

LE-Viz: Complementing Text Serialization with Visual Encoding of Structured Temporal Data for Cross-Distribution Generalization

Ahatsham Hayat¹ Bilal Khan² Mohammad Rashedul Hasan³

¹Department of Anesthesiology, Brigham and Women’s Hospital,

Harvard Medical School, Boston, MA, USA

²Lehigh University, PA, USA

³University of Nebraska-Lincoln, NE, USA

afnu@bwh.harvard.edu, bik221@lehigh.edu, hasan@unl.edu

Abstract

When large language models process structured temporal data, serialization into token sequences collapses the two-dimensional organization (*Features* $\times$ *Time*) that encodes meaningful relationships between co-occurring features. We investigate whether visual encoding can recover the structural information that linearization loses. We introduce LE-Viz, a framework that transforms longitudinal experiential data into complementary visual and textual representations for a vision-language model that generates prospective narratives, framing prediction as language generation. Evaluated on cross-distribution tasks across three datasets spanning mental health and educational forecasting, LE-Viz sets a new benchmark on a widely-adopted cross-distribution generalization task, improving over both the traditional state-of-the-art and the best text-only LLM baseline, and consistently outperforms all baselines. Analyses show that spatial organization of visual inputs, not information quantity, drives the generalization gain, and that the multimodal combination exceeds both text-only and vision-only conditions.

1 Introduction

Large language models (LLMs) are widely applied to structured data that originates outside natural language, including tabular records, clinical time series, and behavioral sensing streams (Hegselmann et al., 2023; Herzig et al., 2020; Kim et al., 2024; Thach et al., 2025; Hayat et al., 2024b). A common strategy is to serialize such data into token sequences so that it can be processed within the autoregressive architecture of LLMs. While effective for short or low-dimensional data, serialization often obscures structural relationships when temporal and feature interactions become more complex.

Longitudinal experiential (LE) data is a setting where this mismatch is both pronounced and consequential. LE data combines subjective self-reports

(e.g., mood, stress) with objective sensor measurements (e.g., physical activity, sleep, phone usage), collected repeatedly from individuals over extended periods through established paradigms such as Experience Sampling Method (Larson and Csikszentmihalyi, 1983) and Ecological Momentary Assessment (Stone and Shiffman, 1994; Shiffman et al., 2008; Mohr et al., 2017). The resulting data exhibits a *Features* $\times$ *Time* organization where concurrent patterns across modalities bear semantic weight. For instance, simultaneous declines in physical activity, shifts in communication behavior, and lower self-reported affect may collectively signal behavioral change requiring contextual interpretation beyond statistical correlation (Saeb et al., 2015; Patharkar et al., 2024). When this two-dimensional structure is serialized into a token sequence, co-occurring feature values occupying the same time step in the original data matrix become scattered across distant token positions. The patterns distinguishing one behavioral trajectory from another become difficult to recover. Both standard machine learning (ML) and deep learning (DL) approaches (Saeb et al., 2015; Xu et al., 2023a) and LLM-based methods (Kim et al., 2024; Thach et al., 2025; Hayat et al., 2025) exhibit limited cross-distribution generalization on LE data, a failure attributable in part to this representational collapse.

When LE data with N features across T time steps is serialized, temporally adjacent values of the same feature become separated by at least $N-1$ intervening tokens. LLM-based approaches bring the advantage of pre-trained language understanding and contextual reasoning, but they inherit this constraint of one-dimensional processing. Temporal proximity in the original data does not translate to positional proximity in the token sequence. Vision-language models (VLMs) have also been applied to time-series forecasting by converting numerical sequences into visual format (Zhong et al., 2025).

Such approaches target purely numerical univariate data and do not accommodate the heterogeneity of LE data, which blends numerical streams, categorical logs, and qualitative subjective reports (Mohr et al., 2017; Shiffman et al., 2008). Our experiments (Section 4) show that both text-only LLM approaches and generic time-series VLMs exhibit limited cross-distribution generalizability on LE data.

Given these limitations, we investigate whether visual encoding can recover the structural information that serialization loses, not by replacing text-based processing but by complementing it with a second modality that preserves what linearization discards. Research on visuospatial reasoning shows that external visualizations such as charts and diagrams do not simply display information in an alternative format; they reorganize it so that structural relationships become perceptually accessible through spatial proximity, alignment, and grouping (Tversky, 2005, 2011). We apply this principle to structured temporal data. By encoding the data matrix as a visual image, structural relationships that are invisible in token sequences become spatial patterns that vision encoders are equipped to detect. Pairing such visual representations with textual descriptions that capture interpreted semantic context creates a multimodal input in which each modality contributes what the other lacks.

To realize this approach, we introduce **LE-Viz**, a framework that transforms LE data into two complementary modalities for a pre-trained VLM. (1) Spatially-encoded visualizations (line charts or heatmap chains) that map the $Features \times Time$ matrix into interpretable images, preserving temporal adjacency as pixel-level spatial proximity. (2) Textual meta-narratives that summarize significant behavioral trends and global context. A VLM (LLaVA-NeXT 7B (Liu et al., 2024a)) is fine-tuned to process these multimodal inputs and generate *prospective narratives* (Moulton and Kosslyn, 2009; Schacter et al., 2008), natural language descriptions of forecasted states in place of classification labels. This formulation yields outputs that can be parsed for categorical predictions while retaining the reasoning trace behind them.

