude.

Taken together, these findings suggest that the depression and insomnia dimensions of the HAMD-17 exhibit the most robust and consistent associations with both VAD abnormalities, while the insomnia subscale emerges as particularly prominent across all three anomaly score types, reinforcing its central role in the neural abnormalities captured by VAD.

In addition, to further examine whether the association between anomaly scores and clinical severity generalized to the HC group, we performed an exploratory correlation analysis using the available HAMD-17 scores in HC. Because HAMD-17 scores in HC group were very low and exhibited a restricted clinical range, this analysis was intended only to assess whether the patient-level association extended to healthy individuals. No significant correlations were observed between HAMD-17 scores and the visual, emotional, or combined anomaly scores after BH-FDR correction (all $q > 0.98$; Supplementary Table 10). These findings suggest that significant associations between VAD-derived anomaly scores and HAMD-17 severity were observed only among MDD patients, whereas no corresponding associations were detected in HC. This lack of association in HC should be interpreted cautiously, as the HAMD-17 is primarily designed to assess depressive symptom severity and showed strong floor effects in healthy controls, limiting its ability to capture meaningful variability within the non-clinical range.

Neuroscience interpretation results

Finally, to further delineate the cortical regions associated with visual and affective abnormalities in MDD, we conducted a fine-grained parcel-level contribution analysis based on the fsaverage5 cortical surface parcellation comprising 400 parcels across 17 functional networks[67] using the ENIGMA Toolbox for MATLAB[68] (Fig. 5, network-level analyzes are presented in Supplementary Fig. 5). Unlike conventional network-level masking, this analysis quantified how individual parcels modulated branch-specific anomaly scores, allowing us to disentangle the cortical substrates associated with visual decoding and affective decoding, respectively.

Specifically, for each parcel, we measured its contribution to visual and emotional anomaly scores by calculating the change in predicted anomaly scores (Δscore) after selectively masking that parcel while keeping all other cortical regions intact. Thus, the parcel-level contribution score should be interpreted primarily as a model-sensitivity index, reflecting the extent to which the trained VAD model relied on information from a given parcel when estimating visual or affective anomaly scores, rather than as evidence of causal effects between specific brain regions and depression. To further examine whether these model-sensitive parcels also carried neural relevance in the original fMRI signals, we performed a model-independent RSA analysis based solely on raw BOLD responses. The 400 parcels were ranked by ablation-derived contribution scores and divided into four quartiles from low to high contribution. As summarized in Supplementary Table 11, parcels with higher ablation-derived contribution scores showed larger HC–MDD RSA normative deviations, with the strongest effect observed in the highest-contribution quartile. This pattern suggests that the ablation-derived parcel contributions reflect model sensitivity with convergent evidence of neural relevance, but should still be interpreted as associative rather than causal.

As shown in Fig. 5a, the ten parcels exerting the strongest influence on visual anomaly scores were predominantly located within the Visual Peripheral (VisPeri) and Visual Central (VisCent) networks, with additional involvement of regions belonging to the salience/ventral attention system. Figure 5b further visualizes the contribution weights of all parcels across the cortex. These parcels largely correspond to extrastriate and striate visual cortices, which are critical for early and intermediate stages of visual feature extraction, including luminance, contrast, and spatial structure. Prior neuroimaging studies have consistently reported altered visual cortex responsivity and reduced sensory gain in depression, particularly during the processing of emotionally salient or dynamic visual stimuli[55,69]. The prominent contribution of peripheral and central visual parcels in our results aligns with the general pattern of attenuated visual cortical responsiveness reported in depression.

In contrast, the parcels contributing most strongly to emotional anomaly scores were primarily distributed across higher-order associative networks, including the Default Mode Network (DMN), salience/ventral attention networks, and somatomotor–insular regions (Fig. 5c, d). These areas are strongly implicated in self-referential processing, emotional appraisal, interoceptive awareness, and the integration of affective salience with cognitive control[70,71]. Large-scale existing studies of depression have consistently demonstrated hyperactivity and altered functional connectivity

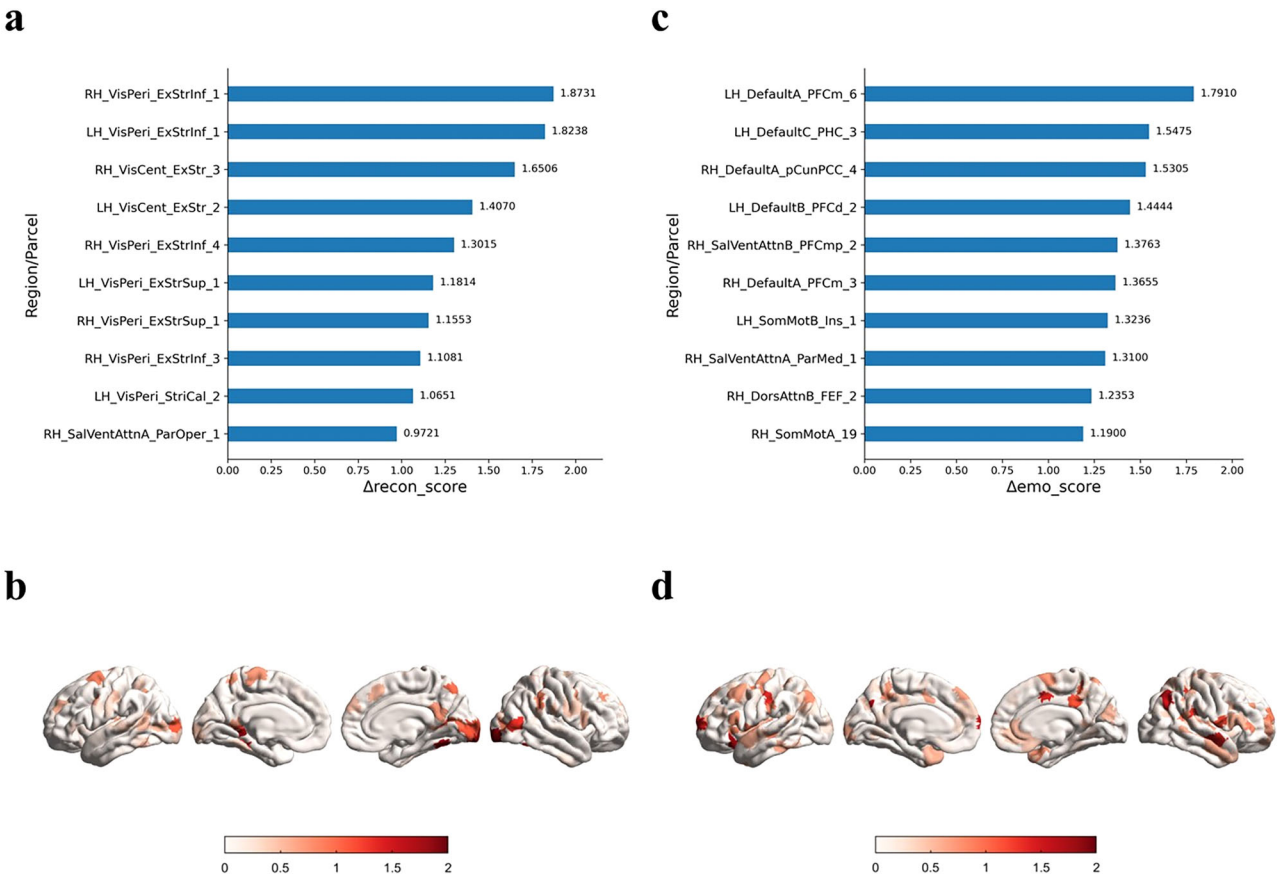


Fig. 5 | Neuroscience interpretation results. a The ten parcels with the strongest contributions to visual anomaly scores, which are predominantly distributed within the Visual Peripheral (VisPeri) and Visual Central (VisCent) networks, with additional involvement of regions from the salience/ventral attention network. **b** The corresponding visual anomaly scores (Δrecon_score) for the parcels shown in (a), visualized on the lateral and medial cortical surfaces. **c** The ten parcels exerting the strongest influence on emotional anomaly scores, primarily located within the default mode network, the salience/ventral attention network, and somatomotor–insular regions. All regional analyses were based on the fsaverage5 cortical template with 400 parcels across 17 functional networks. **d** The corresponding emotional anomaly scores (Δemo_score) for the parcels shown in (**c**), projected onto the lateral and medial cortical surfaces.

within medial prefrontal, posterior cingulate/precuneus, and parahippocampal regions of the DMN, as well as aberrant engagement of salience and insular cortices during emotional processing tasks[56,72,73]. Importantly, the involvement of DMN parcels even during continuous naturalistic viewing likely reflects excessive internally oriented processing and affective self-referential activity, rather than classic resting-state DMN alterations. The prominent contributions of these parcels in our results are consistent with established affective circuits reported in previous studies, suggesting alignment between the affective decoding branch and known neurobiological mechanisms.