We evaluate LE-Viz on cross-distribution forecasting tasks across three diverse LE datasets spanning mental health and educational behavior prediction. GLOBEM (Xu et al., 2023a), a widely-adopted benchmark for cross-dataset generalization; LifeSnaps (Yfantidou et al., 2022); and

MFAFY (Hayat et al., 2024a,b; Thach et al., 2025).

Our main contributions are:

- I. Evidence that visual encoding of structured temporal data complements text-based LLM approaches by preserving two-dimensional relational structure that linearization destroys, leading to substantially stronger cross-distribution generalization.
- II. A multimodal framework (LE-Viz) that pairs spatially-encoded visualizations with textual meta-narratives and extends prospective narrative generation to multimodal inputs, conditioning forecasting on complementary visual and textual representations.
- III. New state-of-the-art (SOTA) on GLOBEM with 72.86% out-of-distribution accuracy, improving over the traditional SOTA of 52.8% by over 20 percentage points and over the best text-only LLM baseline by 5.46 percentage points, and over a generic time-series VLM baseline, with consistent gains across all datasets.
- IV. Ablation evidence that spatial organization of visual inputs, not information quantity, drives the generalization gain, and that text-vision complementarity yields improvements beyond what either modality achieves alone.

2 Related Work

LLMs for Structured and Longitudinal Data.

Several approaches apply LLMs to LE data through prompting and fine-tuning on health datasets (Kim et al., 2024), behavioral forecasting with diverse architectures (Hayat et al., 2024a,b), zero-shot multi-label prediction (Thach et al., 2025), and contextual meta-narrative generation for cross-distribution forecasting (Hayat et al., 2025). These approaches serialize structured temporal data into token sequences, inheriting the linearization limitations discussed in Section 1. Traditional ML and DL methods trained on LE data similarly struggle with cross-distribution generalization (Xu et al., 2023a,b), and domain generalization techniques (Pan and Yang, 2010; Zhou et al., 2022) have shown limited success on complex behavioral data where distribution shifts span cohorts, time periods, and contexts.

Visual Encoding and Multimodal Approaches for Temporal Data. An alternative to text serialization is to encode structured data visually, allowing VLMs to process spatial patterns alongside linguistic context. VLMs pre-trained on large-scale image-

text corpora (Radford et al., 2021; Alayrac et al., 2022) have been extended to non-photographic visual content including documents (Xu et al., 2020, 2021b) and charts (Masry et al., 2022; Islam et al., 2024). VLMs have also been applied to time-series forecasting by converting numerical sequences into visual format (Zhong et al., 2025). Our approach differs from this line of research in **three respects**. We transform heterogeneous data combining numerical, categorical, and qualitative modalities; we target cross-distribution generalization across cohorts and time periods; and we design visual encodings that preserve two-dimensional relational structure beyond simple temporal trajectories. Our representational design draws on cognitive research on visuospatial reasoning (Tversky, 2005, 2011), as discussed in Section 1. An extended review is provided in Appendix A.7.

3 Methodology

3.1 Problem Formulation

We address cross-distribution generalization for heterogeneous sequential data, where models trained on a source distribution (T) are evaluated on a target distribution (T') drawn from different cohorts or time periods. As discussed in Section 1, standard approaches serializing temporal sequences into one-dimensional token streams lose the two-dimensional structural relationships supporting generalization across distributions.

For data collected from N individuals over K weeks (constituting T_{total} daily time steps), each individual $i \in \{1, \dots, N\}$ at daily time step $j \in \{1, \dots, T_{total}\}$ has a feature vector $x_{i,j}$ encompassing diverse modalities (numerical sensor streams, categorical logs, qualitative reports). Following standard practice (Xu et al., 2023a), we segment each individual’s K -week data into overlapping k -week input sequences using a sliding window, starting at week s and ending at week $s + k - 1$, generating $K - k$ sequences per individual. Each sequence corresponds to raw data $\{x_{i,j}\}$ for all daily time steps within that window.

Our approach transforms raw LE sequences into multimodal representations $(X_{i,s\dots s+k-1}^{text}, X_{i,s\dots s+k-1}^{visual})$ where visual encodings preserve temporal structure through spatial locality. The textual component X^{text} provides semantic context through meta-narratives, while the visual component X^{visual} encodes temporal trajectories as spatial patterns in charts and

heatmaps. Instead of classification, we formulate forecasting as prospective narrative generation, where the model outputs natural language descriptions $y_{i,s+k}^{text}$ of the subsequent week’s state. The problem is to learn a function f instantiated by a fine-tuned VLM that maps multimodal representations to prospective narratives with effective cross-distribution generalization:

$$f : (X_{i,s\dots s+k-1}^{text}, X_{i,s\dots s+k-1}^{visual}) \rightarrow y_{i,s+k}^{text}$$

3.2 The LE-Viz Framework

LE-Viz is a multimodal framework designed for generalizable modeling of LE data through three stages (Figure 1): Multimodal Input Transformation, Prospective Narrative Generation, and Model Adaptation and Inference. The first stage transforms raw heterogeneous sequential data into complementary textual and visual representations that preserve the data’s two-dimensional structure.

Multimodal Input Transformation. For each k -week input sequence, we transform raw LE data into two components that capture both the two-dimensional *Feature* $\times$ *Time* organization and qualitative nuances discarded by standard serialization approaches, leveraging distinct VLM processing capabilities.

Textual Meta-Narrative Generation. Previous approaches to textualizing LE data include verbalizing the complete detailed sequence (Hayat et al., 2024a,b), generating a concise summary (Thach et al., 2025), structuring in a clustered format (Kim et al., 2024), and creating an interpretative meta-narrative of global context and significant trends (Hayat et al., 2025). While the first three methods provide 1D text inputs that may not fully capture interwoven temporal trends due to token-level linearization, the **Meta-Narrative** approach generates a concise summary across all features (e.g., “Mood declined as sleep decreased”), distilling high-dimensional data into a semantically rich form that emphasizes (i) temporal patterns and trends observed in individual features; (ii) relationships and interactions between different features over time; and (iii) contextual interpretations or anomalies based on domain knowledge. LE-Viz adopts this meta-narrative as its textual component and pairs it with visual encodings. We use GPT-4o (OpenAI, 2024) to generate these meta-narratives, which also serve as a standalone text-only baseline in our experiments.

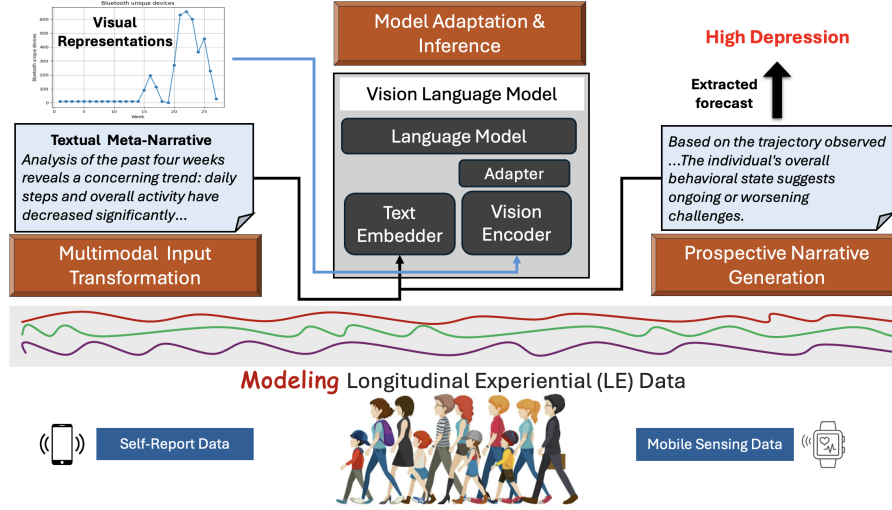


Figure 1: Overview of the LE-Viz Framework. Raw heterogeneous sequential data is transformed into multimodal representations combining textual meta-narratives and visual encodings (charts, heatmaps) through **Multimodal Input Transformation**. A pre-trained VLM is fine-tuned with prospective narratives as training targets for the **Prospective Narrative Generation** task. During inference, the fine-tuned VLM generates predictive narratives from which forecasts are extracted.

Visual Representation Generation. The visual component explicitly encodes feature dynamics through spatial organization, preserving temporal adjacency as spatial adjacency so that vision encoders can process temporal patterns through local patch operations. We generate two complementary visualization types (Figure 2):

- **Feature Trajectory Line Charts:** A standard two-dimensional plot visualizing the feature’s value over time, mapping time to the horizontal axis and value to the vertical axis. This format effectively captures temporal trajectories as spatial patterns (e.g., slope, curvature) and explicitly shows exact values and trend details.
- **Feature Heatmap Chains:** A visual chain (or sequence) of colored markers organized along a temporal axis. Each marker corresponds to a time step, and its color intensity encodes the feature’s value at that time step. This format captures the temporal progression of values through color variations, emphasizing relative changes via color perception rather than position on a second spatial axis, and can be particularly effective for VLMs trained on diverse visual patterns.

We use mean imputation for missing values following standard practice (Xu et al., 2023a). The multimodal input is structured in an **interleaved format**. The global meta-narrative followed by feature-specific modules, each comprising a brief feature-specific text summary and its corresponding visualization. This organization maintains spa-

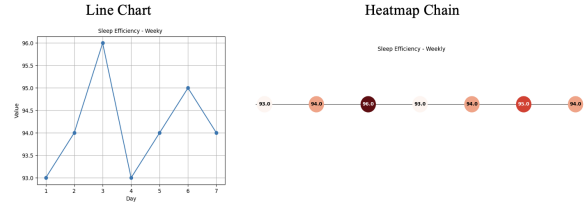


Figure 2: Examples of per-feature visual representations for a single feature’s k-week trajectory. **Left:** Feature trajectory line chart showing detailed value variations over time. **Right:** Feature heatmap chain displaying temporal progression through color intensity.

tial proximity between related textual and visual information, facilitating cross-modal attention in the VLM. An alternative combined format, where all feature visualizations are arranged in a single image, is evaluated in Section 4.2.

We note that the textual and visual components encode overlapping content from the same underlying data, but through different representational structures. The meta-narrative compresses high-dimensional trajectories into interpreted semantic summaries, while the visualizations preserve local temporal structure as spatial patterns. Whether these complementary affordances lead to gains beyond what either modality provides alone is an empirical question we address through component isolation ablations in Section 4.

Prospective Narrative Generation. We formulate LE data forecasting as prospective narrative generation in natural language (Hayat et al., 2025), instead of traditional classification or regression (Xu et al., 2023a). LE-Viz extends this formulation to multimodal inputs, leveraging VLMs’ alignment

between spatial visual patterns and linguistic descriptions.