Discussion

Addressing the critical need for objective, generalizable, and ecologically valid biomarkers for adolescent depression, this study proposes a Visual–Affective Decoding (VAD) framework grounded in naturalistic t-fMRI. By explicitly modeling how the adolescent brain processes real-world visual and emotional information, VAD directly responds to the limitations of conventional rs-fMRI and abstract task paradigms highlighted in recent neuroimaging literature. Using whole-brain t-fMRI data acquired during naturalistic movie viewing, our model achieved strong performance in distinguishing adolescents with MDD from HC, with an AUC of 0.87, sensitivity of 95.56%, specificity of 72.22%, and an overall accuracy of 83.89%. Moreover, the framework demonstrated robust interpretability, providing valuable insights into the model-inferred neurofunctional alterations associated with depression through the analysis of visual and emotional responses to movie stimuli. In addition to focusing on the identification of MDD neural biomarkers and regions warranting future interventional validation, we conducted ablation experiments to analyze the contribution of different brain regions to MDD-related anomaly scores. This allowed us to validate the neurobiological plausibility of the proposed model by comparing it with large-scale meta-analytic depression brain maps. These findings indicate that VAD excels not only in diagnosing depression but also in offering valuable neurobiological insights into its pathophysiology.

Our approach departs from conventional rs-fMRI-based detection methods, which often rely on graph neural networks to model functional connectivity (e.g., BrainGNN[74], LG-GNN[75]) or on predefined emotional atlases. However, these methods are constrained by their reliance on rs-fMRI and static, predefined regions of interest, limiting their ability to capture the dynamic neural responses to affective stimuli and restricting their generalizability. Such frameworks are further challenged by the pronounced inter-individual heterogeneity and context-sensitive nature of the adolescent brain, which may render adult-derived models inadequate for capturing the unique trajectories of developmental cohorts[76]. To address these limitations, the VAD framework leverages naturalistic stimuli fMRI data and adopts an anomaly detection strategy trained on healthy adolescents, capitalizing on the observation that patients with MDD exhibit altered neural responses during the processing of naturalistic visual and emotional content[77]. The central challenge in utilizing t-fMRI for clinical application is the high inter-subject variability of brain activity[46], which has traditionally impeded the development of generalizable, cross-subject decoding models. Our framework is designed to explicitly overcome this barrier by integrating

two complementary decoding pipelines within a unified anomaly detection architecture, trained exclusively on HC data.

First, we address the visual decoding component (Fig. 1a). Building on recent advances in cross-subject brain decoding such as MindEye2[49] and fMRI-to-video decoding such as NeuroClips[44], we utilized whole-brain fMRI data rather than restricting analysis to the visual cortex, a strategy shown to improve generalization[53]. We employed an Inception Extension module for temporal upsampling and a shared MLP backbone to project fMRI features into a common latent space with visual frame embeddings, which were derived from a pretrained ViT/bigG-14 (CLIP[47]) encoder. This alignment is optimized via a contrastive learning objective (L_{contra}), effectively learning a normative map of healthy visual processing. Moreover, after completing the whole end-to-end pretraining, a diffusion prior module is introduced to further enhance the interpretability of visual reconstruction. This module projects the learned fMRI embeddings back into the CLIP image space, from which high-quality video frames are reconstructed using the SDXL unCLIP model (Fig. 1b). This enables intuitive visual comparison between MDD patients and healthy controls during inference. Second, we integrated a parallel affective decoding pipeline (Fig. 1c). Recognizing that depression is profoundly characterized by affective dysregulation, we repurposed the powerful fMRI embeddings learned by the shared visual backbone. These embeddings were temporally downsampled (via average pooling) to match the resolution of emotional scores and subsequently fed into a dedicated MLP to predict continuous valence and arousal scores. By jointly optimizing for both visual-semantic alignment and affective prediction (L_{emo}), VAD learns a comprehensive, objective and robust representation of healthy neural function. For unseen subjects, the pretrained model estimates visual abnormality (e_{recon}) based on contrastive loss deviation and affective abnormality (e_{emo}) via mean-squared error. The clip-level abnormalities are averaged and weighted to compute a subject-level overall anomaly score (e), to which a fixed threshold determined by the Youden index is subsequently applied for final discrimination between HC and MDD (Fig. 1d). By integrating VAD within an anomaly-based architecture, this VAD framework enhances the objectivity and cross-cohort generalizability of neurofunctional signatures for diagnosing adolescent depression in unseen populations, surpassing approaches based on static connectivity or predefined emotional atlases.

Our results validate the efficacy of this integrated framework. The classification performance in the internal cohort, with an AUC of 0.87, an accuracy of 83.89%, a sensitivity of 95.56%, and a specificity of 72.22%, demonstrates that the model can reliably distinguish unseen adolescent MDD patients from HC based on their neuro-visual and neuro-affective deviations. Notably, the model achieved comparable performance in an independent external cohort, with an AUC of 0.81, an accuracy of 75.36%, a sensitivity of 81.43%, and a specificity of 69.12% using the same fixed threshold ($t = 13.135$) determined in the internal cohort. This external validation provides further evidence for the robustness and cross-site generalizability of the VAD framework in capturing clinically relevant neurofunctional deviations in adolescent MDD. To further assess the stability of the diagnostic decision rule, we performed a threshold sensitivity analysis across a range of overall anomaly-score cutoffs. The internally selected Youden threshold ($t = 13.135$) was located within a relatively stable region of the internal performance curves and maintained comparable diagnostic performance when directly applied to the independent external cohort. Moreover, the cohort-specific Youden threshold estimated from the external data ($t = 13.175$) was close to the internally derived threshold, with only minor differences in accuracy, sensitivity, and specificity between the two operating points (Supplementary Fig. 1, and Supplementary Table 4). These findings support the robustness and transferability of the fixed diagnostic threshold across the two clinical sites under the present naturalistic-viewing paradigm.

Importantly, our framework offers multi-faceted interpretability. The analysis of emotional patterns (Fig. 3) revealed that MDD patients exhibited significantly blunted emotional responses, particularly to high-valence and high-arousal stimuli (Pearson's $r = 0.724$ and $r = 0.545$, respectively). This finding aligns perfectly with the ECI hypothesis[57], indicating a generalized attenuation of emotional reactivity. This was further supported by the significantly reduced variability (standard deviation) in predicted valence and arousal for the MDD group, which itself served as a powerful classifier (AUCs of 0.86 and 0.84). Moreover, the visual decoding analysis (Fig. 4) uncovered clear aberrancies in MDD. Reconstructed images from MDD patients showed significantly reduced brightness and color saturation (Table 2), suggesting a diminished perceptual representation. In addition, the correlation analysis between visual and emotional anomaly scores showed that visual and affective abnormalities in MDD are intertwined, further supporting the theoretical rationale of the dual-pipeline VAD framework. Additionally, the retrieval analysis (Fig. 4c) provided quantitative evidence for the pattern consistent with emotional rumination, a hallmark of depression, where MDD patients exhibited a higher probability of incorrectly retrieving frames associated with earlier, often negative, stimuli. Besides, the clinical relevance of the VAD framework was substantiated by robust correlations between the model-derived anomaly scores and clinical measures. Specifically, we observed a significant positive correlation with the total HAMD-17 score ($r = 0.527$, uncorrected $p = 0.028$, FDR-corrected $q = 0.047$), confirming that higher anomaly scores reflect greater symptom severity. Crucially, the strongest associations were found with the "depression" and "insomnia" sub-factors, suggesting that the neural abnormalities captured by VAD are most saliently linked to the core affective and sleep-related disturbances characteristic of MDD. Finally, our dual-pipeline parcel-level analysis demonstrated a clear functional dissociation between the cortical regions associated with visual and affective abnormalities in MDD. Visual decoding abnormalities are most sensitive to occipital and extrastriate visual regions, whereas affective abnormalities are dominated by default-mode, salience, and interoceptive systems. Importantly, these findings closely align with independent neuroimaging and meta-analytic evidence on depression-related dysfunctions, supporting both the biological plausibility and interpretability of our VAD framework.

Despite the promising results, several limitations should be considered, which also highlight important directions for future work. First, our interpretability analysis, while informative, remained at the level of large-scale functional networks. We did not perform fine-grained, voxel-level or region-specific decoding within these networks, which could obscure the contributions of smaller, critical subregions. Future work should explore more granular decoding to pinpoint specific neural substrates of visual and affective processing deficits. Second, the study is focused on detection and model-inferred neurofunctional interpretation. Although some model-sensitive regions, such as those within the DMN, overlapped with brain networks previously implicated in depression, the current framework does not directly model or predict treatment response, nor does it inform specific neuromodulation strategies. Future studies incorporating longitudinal follow-up, interventional designs, and formal causal modeling will be necessary to determine whether the model-sensitive regions identified here play a causal role in depression-related visual and affective abnormalities or treatment response. Integrating this model with longitudinal data from clinical interventions, such as TMS, psychotherapy, or pharmacological treatment, could therefore be a valuable future direction. Third, both the internal and external cohorts viewed the same naturalistic movie stimulus. This shared stimulus context may partly account for the close agreement between the internally derived Youden threshold and the cohort-specific threshold estimated from the external data. Accordingly, the present findings support cross-site transferability of the diagnostic threshold under the same naturalistic-viewing paradigm, but do not establish its transferability to cohorts acquired using substantially different stimulus sets. Because the VAD framework learns visual and affective response patterns associated with the movie used during model development, application to a different naturalistic stimulus may require model retraining or recalibration, together with re-estimation of the diagnostic threshold. For researchers interested in evaluating new participants using the same paradigm, access to the stimulus material and relevant experimental details may be requested from the corresponding authors, subject to applicable sharing restrictions. Finally,

our adolescent sample, while crucial for understanding depression at a vulnerable developmental stage, may limit the generalizability of these specific neural patterns to adult or late-life depression.