For each k -week sequence, the VLM generates a prospective narrative incorporating the forecasted state for the subsequent week. This narrative serves as a holistic assessment that can be later parsed for specific predictions. To create training targets ($y_{i,s+k}^{\text{text}}$), we use GPT-4o to generate narrative reflections on past trajectories aligned with known actual states for week $s + k$.

This generative formulation requires the VLM to integrate information across modalities. The model attends to visual patterns in the charts (capturing fine-grained temporal dynamics), interprets the meta-narrative (providing global context), and synthesizes this information into a coherent linguistic forecast. Through the causal self-attention mechanism operating over the interleaved text and visual token sequence, the model learns to ground linguistic concepts (e.g., “declining mood”) in visual patterns (e.g., downward slopes in mood charts), creating representations that capture both semantic meaning and temporal dynamics. During evaluation, these generated narratives are processed using an LLM to extract specific forecasts (e.g., binary depression level) for comparison with ground truth, as detailed below.

Model Adaptation and Inference. The final stage adapts a pre-trained VLM (LLaVA-NeXT 7B (Liu et al., 2024a)) for our forecasting task through fine-tuning. LLaVA-NeXT combines a SigLIP-SO400M vision encoder (Zhai et al., 2023) with a Qwen1.5-7B-Chat language model (Bai et al., 2023), connected through a two-layer MLP projector that maps visual tokens into the language model’s embedding space (Li et al., 2024). During fine-tuning, we update both the projector parameters and the language model parameters using Low-Rank Adaptation (LoRA) (Hu et al., 2021) with rank 64 and alpha 128, applied to all attention modules in the language model, while keeping the vision encoder frozen. Freezing the vision encoder preserves pre-trained visual features while allowing the projector and language model to adapt cross-modal integration for the specific representational demands of abstract temporal visualizations. The model is trained for 20 epochs with a learning rate of 1×10^{-4} and cosine decay. All visual encodings use fixed axis scales, standardized feature ordering, and uniform color mappings across samples to ensure that learned patterns reflect structural rela-

tionships in the data, not artifacts of visualization formatting.

The VLM processes the multimodal input through the following stages. (1) The vision encoder independently processes each chart or heatmap image, extracting a sequence of patch embeddings. (2) The projector maps these visual embeddings into the language model’s token space. (3) The language model processes the interleaved sequence of projected visual tokens and text tokens (from the meta-narrative and feature descriptions) through causal self-attention, allowing cross-modal interaction within the unified token sequence. (4) The model generates the prospective narrative autoregressively, conditioned on this multimodal representation. At inference, we use greedy decoding (temperature 0) with a maximum generation length of 2000 tokens.

At test time, we apply the same input transformation pipeline to create multimodal representations. The fine-tuned VLM generates predictive narratives for the subsequent period. For quantitative evaluation, an external LLM (GPT-4o) performs structured extraction on these narratives, parsing them into categorical predictions (e.g., “high” or “low” depression level) that are compared against ground truth labels. This extraction step is a constrained parsing task (mapping narrative text to predefined categories) that is independent of the narrative generation used to create training targets; the evaluation comparison is between extracted predictions and held-out ground truth, not between model outputs and GPT-4o’s own generations. To verify that evaluation results are not sensitive to the extraction model, we repeat the extraction using the open-source Llama 3.1 8B Instruct (Grattafiori et al., 2024). The two extraction pipelines show strong agreement, with BERTScore (Zhang et al., 2020) similarity exceeding 95%, confirming that the evaluation is not dependent on a proprietary model. We discuss reproducibility considerations regarding GPT-4o usage for meta-narrative and target generation in Section 6. For implementation details including optimization settings and full hyperparameter configuration, see Appendix A.1.

4 Experiments and Results

Datasets. We evaluate LE-Viz on three diverse LE datasets that pose distinct challenges for cross-distribution generalization from structured temporal data. **GLOBEM** (Xu et al., 2023a) comprises

661 person-years from 497 participants across two institutions and four years, with 15 mobile sensing features (location, phone usage, physical activity, sleep) for mental health forecasting with cross-cohort and cross-temporal generalization. **LifeSnaps** (Yfantidou et al., 2022) includes 71 participants over 4 months with over 35 Fitbit sensing and Ecological Momentary Assessment features for anxiety prediction with cross-temporal and cross-participant generalization. **MFAFY** (Hayat et al., 2024a,b; Thach et al., 2025) spans 96 college students across two academic years with 28 qualitative non-cognitive dimensions for educational engagement forecasting with cross-year generalization. While GLOBEM and LifeSnaps contain structured numerical sensor data, MFAFY consists purely of qualitative self-reports, representing a particularly challenging case for visual encoding.

All experiments follow strict cross-distribution evaluation protocols, with 15% of data from the training period (T) forming the in-distribution test set and 100% of data from the distinct period (T') constituting the out-of-distribution (OOD) test set. This protocol ensures no overlap between training and evaluation samples. Detailed preprocessing, feature sets, and split configurations appear in Appendix A.3.

Baselines. We compare LE-Viz against baselines spanning three categories. **Time-series models:** PatchTST (Nie et al., 2022) and iTransformer (Liu et al., 2024c), two SOTA transformer-based architectures that process numerical sequences without language or vision components. **LLM baselines** using different textualization strategies: (i) *LLM-Complete* (Hayat et al., 2024a,b), which verbalizes the full sequence into natural language; (ii) *LLM-Summary* (Thach et al., 2025), which generates concise summaries using standard techniques (Zhang et al., 2023); (iii) *LLM-Meta* (Hayat et al., 2025), which uses a textual meta-narrative summarizing behavioral trends; this representation also serves as the textual component of our multimodal input (Section 3.2); and (iv) *LLM-Clustered* (Kim et al., 2024), which presents features sequentially with their values. **Multimodal time-series model:** Time-VLM (Zhong et al., 2025), which integrates numerical time-series inputs with vision-language modeling components. For additional context, we reference the best reported traditional ML/DL result on GLOBEM (52.8% OOD accuracy) (Xu et al., 2023b). Detailed baseline descriptions ap-

pear in Appendix A.3.

Evaluation Metrics. As detailed in Section 3.2, an external LLM (GPT-4o) performs structured extraction on each generated narrative ($y_{i,s+k}^{\text{text}}$), parsing it into a categorical prediction (e.g., binary depression, anxiety, or engagement level). We compute standard classification metrics by comparing extracted forecasts to ground truth labels ($y_{i,s+k}^{\text{actual}}$). Given the nearly balanced datasets, we report accuracy, precision, recall, and F1-score.

Implementation Details. LE-Viz is implemented using the Hugging Face Transformers library. The core VLM is LLaVA-NeXT 7B (Liu et al., 2024a). Baseline LLM experiments use Llama 3.1 8B Instruct (Grattafiori et al., 2024). Textual meta-narrative generation, target narrative creation for training, and forecast extraction for evaluation use the GPT-4o API. Visual representations (charts and heatmaps) are generated using Matplotlib and Seaborn. Experiments run on a server with $8 \times$ NVIDIA A40 48GB GPUs. Specific hyperparameters are provided in Appendix A.1.

4.1 Results

GLOBEM. Table 1 presents results for the GLOBEM mental health forecasting task. Time-series baselines (PatchTST, iTransformer) achieve OOD accuracies near 50%, comparable to the best reported traditional ML/DL result of 52.8% (Xu et al., 2023b). Time-VLM, despite incorporating vision-language components, performs similarly (51.43% OOD), suggesting that generic multimodal time-series architectures do not address the specific representational challenges of LE data. Among text-only methods, OOD performance ranges from 62.43% to 67.40%, with LLM-Meta (Hayat et al., 2025) achieving the strongest text-only OOD Accuracy (67.40%) and F1 (69.40%). LE-Viz (Chart) achieves the highest OOD performance (Acc **72.86%**, F1 **74.92%**), a gain of 5.46 percentage points over the best text-only baseline. LE-Viz (Heatmap) also surpasses all baselines (Acc 70.98%, F1 73.16%).

LifeSnaps. For the LifeSnaps stress level forecasting task (Table 1), time-series baselines and Time-VLM again achieve limited OOD performance (43–48%), while LLM-Meta achieves the strongest text-only OOD results (Acc 67.19%, F1 68.66%). LE-Viz (Chart) attains the highest OOD performance (Acc **71.88%**, F1 **76.92%**), surpassing all baselines across both metrics. LE-Viz (Heatmap) also

Table 1: Cross-distribution generalization results ($T \rightarrow T'$) across all datasets. Bold indicates best performance for each dataset and metric category.

Dataset	Shift	Method	In-Distribution (ID) Test				Out-of-Distribution (OOD) Test			
			Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
GLOBEM	Years 1&2 $\rightarrow$ Years 3&4	<i>Time-Series Models</i>								
		PatchTST (Nie et al., 2022)	53.58	53.73	53.58	53.01	49.88	49.24	49.88	49.16
		iTransformer (Liu et al., 2024c)	54.61	54.61	54.61	54.61	51.06	51.07	51.06	51.07
		<i>LLM Models</i>								
		LLM-Complete (Hayat et al., 2024a,b)	69.96	71.56	68.42	69.96	65.94	67.95	68.52	68.23
		LLM-Summary (Thach et al., 2025)	69.51	72.22	67.24	69.65	62.43	65.97	63.57	64.75
		LLM-Clustered (Kim et al., 2024)	70.05	71.30	69.37	70.32	66.44	67.92	69.09	68.50
		LLM-Meta (Hayat et al., 2025)	73.99	75.93	71.93	73.87	67.40	68.81	70.00	69.40
		<i>VLM Models</i>								
		Time-VLM (Zhong et al., 2025)	57.34	55.92	56.48	56.20	51.43	49.87	50.21	50.04
		LE-Viz (Text: Meta + Visual: Heatmap)	76.68	72.83	71.28	72.04	70.98	73.19	73.13	73.16
		LE-Viz (Text: Meta + Visual: Chart)	79.37	83.33	80.77	82.03	72.86	75.58	74.28	74.92
LifeSnaps	First 2 Months $\rightarrow$ Last 2 Months	<i>Time-Series Models</i>								
		PatchTST (Nie et al., 2022)	47.83	47.83	47.83	47.83	43.75	41.23	43.75	40.47
		iTransformer (Liu et al., 2024c)	52.17	46.81	52.17	48.83	48.44	48.14	48.44	47.99
		<i>LLM Models</i>								
		LLM-Complete (Hayat et al., 2024a,b)	58.82	77.78	58.33	66.67	54.84	50.00	57.14	53.33
		LLM-Summary (Thach et al., 2025)	47.06	40.00	57.14	47.06	46.88	36.67	42.31	39.29
		LLM-Clustered (Kim et al., 2024)	70.59	80.00	72.72	76.19	62.50	52.94	69.23	60.00
		LLM-Meta (Hayat et al., 2025)	64.71	77.78	63.64	70.00	67.19	63.89	74.19	68.66
		<i>VLM Models</i>								
		Time-VLM (Zhong et al., 2025)	49.28	47.11	47.86	47.48	47.36	45.02	45.71	45.36
		LE-Viz (Text: Meta + Visual: Heatmap)	76.47	83.33	62.50	71.43	69.23	72.41	63.64	67.74
		LE-Viz (Text: Meta + Visual: Chart)	82.35	85.71	75.00	80.00	71.88	83.33	71.43	76.92
MFAFY	Year 1 $\rightarrow$ Year 2	<i>Time-Series Models</i>								
		PatchTST (Nie et al., 2022)	67.39	63.66	67.39	64.34	50.57	49.68	50.57	44.31
		iTransformer (Liu et al., 2024c)	53.26	51.77	53.26	52.47	44.29	43.05	44.29	42.42
		<i>LLM Models</i>								
		LLM-Complete (Hayat et al., 2024a,b)	60.66	56.67	60.71	58.62	57.14	50.55	60.53	55.09
		LLM-Summary (Thach et al., 2025)	57.38	48.28	56.00	51.85	53.43	52.02	52.94	52.48
		LLM-Clustered (Kim et al., 2024)	63.93	62.96	69.37	60.71	62.86	57.47	64.10	60.61
		LLM-Meta (Hayat et al., 2025)	70.49	65.22	60.00	62.50	64.86	61.11	67.48	64.14
		<i>VLM Models</i>								
		Time-VLM (Zhong et al., 2025)	53.41	51.76	52.33	52.04	48.27	46.14	46.82	46.48
		LE-Viz (Text: Meta + Visual: Heatmap)	72.13	69.57	61.54	65.31	65.43	62.94	56.96	59.80
		LE-Viz (Text: Meta + Visual: Chart)	77.05	75.76	80.65	78.12	66.57	64.47	60.87	62.62