In conclusion, this study introduces VAD, a novel visual–affective decoding framework that integrates cross-subject visual and emotional representations from whole-brain t-fMRI acquired under naturalistic viewing conditions for adolescent MDD detection. By training a dual-pipeline neural network exclusively on healthy controls, VAD establishes a normative model of brain function, enabling depression to be robustly identified as systematic deviations from this norm in previously unseen subjects. Beyond achieving high diagnostic accuracy in both internal and external cohorts, the framework provides mechanistically interpretable and objective neurofunctional signatures, quantitatively capturing core features of depression, including emotional blunting, attenuated visual representation, and retrieval-based affective perseveration patterns. By bridging ecologically valid brain responses with generalizable and interpretable biomarkers, VAD represents a substantive step toward clinically relevant, mechanism-informed tools for psychiatric research and early diagnosis.

Methods

Participants

Two independent cohorts were included in this study: an internal cohort (Dataset 1) and an external validation cohort (Dataset 2). The internal cohort was recruited at Harbin First Specialized Hospital, whereas the external validation cohort was recruited at Tianjin Anding Hospital, Mental Health Center of Tianjin Medical University. In both cohorts, participants were classified into either an HC group or an MDD group according to predefined eligibility criteria. For all participants, demographic information and medical histories were systematically recorded, standardized symptom rating scales were administered, and structural and functional MRI data were acquired while participants viewed the same naturalistic movie stimulus (see "Movie stimuli, emotion rating paradigm, and procedure").

Dataset 1 (internal cohort): HC participants were recruited from local communities and schools via announcements, while adolescents with MDD were consecutively enrolled from the hospital's outpatient and inpatient departments. This study was approved by the Ethics Committee of Harbin First Specialized Hospital and Beijing Anding Hospital (Approval No. 2023-92). Written informed consent was obtained from all participants and their parents or legal guardians prior to enrollment. MDD diagnoses were made according to the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-V) criteria by a senior pediatric psychiatrist and confirmed through structured interviews using the Mini-International Neuropsychiatric Interview for Children and Adolescents, 5th Edition (MINI-Kid 5.0), including both child and parent versions. Participants received a transportation allowance of approximately 27.86 USD.

For adolescents in the MDD group, inclusion criteria (all six criteria must be fulfilled): (1) Aged 12–14 years; (2) Confirmed DSM-V diagnosis of MDD, established by agreement between two psychiatrists and verified using MINI-Kid 5.0; (3) Education beyond primary school and ability to complete all study assessments; (4) Medication washout, defined as no psychotropic medication within the past 2 weeks or use for no more than 1 week; (5) Absence of current impulsive or self-harm risk as assessed by attending clinicians; (6) Right-handedness. Exclusion criteria (any one condition is sufficient for exclusion): (1) Presence of mania or inability to complete study assessments; (2) Any prior neurological or psychiatric impairment; (3) A score of 17 or below on the Hamilton Depression Rating Scale (HAMD); (4) History of substance use or dependence; (5) Current pregnancy; (6) Serious medical conditions; (7) Contraindications to MRI scanning, such as metal implants, pacemakers, or other metallic devices.

For adolescents in the HC group, inclusion criteria (all six criteria must be fulfilled): (1) Aged 12–14 years, with no restriction on sex; (2) Right-hand dominance; (3) Educational level above primary school and sufficient understanding to complete the study questionnaires; (4) HAMD-17 score of 7 points or lower; (5) Ability and willingness to participate in all study-related assessments; (6) Written informed consent obtained from both the participant and their parent or guardian. Exclusion criteria (any one condition is sufficient for exclusion): (1) Past or current diagnosis of depression or any other psychiatric disorder; (2) History of neurological disease or serious physical illness; (3) History of substance abuse, dependence, or acute toxic exposure; (4) Left-handedness, MRI contraindications, or structural brain anomalies detected on MRI scans.

Dataset 2 (external validation cohort): This study was approved by the Ethics Committee of Tianjin Anding Hospital (Approval Nos. 2023-25 and 2024-64). Written informed consent was obtained from all participants and their parents or legal guardians prior to enrollment. MDD diagnoses were made according to DSM-V criteria and confirmed through MINI-Kid.

For adolescents in the MDD group, inclusion criteria (all four criteria must be fulfilled): (1) Aged 8–19 years; (2) Confirmed DSM-V diagnosis of MDD, established by a psychiatrist and verified using MINI-Kid; (3) A Montgomery-Åsberg Depression Rating Scale (MADRS) score of 22 or above; (4) Provision of written informed consent. Exclusion criteria (any one condition is sufficient for exclusion): (1) Presence of metal implants above the chest, particularly intracranial or cardiac implants; (2) Severe organic brain disease (e.g., epilepsy, stroke, encephalitis, traumatic brain injury) or serious physical illness (e.g., malignant tumor, acute heart failure, multi-organ failure); (3) Presence of psychotic symptoms; (4) A history of substance or alcohol abuse within the preceding year; (5) Severe communication difficulties precluding normal interaction, comprehension, or adherence to instructions, or an inability to cooperate with treatment and assessment; (6) Judged unsuitable for participation by the treating physician or investigator; (7) Absence of written informed consent.

For adolescents in the HC group, inclusion criteria (all four criteria must be fulfilled): (1) Matched to the MDD group in age and sex; (2) Absence of psychiatric disorders, screened using MINI-Kid; (3) No personal or family history of psychiatric illness; (4) No metal implants and eligibility for MRI scanning.

Movie stimuli, emotion rating paradigm, and procedure

The study employed a continuous, naturalistic paradigm to deliver emotional stimuli. Participants were not required to perform explicit tasks, such as pressing buttons. For analysis purposes, the fMRI data were segmented into 10-s clips, aligned according to the time-locked emotional valence and arousal scores of the video. The stimulus consisted of a silent compilation of five public-service advertisements, each ranging from 1 min 16 s to 1 min 42 s, with an overall duration of 7 min and 50 s. The clips were presented continuously, without any breaks between them. The videos depicted typical adolescent social experiences with parents and peers, including challenges like adapting to new environments and joining peer groups. These scenarios were designed to evoke strong socio-emotional responses in children and adolescents. Thirty-three college students provided rating scores of the videos' emotional content, including valence and arousal. The rating scores confirmed that the videos elicited a range of positive, negative, and neutral emotional reactions. We normalized the valence and arousal scores of the video clips to values ranging from 0 to 1. For instance, clips showing peer support or warm family moments received positive ratings, conflictual or sad scenes were rated negatively, and sequences depicting passive social observation were rated neutrally. This set of videos has been employed in our earlier research[78,79].

Prior to the MRI scan, participants received standardized instructions. They were informed that the scanner would generate noise during operation, which was normal, and were instructed to remain calm, keep their head as still as possible, and watch the video attentively throughout the scan. Participants were also told that they could signal the experimenter by raising their hand if they experienced any discomfort, at which point the scan would be terminated. After completion of the scan, participants were asked about their physical condition and were reassured that transient sensations such as mild dizziness could occur after prolonged immobility and would resolve after a short rest period. To ensure that participants had attended to and understood the video content, a post-task assessment was administered. In this assessment, participants were asked to judge whether specific images

had appeared in the video and whether brief descriptions accurately reflected the video content. Specifically, the post-task assessment combined a recognition-based recall task with questions probing comprehension of key story details. In the recognition task, participants judged whether selected plots had appeared in the movie, whereas the comprehension questions assessed their understanding and memory of the main narrative content. Performance on these components provided a behavioral index of whether participants had adequately attended to and followed the movie during scanning. Only participants with post-task assessment scores above 60 were included in the present study, indicating that the analyzed participants generally showed adequate attention to and comprehension of the movie content. This criterion helped minimize the possibility that the observed group differences were primarily driven by lower attention to the movie, poorer task engagement, or insufficient understanding of the stimulus, rather than by differences in visual–affective processing associated with MDD.

fMRI data acquisition

For the internal cohort, functional MRI (fMRI) data were acquired using a 3.0 Tesla MRI scanner (GE Healthcare, USA) at the Harbin First Specialized Hospital. Task stimuli were presented via the MedAV auditory–visual stimulation system for functional brain imaging.

FMRI images were obtained using an echo-planar imaging (EPI) sequence with the following parameters: repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, field of view (FOV) = 192 mm, flip angle (FA) = 90°, matrix = 64 × 64, slice thickness = 4 mm, slice gap = 0.2 mm, number of slices = 33, a total of 255 volumes acquired over 8 min and 30 s. Structural images were obtained using a high-resolution T1-weighted three-dimensional BRAVO sequence, consisting of 192 sagittal slices, FOV = 256 mm, matrix = 256 × 256, slice thickness = 1.0 mm with no interslice gap.

Functional MRI data for Dataset 2 were acquired on a Siemens 3.0 Tesla scanner (MAGNETOM Prisma, Siemens, Munich, Germany) at Tianjin Anding Hospital. Naturalistic-viewing N-fMRI scans were acquired using a gradient-echo EPI sequence with the following parameters: repetition time (TR) = 800 ms, echo time (TE) = 30 ms, field of view (FOV) = 208 × 208 mm^2, matrix size = 104 × 104, flip angle = 56°, slice thickness = 2 mm, 72 slices, voxel size = 2 × 2 × 2 mm^3. T1-weighted structural images were acquired using an MP-RAGE sequence (TR = 2000 ms, TE = 2.32 ms, FOV = 230 × 230 mm^2, matrix = 256 × 256, flip angle = 8°, slice thickness = 0.9 mm, 208 slices, voxel size = 0.9 × 0.9 × 0.9 mm^3). Participants viewed the same naturalistic movie stimulus described above (see "Movie stimuli, emotion rating paradigm, and procedure").

fMRI data preprocessing

fMRI data preprocessing was performed using the PhiPiPe multimodal neuroimaging preprocessing pipeline[80]. The main preprocessing steps included: (1) Removal of the first five volumes to allow for signal equilibration; (2) Head motion correction; (3) Rigid registration to the structural T1-weighted image; (4) Estimation of whole-brain, white matter, and ventricular masks; (5) Motion outlier detection and interpolation: volumes with framewise displacement (FD) > 0.5 mm were identified as motion outliers and interpolated using adjacent volumes; (6) Nuisance regression, including the mean signals from white matter and ventricles, as well as Friston's 24-parameter motion model; (7) Temporal band-pass filtering (0.01–0.1 Hz); (8) Global mean scaling to a mean intensity value of 10,000; (9) Spatial normalization to the MNI152 standard space by combining T1-to-MNI and fMRI-to-T1 registration transforms, followed by resampling to 3 × 3 × 3 mm^3 isotropic voxels. Bold fMRI data were preprocessed in individual volume space and then transformed onto the fsaverage space.

Then we calculated the voxel-wise correlation between the fMRI voxel signals of each training movie repetition for each subject. The Pearson correlation coefficient for each voxel underwent Fisher z-transformation, and the average z-scores across all training movie clips were tested using a one-sample *t*-test. The significant voxels (Bonferroni correction, $P < 0.05$) shared across subjects were considered stimulus-activated voxels and were used for subsequent analysis. A total of 13,148 common activated voxels across subjects were observed within the 20,484 voxels in the fsaverage5 template. Following previous works[25,26], we used the BOLD signals with a 4-s delay as the movie stimulus responses to account for the hemodynamic delay.

fMRI data inter-subject correlation analysis

To quantify inter-individual neural similarity during naturalistic stimulation, we performed an ROI-level leave-one-out ISC analysis using the preprocessed movie-viewing fMRI time series in the full fsaverage5 surface space. Cortical vertices were assigned to Schaefer 400 parcels. For each participant, vertex-wise BOLD time series were first standardized across time, averaged within each parcel, and then standardized again at the ROI level before ISC computation.

For HC-to-HC ISC, each HC's ROI time series was correlated with the average ROI time series of all remaining HC to avoid self-inclusion bias. For MDD-to-HC ISC, each MDD's ROI time series was correlated with the average ROI time series of the HC group to estimate similarity to the HC normative response pattern. For MDD-to-MDD ISC, each MDD's ROI time series was correlated with the average ROI time series of all remaining MDD patients. Pearson correlation coefficients were computed across time for each ROI and Fisher-z transformed. Subject-level global ISC values were obtained by averaging Fisher-*z* values across ROIs, and back-transformed Pearson *r* values were reported for interpretability. Group comparisons were performed on subject-level global Fisher-z ISC values using two-tailed Welch's *t*-tests, with BH-FDR correction applied across the two global ISC comparisons.

Overall model design and computational pipeline

In this study, we propose an adolescent depression diagnosis framework that integrates coupled VAD of naturalistic paradigm t-fMRI. The framework integrates semantic and neural representations derived from t-fMRI during both visual and emotion decoding, thereby enhancing the accuracy and interpretability of MDD prediction. An overview of the model design and computational pipeline is presented in Fig. 1.

First, to achieve better cross-subject generalization in decoding and to incorporate a broader range of brain regions for VAD to assist MDD detection, we differ from most previous brain visual decoding algorithms that rely solely on fMRI data from visual cortices. Instead, we advocate the use of whole-brain fMRI data to enhance the model's generalization ability. This strategy of employing whole-brain fMRI data for cross-subject general modeling has been experimentally validated to improve generalization performance[53].

During training, VAD takes as input whole-brain t-fMRI signals, the corresponding visual frames of the presented videos, and the emotional scores of each video clip (valence and arousal). Through joint visual–affective decoding, the network learns the normative neural decoding patterns from healthy controls. In the modality alignment stage, the temporal resolutions of video clips and fMRI data are first synchronized, with the batch size and feature dimensionality adjusted to match the emotional embedding space. Subsequently, a patch-based segmentation and tokenization strategy is adopted to establish a voxel-level correspondence across individual brains, thereby achieving voxel alignment. The anatomically registered voxel features are then fed into a shared MLP backbone containing four residual blocks. The backbone shares parameters across all healthy controls and employs a contrastive learning objective to align the fMRI-derived features with the visual frame representations extracted by a pretrained CLIP encoder (ViT/bigG-14) in a shared latent space.

To incorporate affective information, the fMRI embeddings are then average-pooled to match the temporal resolution of the affective annotations (valence and arousal). These pooled embeddings are passed through a single-layer MLP followed by a fully connected layer to output the predicted affective values. By fusing visual reconstruction features with the predicted

affective representations, VAD jointly learns perceptual-level and emotion-level semantics, improving both generalization and interpretability.

After completing end-to-end pretraining, a diffusion prior module is introduced to further enhance the interpretability of visual reconstruction. This module projects the learned fMRI embeddings back into the CLIP image space, from which high-quality video frames are reconstructed using the SDXL unCLIP model. This enables intuitive visual comparison between MDD patients and healthy controls during inference.

During inference, the model processes t-fMRI data from unseen subjects who viewed the same video stimuli as those used during training. The pretrained VAD network outputs two types of signals: (1) fMRI embeddings aligned to the CLIP feature representations of the video frames via the shared MLP backbone, and (2) predicted emotional embeddings (valence and arousal) produced by the fully connected decoder.

The deviations between these predicted embeddings and their corresponding ground-truth values are quantified as anomaly scores, which serve as discriminative indicators for end-to-end MDD detection. In parallel, the reconstructed images obtained via the diffusion prior and SDXL unCLIP modules are used for qualitative visualization and interpretability analysis.

Through this design, VAD establishes an end-to-end t-fMRI-based detection framework grounded in joint visual–affective decoding. By leveraging neural differences between MDD patients and healthy controls in visual and emotional processing, the model enhances both multimodal predictive accuracy and neurobiological interpretability.

In the following, we describe the key components of the framework and elaborate on the underlying computational workflow.

Multimodal data alignment strategy

We denote by X the preprocessed and registered whole-brain fMRI data obtained from a healthy control in the training set after viewing a 470-s video, excluding the first 5 time points (10 s). The data are registered to the fsaverage5 cortical surface template[67], with a dimensionality of (N_t, N_v), where N_t denotes the temporal dimension sampled at the fMRI acquisition rate, and N_v denotes the number of significant voxels retained after preprocessing.

The video is divided into clips at 10-s intervals, corresponding to the temporal windows used for affective scores of valence and arousal. For each clip c, we perform temporal downsampling and retain a subset of frames at fixed intervals, resulting in a set of frames $Y = \left[Y_1, Y_2, ..., Y_{N_f}\right]$. Y_i is the ith retained video frame and N_f is the total number of retained frames. The corresponding fMRI signal for clip c is denoted as X_c.

To achieve a fine-grained correspondence between visual and neural data, we introduce an Inception Extension module that expands one fMRI clip into N_f temporally aligned fMRI samples, denoted as $X_c \to X, X = \left[X_1, X_2, ..., X_{N_f}\right]$.