outperforms text-only methods on OOD Accuracy (69.23%), though with lower F1 (67.74%). These results support the effectiveness of visual encoding for cross-distribution generalization when the underlying data contains continuous numerical sensor readings with temporal trajectories that benefit from spatial representation.

MFAFY. On the MFAFY educational engagement forecasting task (Table 1), the picture shifts. MFAFY differs from other datasets in that its inputs consist purely of qualitative self-reports with no numerical sensor data, providing a test of visual encoding’s boundary conditions. Time-series baselines and Time-VLM show limited OOD performance (44–50%), consistent with the absence of continuous numerical signals. Both LE-Viz variants achieve higher OOD Accuracy than the best text-only baseline (Chart 66.57%, Heatmap 65.43% vs. LLM-Meta 64.86%). However, LLM-Meta achieves the highest OOD F1 (64.14%), while LE-Viz Chart reaches 62.62%. This pattern suggests

that visual encoding provides consistent accuracy gains across data types, but the magnitude of improvement is modulated by whether the underlying features contain continuous temporal dynamics that benefit from spatial representation. When features are categorical, the visual channel’s advantage narrows.

4.2 Ablation Study

We conduct ablation studies analyzing three dimensions of LE-Viz’s design. Table 2 summarizes key results; full tables and detailed analysis are in Appendix A.8.

(1) VLM architecture. Using the optimal input configuration (Text: Meta-Narrative + Visual: Interleaved Charts), we compare three VLMs on GLOBEM, selected to vary along the architectural dimensions most relevant to visual encoding: vision encoder type (SigLIP vs. CLIP), language model integration (instruction-tuned vs. frozen), and parameter scale (7B–13B). Not all ben-

Table 2: Summary of ablation results. All values are OOD Accuracy (%).

Comparison	OOD Acc (%)
(1) VLM Architecture (GLOBEM)	
LLaVA-NeXT 7B	72.86
MiniGPT-4 13B	66.68
mPLUG-Owl2 8.2B	65.74
Text-only Baseline	67.40
(2) Visual Organization (GLOBEM)	
Interleaved Chart	72.86
Combined Chart	62.48
(3) Component Analysis (LifeSnaps)	
Text-only	67.0
Vision-only	57.3
Multimodal (Text+Vision)	71.9

efit equally from visual encoding. LLaVA-NeXT 7B (Liu et al., 2024a) (72.86% OOD) outperforms both MiniGPT-4 13B (Zhu et al., 2023) (66.68%) and mPLUG-Owl2 8.2B (Ye et al., 2023) (65.74%) despite fewer parameters. Both MiniGPT-4 and mPLUG-Owl2 underperform the text-only baseline on OOD F1 (see Appendix A.8, Table 4), suggesting that a poorly aligned vision encoder can degrade generalization. Vision encoder quality and cross-modal alignment are important when applying visual encoding to structured temporal data.

(2) Visual organization. Interleaved presentation (72.86% OOD) outperforms combined single-image format (62.48% OOD) by over 10 percentage points, despite containing identical visual information. The same pattern holds for Heatmap representations (70.98% vs. 66.73%, see Appendix A.8, Table 5), indicating that the organizational effect generalizes across visualization types. Since both formats encode the same data, the gap cannot be attributed to the visual channel providing additional information; it indicates that spatial representational structure drives the generalization gain. In the interleaved format, visual tokens for each feature appear adjacent to the corresponding text, supporting local cross-modal attention; the combined format groups all visual tokens together, requiring long-range correspondences that are less robust under distribution shift.