Each expanded fMRI sample X_i is then partitioned into n non-overlapping patches $X_i = \left[X_{i_1}, X_{i_2}, ..., X_{i_n}\right], i = 1, 2, ..., N_f$. Each patch contains an equal number of voxels, and we use zero padding to fill the last patch. Subsequently, we apply 1-D convolutions combined with positional encoding to generate a set of voxel-wise tokens.

$$T_i = \left[T_{i_1}, T_{i_2}, \ldots, T_{i_n}\right] = Conv\left(X_i\right) + PE, i = 1, 2, \ldots, N_f \quad (1)$$

After the voxels alignment, multimodal data alignment is completed, and the aligned voxel tokens $T_i, i = 1, 2, ..., N_f$ serve as the input to the neural network model.

Visual–affective joint decoding framework

Given that vision forms the foundation of human perception and provides insight into various factors such as emotion and memory, it serves as a crucial entry point for understanding the pathological mechanisms underlying depression[81]. Therefore, we first perform visual decoding.

In order to achieve high-quality pixel-level reconstruction, it is essential to utilize a substantial quantity of CLIP image embeddings. We employ OpenCLIP ViT-bigG/14 for feature alignment, with image token embeddings of dimension 256×1664. The aligned input tokens T are then fed into an MLP backbone network comprising four residual blocks and a tokenization layer (hidden dimension = 1024), which transforms them to 256×1664. The MLP backbone transforms fMRI features into an intermediate embedding space, denoted as:

$$Z_i = MLP_{backbone}\left(T_i\right), i = 1, 2, \ldots, N_f \quad (2)$$

All fMRI data from healthy controls in the training set share the same MLP backbone after multimodal alignment. Subsequently, these backbone embeddings are passed into the affective decoding module and later into the high-level image reconstruction module following end-to-end training.

In the affective decoding module, we employ a relatively simple MLP network, but dimensional alignment is first required. Since each video clip corresponds to a single pair of emotional dimensions (valence and arousal), we first perform temporal downsampling. Specifically, we apply average pooling across the N_f fMRI embeddings $\left(Z_1, Z_2, ..., Z_{N_f}\right)$ obtained from the visual decoding stage to produce a single aggregated feature vector:

$$\bar{Z} = \frac{1}{N_f}\sum_{i=1}^{N_f} Z_i, \bar{Z} \in \mathbb{R}^{256 \times 1664} \quad (3)$$

This pooled embedding $\bar{Z}$ is then passed through an MLP network consisting of two residual blocks and a tokenization layer to generate the predicted affective vector E:

$$E = MLP_{affect}\left(\bar{Z}\right) \quad (4)$$

The output $E = [\hat{v}, \hat{a}]$ represents the predicted valence ($\hat{v}$) and arousal ($\hat{a}$) scores corresponding to each video clip.

After completing the end-to-end training, we further introduced a diffusion prior module to enhance the interpretability of visual reconstruction. Because the CLIP ViT image-space embeddings are more structurally aligned with real visual representations than the fMRI embeddings, transforming the fMRI embedding Z_c into the CLIP ViT image embedding space substantially improves the quality and realism of the reconstructed frames.

Inspired by DALLE·2[82], the diffusion prior serves as an efficient mapping mechanism between latent embedding spaces. Accordingly, we map the fMRI embedding Z_c into the OpenCLIP ViT-bigG/14 image feature space to obtain a reconstructed latent representation Z_c^{re}. This representation is then passed into the SDXL unCLIP model to achieve high-fidelity video frame reconstruction, thereby enabling direct visual comparison between MDD patients and healthy controls during inference.

The SDXL unCLIP model used here was originally introduced in MindEye2[49], which experimentally validated its capacity to reconstruct highly realistic images from genuine OpenCLIP ViT-bigG/14 image embeddings, demonstrating that the reconstructed frames were nearly indistinguishable from the original stimuli.

Training and implementation details

The optimization objectives of the proposed VAD model consist of two major components: (1) The end-to-end training of the visual–emotional joint decoding and detection framework, which integrates two complementary loss terms, a contrastive alignment loss that maps fMRI features to the shared CLIP embedding space in correspondence with real video frames and a regression loss for the video-clip–level emotional decoding task, formulated as Mean Squared Error (MSE); (2) A diffusion prior loss applied after model convergence, used to guide the high-fidelity

reconstruction of video frames from the learned embeddings.

$$L = L_{emo} + \lambda L_{contra} \tag{5}$$

$$L_{recon} = L_{prior} \tag{6}$$

For the selection of the weighting parameter λ, because the visual reconstruction and affective prediction branches used different loss functions and operated on different numerical scales, λ was introduced to balance their relative contributions during training. λ was selected through grid search within the fivefold cross-validation procedure, with the mean validation AUC across folds used as the selection criterion. As shown in Supplementary Fig. 6, the best performance was achieved around $\lambda = 100$; therefore, λ was fixed at 100 in the final experiments.

During the end-to-end training phase, to achieve optimal cross-modal semantic alignment between brain and visual representations, we employ a contrastive learning objective that maximizes the cosine similarity between positive (matched) pairs while minimizing it for negative (unmatched) pairs. This approach encourages the model to establish a robust shared latent space linking fMRI signals and visual embeddings.

Specifically, following the implementation strategy of MindEye1[37], we adopt a combination of BiMixCo and SoftCLIP losses. In the first one-third of training iterations, we apply bidirectional MixCo data augmentation with hard labels[83], enhancing discriminative alignment. In the remaining two-thirds of training, we employ soft-label optimization without augmentation to stabilize training and refine semantic correspondence.

The MixCo voxel mixture paradigm applied during the first one-third of training iterations is restricted within individual video clips, rather than across clips. This design prevents potential cross-clip interference between fMRI signals, which could otherwise confound subsequent emotional decoding tasks.

Specifically, for a given video clip c, two fMRI voxel representations $X_i \in X_c$ and $X_j \in X_c$ are randomly selected and mixed using a weighting coefficient λ sampled from a Beta distribution, yielding the mixed voxel representation $X_{mix_{i,j}}$:

$$\begin{aligned} X_{mix_{i,j}} &= \lambda_i X_i + (1-\lambda_i) X_j, \\ T_{mix_{i,j}} &= Conv\left(X_{mix_{i,j}}\right) + PE, \\ Z_{mix_{i,j}} &= MLP_{(backbone)} T_{mix_{i,j}} \end{aligned} \tag{7}$$

where j denotes any randomly selected voxel index within the same clip. The forward MixCo contrastive loss for voxel-level embedding alignment within the same clip is then defined as:

$$L_{MixCo} = -\frac{1}{|N_f|}\sum_{i=1}^{|N_f|}\left[\lambda_i \cdot \log\frac{\exp\left(Z_{mix_{i,j}} \cdot \frac{g_i}{\tau}\right)}{\sum_{m=1}^{|N_f|}\exp\left(Z_{mix_{i,j}} \cdot \frac{g_m}{\tau}\right)} + (1-\lambda_i) \cdot \log\frac{\exp\left(Z_{mix_{i,j}} \cdot \frac{g_j}{\tau}\right)}{\sum_{m=1}^{|N_f|}\exp\left(Z_{mix_{i,j}} \cdot \frac{g_m}{\tau}\right)}\right] \tag{8}$$

where g denotes ground-truth CLIP image embeddings and τ denotes a temperature hyperparameter. Here we use the bidirectional loss $L_{BiMixCo}$.

To further enhance the performance of subsequent reconstruction tasks and to facilitate more effective optimization toward its own learning objective, we additionally employ the SoftCLIP loss within each video clip for all fMRI samples. The training procedure transitions from the MixCo strategy to SoftCLIP loss after the first one-third of the total training iterations:

$$L_{SoftCLIP} = -\sum_{i=1}^{N_f}\sum_{j=1}^{N_f}\left[\frac{\exp\left(g_i \cdot \frac{g_j}{\tau}\right)}{\sum_{m=1}^{N_f}\exp\left(g_i \cdot \frac{g_m}{\tau}\right)} \cdot \log\frac{\exp\left(Z_i \cdot \frac{g_j}{\tau}\right)}{\sum_{m=1}^{N_f}\exp\left(Z_i \cdot \frac{g_m}{\tau}\right)}\right] \tag{9}$$

Therefore, we take the average of the loss values across all video clips to obtain the final contrastive loss optimization function.

$$L_{contra} = \frac{1}{N_c}\sum_{c=1}^{N_c} L_{c_{BiMixCo|SoftCLIP}} \tag{10}$$

For the regression-based emotional decoding task using valence and arousal, we simply constrain the predicted values $\hat{v}$ and $\hat{a}$ for each video clip with a standard Mean Squared Error (MSE) loss, defined as follows:

$$L_{emo} = \frac{1}{N_c}\sum_{c=1}^{N_c}\left[||v_c - \hat{v}_c||_2^2 + ||a_c - \hat{a}_c||_2^2\right] \tag{11}$$

where N_c denotes the number of video clips.

Finally, for the diffusion prior loss, we constrain the high-fidelity reconstructed video frames $\hat{g}_i$ using the MSE loss with respect to the ground-truth CLIP ViT image embeddings g_i:

$$L_{prior} = \frac{1}{B}\sum_{i=1}^{B}||g_i - \hat{g}_i||_2^2 \tag{12}$$

where B denotes the batch size used during the reconstruction process. Note that, in this case, the batch does not necessarily correspond to the number of frames within a video clip.

The proposed model was implemented in Python 3.11 using PyTorch 2.1.0 and trained for 100 epochs on four NVIDIA RTX 4090 GPUs. For optimization, we employed the Adam optimizer with an initial learning rate of 3×10^{-4}, combined with a OneCycle learning rate scheduling strategy to ensure stable convergence. Detailed architectural configurations and implementation specifications are provided in Supplementary Table 12.

Anomaly inference framework for unseen subjects

During the inference phase, the model takes as input the t-fMRI data collected from new adolescent subjects while they view the same video stimuli. After identical preprocessing, the data are fed into the pretrained visual–emotional joint decoding model for unified prediction. As shown in our previous analyzes, since the model is trained exclusively on healthy controls, MDD patients typically exhibit higher anomaly scores when their data are processed by the pretrained model. These deviations reflect atypical neural patterns that diverge from the normative visual and emotional representations observed in healthy controls, thereby serving as the basis for MDD detection.

Specifically, in anomaly estimation, we follow the same optimization objectives as those used during training, consisting of both visual anomaly scores and emotional anomaly scores. The final detection score is computed as a weighted combination of these two components. For visual anomaly estimation, we calculate the deviation between the model-predicted fMRI embeddings for a given clip c, denoted as $\hat{Z}_i, i = 1, 2, \ldots, N_f$, and the CLIP image embeddings g_i of the corresponding downsampled video frames within the same clip, using the contrastive loss error. This yields the visual anomaly score e_{recon}:

$$e_{c_{recon}} = -\frac{1}{N_f}\sum_{i=1}^{N_f}\log\frac{\exp\left(\hat{Z}_i \cdot \frac{g_i}{\tau}\right)}{\sum_{m=1}^{N_f}\exp\left(\hat{Z}_i \cdot \frac{g_m}{\tau}\right)} \tag{13}$$

$$e_{recon} = \frac{1}{N_c}\sum_{c=1}^{N_c} e_{c_{recon}} \tag{14}$$

For the emotional anomaly score, we compute the mean squared error between the predicted and ground-truth emotional indicators, valence and arousal, across all video clips. This quantifies the degree of deviation in the subject's emotional decoding from the normative patterns observed in

healthy controls:

$$e_{emo} = \frac{1}{N_c}\sum_{c=1}^{N_c}\left[||v_c - \hat{v}_c||_2^2 + ||a_c - \hat{a}_c||_2^2\right] \quad (15)$$

Then, the same weighting parameter λ was also used when combining the visual and emotional anomaly scores during inference. This choice was made because the anomaly scores were derived from the corresponding branch-specific training objectives and therefore retained similar scale differences between the visual and affective components. Using the same λ ensured consistency between the training objective and the inference-stage anomaly score integration.

$$e = e_{emo} + \lambda e_{recon} \quad (16)$$

Finally, the integrated anomaly score e was converted into a binary diagnostic decision using a fixed threshold. Specifically, a subject was classified as MDD when $e \geq t$, and as HC otherwise. In the internal development cohort, the threshold t was determined by maximizing the Youden[54] index,

$$J(t) = Sensitivity(t) + Specificity(t) - 1 \quad (17)$$

Using the internal fivefold cross-validation results, the optimal threshold was identified as $t = 13.135$. This threshold was then fixed and directly applied to the independent external validation cohort without further adjustment.

Performance validation

To ensure the reliability and robustness of feature selection, model performance was rigorously evaluated using fivefold cross-validation in the internal cohort, incorporating standard quantitative metrics such as the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, and specificity. Specifically, the training cohort of healthy controls was randomly partitioned into five equal subsets. In each iteration, four subsets were used as the training data to fit the model, while the remaining subset, together with the MDD patients, was used to evaluate model generalization. This procedure was repeated five times so that each subset served once as the validation fold. The optimal set of hyperparameters and the anomaly threshold determined during cross-validation were subsequently employed to train the final model on the complete training cohort, which was then evaluated on the independent external validation cohort without further adjustment. To clarify whether the random fold assignment introduced substantial demographic or clinical imbalance, we further examined the demographic and clinical characteristics of the held-out test samples across the five folds. As shown in Supplementary Table 13, the fold-wise summaries included the female proportion in HC and MDD groups, mean age in HC and MDD groups, and mean HAMD-17 and HAMA scores in MDD group. This analysis was added to clarify the composition of each cross-validation fold and to assess whether the random fold assignment resulted in apparent demographic or clinical imbalance. Overall, the fold-level summaries suggested that the test folds were broadly comparable in age, sex and clinical severity. All computational and statistical analyzes were implemented in Python 3.11.

Subsequently, to validate the Emotion Context Insensitivity (ECI) hypothesis, we computed clip-wise Pearson correlation coefficients and their significance levels between the model-predicted emotional anomaly scores and the ground-truth valence and arousal scores for each MDD patient. This analysis assessed the relationship between the model-derived emotional anomalies and behavioral emotion responses, thereby verifying the consistency of the model's predictions with the ECI hypothesis. Furthermore, the within-group variances of decoded valence and arousal were compared between MDD and HC groups and used as discriminative features for classification. Performance was evaluated using AUC, accuracy, sensitivity, and specificity to assess differences in emotional variability between groups (see "Abnormal affective patterns in MDD").

Next, for visual decoding validation, we employed the diffusion prior module and the SDXL-unCLIP model built upon the pretrained end-to-end framework to perform high-fidelity video reconstruction. Quantitative analysis of brightness and RGB color-channel intensity using t-tests was conducted to evaluate the relationship between reconstructed image features and underlying perceptual representations. In addition, an image retrieval analysis was performed: one middle frame per second was extracted from the 460 s video, and the cosine similarity in CLIP embedding space between the reconstructed and ground-truth frames was computed. The Top-1 retrieval accuracy was used to evaluate image-matching performance. The retrieval database included only previously viewed frames to prevent future-frame interference. Finally, a comparative analysis of Top-1 accuracy between MDD and HC groups, along with the probability distribution of erroneous retrievals in MDD, was designed to examine the neural-behavioral characteristics associated with persistent cognition and the pattern consistent with emotional rumination (see "Visual decoding feature results").

Finally, the model-derived overall anomaly scores were correlated with participants' (both MDD and HC groups) HAMD-17 clinical ratings to validate the association between predicted neural abnormalities and the severity of depressive symptoms. Additionally, item-wise correlation analyzes were conducted across major HAMD-17 dimensions (including anxiety, depression, insomnia, and somatic symptoms) to further investigate the correspondence between model-derived anomaly patterns and specific clinical features (see "Correlation results between anomaly scores and HAMD-17 clinical measures").

Interpretation of functional imaging

To localize cortical regions contributing most strongly to the model-predicted abnormality patterns, we adopted a parcel-level masking approach grounded on the fsaverage5 cortical surface template[67]. Specifically, we utilized a fine-grained parcellation comprising 400 cortical parcels distributed across 17 functional networks, covering both hemispheres. Each parcel represented a spatially contiguous cortical unit associated with a specific functional system, enabling a detailed characterization of regional contributions beyond coarse network-level effects.

In this analysis, we implemented a parcel-wise ablation procedure: for each parcel, all voxels within the corresponding cortical region were selectively masked, while voxels outside the parcel remained unchanged. The masked fMRI data were then re-input into the pre-trained model to compute the resulting anomaly scores. By comparing this perturbed score with the unmasked baseline, we quantified the relative contribution of each parcel to the overall MDD-related abnormality.

This masking process was performed iteratively across all 400 cortical parcels. The resulting parcel-wise abnormality changes were aggregated and visualized as contribution profiles (see "Neuroscience interpretation results"), allowing us to identify specific cortical regions whose perturbation most strongly altered model predictions. This approach provides an interpretable, fine-grained characterization of how distinct cortical parcels modulate MDD-related functional abnormalities.

To further assess whether the ablation-derived parcel contributions were supported by model-independent neural evidence, we conducted an additional RSA-based normative-deviation analysis using the original BOLD signals. This analysis was performed independently of the VAD model outputs, reconstructed images, and anomaly scores. For each valid parcel, voxel signals were averaged within the parcel for each of the 46 movie clips, yielding a 46-dimensional clip-wise response vector for each participant. For each parcel, an HC normative representational vector was computed by averaging the corresponding 46-dimensional vectors across HC group. For HC group, a leave-one-subject-out procedure was used when computing the HC normative vector to avoid self-inclusion bias. Each participant's representational similarity to the HC normative vector was then calculated using Pearson correlation, and RSA normative deviation

was defined as the reduction in similarity relative to the HC normative representation.