(3) Component analysis. Controlled comparison on LifeSnaps with identical hyperparameters shows that Text-only (67.0% OOD) outperforms Vision-only (57.3%), but the multimodal combination (71.9%) exceeds both, indicating that the visual channel contributes complementary information that text alone does not capture. This pattern is consistent across all three VLMs, with multimodal gains over text-only ranging from 4.9 to 6.7 percentage points OOD.

5 Discussion

The experimental findings point to three complementary mechanisms through which visual encoding supports generalization. **First**, as argued in Section 1, text serialization separates temporally adjacent values by $O(N)$ tokens for N -feature data; visual encoding reduces this to local patch operations, and the interleaved vs. combined ablation confirms that this locality translates to measurable OOD gains. **Second**, pre-trained visual features transfer to abstract visualizations. The 7.12 percentage point OOD gap between LLaVA-NeXT (SigLIP) and mPLUG-Owl2 (CLIP) persists with frozen encoders, suggesting that the quality of low-level visual primitives learned from natural images is consequential for temporal pattern detection in charts. **Third**, interleaved organization supports local cross-modal grounding. The interleaved advantage persists under distribution shift (10.38 percentage points OOD vs. 12.11 ID), confirming that local text-vision alignment remains beneficial on unseen data. The contrast with Time-VLM is instructive. Despite incorporating vision-language components, Time-VLM performs at or below pure time-series baselines on all three datasets. Generic multimodal integration without purpose-designed visual encoding does not address the representational challenges of LE data.

The component analysis shows an important asymmetry. Text alone outperforms vision alone across all VLMs, but the multimodal combination consistently exceeds both. Visual encoding contributes complementary information that text captures poorly, even when vision is not independently sufficient. These mechanisms are strongest when the underlying features contain continuous temporal dynamics (GLOBEM, LifeSnaps); the narrower gains on MFAFY’s categorical features suggest that spatial continuity in the visualized data is one condition for the visual channel to provide substantial benefit. For structured temporal data with inherent two-dimensional organization, spatial visual representations offer a complementary pathway that extends what text-based processing can achieve. A natural next step is to investigate whether learned visual encodings, where the model optimizes the visualization parameters jointly with the forecasting objective, could extend the benefits of spatial representation to data types where fixed chart formats show limited advantage.

6 Limitations

Empirical Scope. While our evaluation spans three datasets across different domains (mental health, stress, educational engagement) and data modalities (numerical sensing, qualitative self-reports), all tasks involve binary prediction with specific distribution shift scenarios. The generality of our findings would be strengthened by evaluation on datasets with different characteristics, including varying sparsity patterns, longer prediction horizons, multi-class or continuous outcomes, and larger participant populations. We evaluate a limited set of VLM architectures (three) and LLM baselines (four textualization strategies); testing with a broader range of models, including more recent open-weight VLMs, would provide a fuller picture of when and why visual encoding benefits cross-distribution generalization. Our only VLM-based time-series baseline is Time-VLM, which was designed for purely numerical forecasting rather than heterogeneous LE data; comparison with additional multimodal temporal architectures as they become available would strengthen the evaluation. As reported in Section 4.1, the text-only baseline achieves higher OOD F1 than LE-Viz on MFAFY, indicating that the multimodal approach does not uniformly dominate across all metrics and data types, particularly for purely qualitative features.

Dependence on GPT-4o. The framework relies on GPT-4o for three components: meta-narrative generation (input creation), prospective narrative generation for training targets, and forecast extraction during evaluation. This introduces a dependency on a proprietary model, raising concerns about reproducibility and cost. Meta-narrative generation and training target creation are preprocessing steps whose outputs can be cached and distributed with the dataset, and our Llama 3.1 extraction validation (Section 3.2) confirms that the evaluation is not sensitive to the extraction model. Nonetheless, evaluating the full pipeline with open-weight alternatives for all three components is an important direction for ensuring broader reproducibility. Additionally, the quality of generated meta-narratives and training targets may be sensitive to prompting strategies; we use a fixed set of prompts across all experiments, but systematic analysis of prompt sensitivity remains an open question.

Narrative Generation Formulation. Our ap-

proach formulates prediction as prospective narrative generation followed by structured extraction, introducing an intermediate natural language step between input processing and categorical output. This formulation offers potential interpretability benefits (the generated narrative can be inspected), but it also introduces a potential source of error. Predictions are mediated by the model’s language generation quality and the extraction step’s parsing accuracy. We have not compared this formulation against classification heads on the same multimodal representations, which would clarify whether the generalization gains are attributable to the multimodal representation, the generation formulation, or their interaction.

Visual Encoding Design. The visual representations used in LE-Viz (line charts, heatmap chains) are informed by domain knowledge and cognitive principles of visuospatial reasoning, but they have not been systematically optimized. Design parameters such as axis scaling, color mapping, line thickness, image resolution, and aspect ratio may influence how effectively vision encoders process temporal patterns, and a principled study of these parameters’ effects on downstream performance has not been conducted. The MFAFY results suggest that visual encoding is less beneficial for categorical features lacking continuous temporal dynamics, but we have not characterized the full range of data properties (feature types, dimensionality, temporal granularity) that modulate the effectiveness of visual encoding.

Interpretability of Visual Feature Transfer. The VLM comparison ablation provides indirect evidence that pre-trained visual features transfer to abstract temporal visualizations, but validating what specific visual primitives the encoder applies to temporal charts, and how attention patterns differ between interleaved and combined formats, would require feature attribution analysis or attention visualization methods that we have not yet applied. Such analysis could also clarify whether the vision encoder processes temporal charts through the same mechanisms it uses for natural images or through distinct computational pathways.