The 400 parcels were ranked according to their ablation-derived contribution scores and divided into four quartiles from low to high contribution (Q1–Q4). Within each quartile, participant-level RSA normative deviations were averaged across parcels. We then quantified the association between diagnostic status and RSA normative deviation using Pearson correlation, and the standardized HC–MDD group difference was estimated using Cohen's d. This analysis was designed to test whether parcels with higher model-derived contribution scores also exhibited larger MDD-related deviations in the original BOLD representational structure.

For completeness, a complementary coarse-grained analysis at the 17-network level is provided in Supplementary Fig. 5.

Statistical analysis

All hypothesis tests were two-tailed and used a significance level of 0.05.

The Area Under Curve (AUC), accuracy, sensitivity and specificity were used to measure the model performance. These metrics are defined as follows:

$$accuracy = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}} \tag{18}$$

$$sensitivity = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{19}$$

$$specificity = \frac{N_{TN}}{N_{TN} + N_{FP}} \tag{20}$$

$$AUC = \int_0^1 TPR(FPR)d(FPR) \tag{21}$$

where

$$TPR = sensitivity = \frac{N_{TP}}{N_{TP} + N_{FN}}, FPR = 1 - specificity = \frac{N_{FP}}{N_{TN} + N_{FP}}$$

N_{TP}, N_{TN}, N_{FP} and N_{FN} represent the number of true positive, true negative, false positive, false negative samples, respectively.

To control for FDR arising from multiple parallel comparisons, the Benjamini–Hochberg (BH) procedure[66] was applied to correlation analyzes involving multiple related tests. For example, BH-FDR correction was applied to the correlations between overall anomaly scores and HAMD-17 total/subscale scores, as well as to the correlations between visual or emotional anomaly scores and HAMD-17 clinical variables. BH-corrected p-values, reported as q values, are presented alongside uncorrected p-values throughout the results where applicable.

Within each predefined family, the m uncorrected p-values were first ranked in ascending order. The Benjamini–Hochberg adjusted p-value (q-value) for the ith ordered test was computed recursively as

$$p_{\mathrm{adj}}(i) = \min\left(p_{\mathrm{adj}}(i+1), p_{(i)} \times \frac{m}{i}\right) \tag{22}$$

where m denotes the number of statistical tests within the corresponding family. Statistical significance was determined using an FDR threshold of $q < 0.05$.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to ethical and privacy restrictions associated with human neuroimaging data, but are available from the corresponding author on reasonable request.

Code availability

The full implementation of the proposed VAD model is publicly available at https://github.com/jingxuLiu/VAD.

Received: 13 January 2026; Accepted: 5 October 2026;
Published online: 26 October 2026

References

1. Cassano, P. & Fava, M. Depression and public health: an overview. *J. Psychosom. Res.* **53**, 849–857 (2002).
2. Murray, C. J. et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2197–2223 (2012).
3. Klein, D. N., Goldstein, B. L. & Finsaas, M. Depressive disorders. in *Child and Adolescent Psychopathology* 3rd edn, 610–641 (2017).
4. Stringaris, A. What is depression? *Wiley Online Libr.* **58**, 1287–1289 (2017).
5. Flint, J. The genetic basis of major depressive disorder. *Mol. Psychiatry* **28**, 2254–2265 (2023).
6. Singh, M. K., Gorelik, A. J., Stave, C. & Gotlib, I. H. Genetics, epigenetics, and neurobiology of childhood-onset depression: an umbrella review. *Mol. Psychiatry* **29**, 553–565 (2024).
7. Yuan, M. et al. Epigenetic regulation in major depression and other stress-related disorders: molecular mechanisms, clinical relevance and therapeutic potential. *Signal Transduct. Target. Ther.* **8**, 309 (2023).
8. Organization, W. H. *Depressive Disorder (Depression) World Health Organization. Geneva, Switzerland* (2023).
9. Shen, J. et al. Depression detection from electroencephalogram signals induced by affective auditory stimuli. in *Proc. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* 76–82 (IEEE, 2019).
10. Shen, J., Zhang, X., Wang, G., Ding, Z. & Hu, B. An improved empirical mode decomposition of electroencephalogram signals for depression detection. *IEEE Trans. Affect. Comput.* **13**, 262–271 (2019).
11. Zhong, J. et al. Robust discriminant feature extraction for automatic depression recognition. *Biomed. Signal Process. Control* **82**, 104505 (2023).
12. Liu, M. et al. Hierarchical encoding and fusion of brain functions for depression subtype classification. *IEEE Trans. Affect. Comput.* **15**, 1826–1837 (2024).
13. Liu, M. et al. Randomizing human brain function representation for brain disease diagnosis. *IEEE Trans. Med. Imaging* **43**, 2537–2546 (2024).
14. Mahendran, R. K., Rohini, C. & Kumar, P. DST-M2D: a dualistic-series transformer framework for major depressive disorder analysis using multimodal MRI data. in *Proc. 2025 International Conference on Emerging Smart Computing and Informatics (ESCI)* 1–10 (IEEE, 2025).
15. Drysdale, A. T. et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* **23**, 28–38 (2017).
16. Liu, F. et al. Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. *PLoS ONE* **7**, e40968 (2012).
17. Lin, W., Wu, H., Liu, Y., Lv, D. & Yang, L. A CCA and ICA-based mixture model for identifying major depression disorder. *IEEE Trans. Med. Imaging* **36**, 745–756 (2016).
18. Kustubayeva, A., Eliassen, J., Matthews, G. & Nelson, E. FMRI study of implicit emotional face processing in patients with MDD with melancholic subtype. *Front. Hum. Neurosci.* **17**, 2023 (2023).
19. Akhapkin, R. V., Volel, B. A., Shishorin, R. M., Ustyuzhanin, D. V. & Petelin, D. S. Recognition of facial emotion expressions in patients with depressive disorders: a prospective, observational study. *Neurol. Ther.* **10**, 225–234 (2021).

20. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G. & Pantic, M. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat. Mach. Intell.* **3**, 42–50 (2021).
21. Saarimäki, H. Naturalistic stimuli in affective neuroimaging: a review. *Front. Hum. Neurosci.* **15**, 2021 (2021).
22. Klamer, K. et al. Differential fMRI neural synchrony associated with migraine during naturalistic stimuli with negative emotional valence. *J. Headache Pain.* **26**, 62 (2025).
23. Zhou, F. et al. Capturing dynamic fear experiences in naturalistic contexts: an ecologically valid fMRI signature integrating brain activation and connectivity. *IEEE Trans. Affect. Comput.* 1–18 (2025).
24. Yu, S. et al. A hybrid learning framework for fine-grained interpretation of brain spatiotemporal patterns during naturalistic functional magnetic resonance imaging. *Front. Hum. Neurosci.* **16**, 2022 (2022).
25. Chan, H.-Y., Smidts, A., Schoots, V. C., Sanfey, A. G. & Boksem, M. A. S. Decoding dynamic affective responses to naturalistic videos with shared neural patterns. *NeuroImage* **216**, 116618 (2020).
26. Wang, L. et al. Arousal modulates the amygdala-insula reciprocal connectivity during naturalistic emotional movie watching. *NeuroImage* **279**, 120316 (2023).
27. Xu, S. et al. Ecologically-valid emotion signatures enhance mood disorder diagnostics. *Adv. Sci.* e05524.
28. Horikawa, T., Cowen, A. S., Keltner, D. & Kamitani, Y. The neural representation of visually evoked emotion is high-dimensional, categorical, and distributed across transmodal brain regions. *iScience* **23** (2020).
29. Fu, K., Du, C., Wang, S. & He, H. Multi-view multi-label fine-grained emotion decoding from human brain activity. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 9026–9040 (2022).
30. Anders, S., Lotze, M., Erb, M., Grodd, W. & Birbaumer, N. Brain activity underlying emotional valence and arousal: a response-related fMRI study. *Hum. Brain Mapp.* **23**, 200–209 (2004).
31. Mattek, A. M., Burr, D. A., Shin, J., Whicker, C. L. & Kim, M. J. Identifying the representational structure of affect using fMRI. *Affect. Sci.* **1**, 42–56 (2020).
32. Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A. & Wager, T. D. A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biol.* **13**, e1002180 (2015).
33. Liu, X. et al. A neural signature for the subjective experience of threat anticipation under uncertainty. *Nat. Commun.* **15**, 1544 (2024).
34. Zhang, R. et al. A neurofunctional signature of affective arousal generalizes across valence domains and distinguishes subjective experience from autonomic reactivity. *Nat. Commun.* **16**, 6492 (2025).
35. Lee, S. A. et al. Brain representations of affective valence and intensity in sustained pleasure and pain. *Proc. Natl. Acad. Sci. USA* **121**, e2310433121 (2024).
36. Ozcelik, F. & VanRullen, R. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Sci. Rep.* **13**, 15666 (2023).
37. Scotti, P. et al. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *Adv. Neural Inf. Process. Syst.* **36**, 24705–24728 (2023).
38. Kaur, T. & Gandhi, T. K. Automated brain image classification based on VGG-16 and transfer learning. in *Proc. 2019 International Conference on Information Technology (ICIT)* 94–98 (IEEE, 2019).
39. Lin, S., Sprague, T. & Singh, A. K. Mind reader: Reconstructing complex images from brain activities. *Adv. Neural Inf. Process. Syst.* **35**, 29624–29636 (2022).
40. Liu, Y., Ma, Y., Zhou, W., Zhu, G. & Zheng, N. Brainclip: Bridging brain and visual-linguistic representation via clip for generic natural visual stimulus decoding. Preprint at https://arxiv.org/abs/2302.12971 (2023).
41. Gong, Z. et al. Lite-mind: Towards efficient and robust brain representation learning. in *ACM Multimedia 2024* (2024).
42. Bao, G., et al. Wills Aligner: multi-subject collaborative brain visual decoding. in *Proc. AAAI Conference on Artificial Intelligence*, Vol. 39 14194–14202 (2025).
43. Chen, Z., Qing, J. & Zhou, J. H. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Adv. Neural Inf. Process. Syst.* **36**, 24841–24858 (2023).
44. Gong, Z. et al. NeuroClips: Towards high-fidelity and smooth fMRI-to-video reconstruction. *Adv. Neural Inf. Process. Syst.* **37**, 51655–51683 (2024).
45. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *elife* **9**, e56601 (2020).
46. Wang, Z., Murai, Y. & Whitney, D. Idiosyncratic perception: a link between acuity, perceived position and apparent size. *Proc. R. Soc. B* **287**, 20200825 (2020).
47. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
48. Takagi, Y. & Nishimoto, S. High-resolution image reconstruction with latent diffusion models from human brain activity. in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 14453–14463 (2023).
49. Scotti, P. S. et al. Mindeye2: shared-subject models enable fMRI-to-image with 1h of data. Preprint at https://arxiv.org/abs/2403.11207 (2024).
50. Gong, Z. et al. Mindtuner: Cross-subject visual decoding with visual fingerprint and semantic correction. *Proc. AAAI Conf. Artif. Intell.* **39**, 14247–14255 (2025).
51. Tian, Z., Quan, R., Ma, F., Zhan, K. & Yang, Y. BRAINGUARD: privacy-preserving multisubject image reconstructions from brain activities. *Proc. AAAI Conf. Artif. Intell.* **39**, 14414–14422 (2025).
52. Dai, Y. et al. MindAligner: explicit brain functional alignment for cross-subject visual decoding from limited fMRI data. Preprint at https://arxiv.org/abs/2502.05034 (2025).
53. Kong, X., Huang, K., Li, P. & Zhang, L. Toward generalizing visual brain decoding to unseen subjects. Preprint at https://arxiv.org/abs/2410.14445 (2024).
54. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).
55. Kaiser, R. H., Andrews-Hanna, J. R., Wager, T. D. & Pizzagalli, D. A. Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity. *JAMA Psychiatry* **72**, 603–611 (2015).
56. Miller, C. H., Hamilton, J. P., Sacchet, M. D. & Gotlib, I. H. Meta-analysis of functional neuroimaging of major depressive disorder in youth. *JAMA Psychiatry* **72**, 1045–1053 (2015).
57. Bylsma, L. M., Morris, B. H. & Rottenberg, J. A meta-analysis of emotional reactivity in major depressive disorder. *Clin. Psychol. Rev.* **28**, 676–691 (2008).
58. Kuppens, P. et al. The relation between valence and arousal in subjective experience varies with personality and culture. *J. Personal.* **85**, 530–542 (2017).
59. Kuppens, P., Tuerlinckx, F., Russell, J. A. & Barrett, L. F. The relation between valence and arousal in subjective experience. *Psychol. Bull.* **139**, 917 (2013).
60. Nolen-Hoeksema, S. Responses to depression and their effects on the duration of depressive episodes. *J. Abnorm. Psychol.* **100**, 569 (1991).
61. Nolen-Hoeksema, S. & Morrow, J. Effects of rumination and distraction on naturally occurring depressed mood. *Cogn. Emot.* **7**, 561–570 (1993).
62. Cooney, R. E., Joormann, J., Eugène, F., Dennis, E. L. & Gotlib, I. H. Neural correlates of rumination in depression. *Cogn. Affect. Behav. Neurosci.* **10**, 470–478 (2010).

63. Kragel, P. A., Reddan, M. C., LaBar, K. S. & Wager, T. D. Emotion schemas are embedded in the human visual system. *Sci. Adv.* **5**, eaaw4358 (2019).
64. Bo, K. et al. Decoding neural representations of affective scenes in retinotopic visual cortex. *Cereb. Cortex* **31**, 3047–3063 (2021).
65. Seemüller, F. et al. A factor analytic comparison of three commonly used depression scales (HAMD, MADRS, BDI) in a large sample of depressed inpatients. *BMC Psychiatry* **23**, 548 (2023).
66. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
67. Fischl, B., Sereno, M. I., Tootell, R. B. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
68. Larivière, S. et al. The ENIGMA Toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nat. Methods* **18**, 698–700 (2021).
69. Wu, F., Lu, Q., Kong, Y. & Zhang, Z. A comprehensive overview of the role of visual cortex malfunction in depressive disorders: opportunities and challenges. *Neurosci. Bull.* **39**, 1426–1438 (2023).
70. Sanz-Morales, E. & Melero, H. Advances in the fMRI analysis of the default mode network: a review. *Brain Struct. Funct.* **230**, 22 (2024).
71. Pinto, A. M. et al. Emotion regulation and the salience network: a hypothetical integrative model of fibromyalgia. *Nat. Rev. Rheumatol.* **19**, 44–60 (2023).
72. Hamilton, J. P. et al. Default-mode and task-positive network activity in major depressive disorder: implications for adaptive and maladaptive rumination. *Biol. Psychiatry* **70**, 327–333 (2011).
73. Kim, S. M. et al. Affective network and default mode network in depressive adolescents with disruptive behaviors. *Neuropsychiatr. Dis. Treat.* 49–56 (2015).
74. Li, X. et al. Braingnn: Interpretable brain graph neural network for fMRI analysis. *Med. Image Anal.* **74**, 102233 (2021).
75. Zhang, H. et al. Classification of brain disorders in rs-fMRI via local-to-global graph neural networks. *IEEE Trans. Med. Imaging* **42**, 444–455 (2022).
76. Ordaz, S. J., Foran, W., Velanova, K. & Luna, B. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. *J. Neurosci.* **33**, 18109–18124 (2013).
77. Gruskin, D. C., Rosenberg, M. D. & Holmes, A. J. Relationships between depressive symptoms and brain responses during emotional movie viewing emerge in adolescence. *NeuroImage* **216**, 116217 (2020).
78. Xie, S. et al. A normative model of brain responses to social scenarios reflects the maturity of children and adolescents' social–emotional abilities. *Soc. Cogn. Affect. Neurosci.* **18**, nsad062 (2023).
79. Jin, S. et al. Aberrant functional connectivity of the bed nucleus of the stria terminalis and its age dependence in children and adolescents with social anxiety disorder. *Asian J. Psychiatry* **82**, 103498 (2023).
80. Hu, Y. et al. PhiPipe: a multi-modal MRI data processing pipeline with test–retest reliability and predicative validity assessments. *Hum. Brain Mapp.* **44**, 2062–2084 (2023).
81. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT press, 2010).
82. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with clip latents. Preprint at https://arxiv.org/abs/2204.06125 **1**, 3 (2022).
83. Kim, S., Lee, G., Bae, S. & Yun, S.-Y. Mixco: mix-up contrastive learning for visual representation. Preprint at https://arxiv.org/abs/2010.06300 (2020).

Acknowledgements

This work was supported by National Natural Science Foundation of China (62576251), National Natural Science Foundation of China (62376198, 62406225), National Key Research and Development Program under Grant 2025ZD0219200, Beijing Research Ward Excellence Program (BRWEP2024W072120106), National Natural Science Foundation of China (82471576), the Capital Medical University Laboratory for Clinical Medicine, Beijing High-Level Innovation and Entrepreneurship Talent Support Program-Dengfeng Project (G202512066), and Research Fund from Beijing Anding Hospital (YG202302).

Author contributions

J.L., M.T., Y.L., L.X., D.L., D.M., and Q.Z. conceived and designed the study. K.L., Z.X., H.T., and Z.Y. collected the data. J.L., K.L., Q.Z., and Z.Y. analyzed the data. J.L., Q.Z., and K.L. drafted the manuscript. M.T., Y.K., L.H., L.Z., D.M., and Z.Y. revised the draft. All authors read and approved the final manuscript.

Funding

The authors received no specific funding for this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-03413-8.

Correspondence and requests for materials should be addressed to Duoqian Miao, Qi Zhang or Zhi Yang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.