Privacy, Fairness, and Ethical Considerations. LE data contains sensitive personal information including mental health indicators, behavioral patterns, and educational outcomes. Using powerful generative models on such data raises concerns about potential memorization of sensitive training

examples, amplification of biases present in the data or pre-trained model weights, and the reliability of generated narratives when used for consequential predictions. Our current evaluation does not include fairness audits across demographic subgroups or analysis of whether the model’s generalization patterns differ across populations. Responsible deployment of LE-Viz or similar frameworks in applied settings would require systematic bias evaluation, privacy-preserving training procedures, and domain expert validation of generated narratives before any predictions inform real-world decisions.

Computational Requirements. Fine-tuning LE-Viz requires multiple high-end GPUs ($8 \times$ NVIDIA A40 48GB), and the multimodal input format (interleaved text and images) increases both training time and inference-time memory relative to text-only approaches. The interleaved format, while giving the strongest generalization results, processes more visual tokens per sample than a combined single-image format, creating a tradeoff between representational quality and computational efficiency. This may affect scalability for applications requiring low-latency or real-time processing, and for researchers with constrained computational resources.

Acknowledgments

This research was supported by grants from the U.S. National Science Foundation (NSF DUE 2142558), the U.S. National Institutes of Health (NIH NIGMS P20GM130461 and NIH NIAAA R21AA029231), and the Rural Drug Addiction Research Center at the University of Nebraska-Lincoln.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: a visual language model for few-shot learning](#). *Preprint*, arXiv:2204.14198.
- S. Bae, D. Ferreira, B. Suffoletto, J. C. Puyana, R. Kurtz, T. Chung, and A. K. Dey. 2017. [Detecting drinking episodes in young adults using smartphone-based sensors](#). *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(2):5.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. [Qwen technical report](#). *Preprint*, arXiv:2309.16609.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. [Multimodal machine learning: A survey and taxonomy](#). *Preprint*, arXiv:1705.09406.
- P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, M. Goel, and A. K. Dey. 2021. [Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection](#). *ACM Trans. Comput.-Hum. Interact.*, 28(1):3:1–3:41.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Ahatsham Hayat, Bilal Khan, and Mohammad Hasan. 2024a. [Improving transfer learning for early forecasting of academic performance by contextualizing language models](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 137–148, Mexico City, Mexico. Association for Computational Linguistics.
- Ahatsham Hayat, Bilal Khan, and Mohammad Rashedul Hasan. 2024b. [Leveraging language models for analyzing longitudinal experiential data in education](#). *arXiv:2503.21617*.
- Ahatsham Hayat, Bilal Khan, and Mohammad Rashedul Hasan. 2025. [ConText-LE: Cross-distribution generalization for longitudinal experiential data via narrative-based LLM representations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15335–15360, Suzhou, China. Association for Computational Linguistics.
- Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. [Tabllm: Few-shot classification of tabular data with large language models](#). In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.
- Jonathan Herzig, Pawel K. Nowak, Thomas Müller, Francesco Piccinno, and Julian M. Eisenschlos. 2020. [Tapas: Weakly supervised table parsing via pre-trained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-Rank Adaptation of Large Language Models](#). *arXiv preprint*. ArXiv:2106.09685 [cs].

- Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. [Are large vision language models up to the challenge of chart comprehension and reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA. Association for Computational Linguistics.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. [Scaling up visual and vision-language representation learning with noisy text supervision](#). *Preprint*, arXiv:2102.05918.
- Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. [Health-llm: Large language models for health prediction via wearable sensor data](#). *Preprint*, arXiv:2401.06866.
- Reed Larson and Mihaly Csikszentmihalyi. 1983. The experience sampling method. *New Directions for Methodology of Social and Behavioral Science*, 15:41–56.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. [Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models](#). <https://llava-vl.github.io/blog/2024-06-16-llava-next-interleave/>. Accessed: YYYY-MM-DD.
- Xianling Li, Xiaoyan Zhu, Xiaofei Zhu, Yunkai Ji, and Xiangnan Tang. 2020. [Student academic performance prediction using deep multi-source behavior sequential network](#). In *Advances in Knowledge Discovery and Data Mining. PAKDD 2020*, volume 12084 of *Lecture Notes in Computer Science*, pages 570–582. Springer, Cham.
- Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. [Pmc-clip: Contrastive language-image pre-training using biomedical documents](#). *Preprint*, arXiv:2303.07240.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. [Llava-next: Improved reasoning, ocr, and world knowledge](#).
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2024c. [itransformer: Inverted transformers are effective for time series forecasting](#). *Preprint*, arXiv:2310.06625.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [ChartQA: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- David C. Mohr, Michelle Zhang, and Stephen M. Schueller. 2017. [Personal sensing: Understanding mental health using ubiquitous sensors and machine learning](#). *Annual Review of Clinical Psychology*, 13:23–47.
- Samuel T. Moulton and Stephen M. Kosslyn. 2009. [Imaging predictions: Mental imagery as mental emulation](#). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1273–1280.
- Ehsan Nemati, Xiangyu Xu, Varun Nathan, Kaveh Vatanparvar, Tanzima Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jing Kuang, and Anhong Gao. 2022. [Ubilung: Multi-modal passive-based lung health assessment](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 551–555. IEEE.
- Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. [A time series is worth 64 words: Long-term forecasting with transformers](#). *ArXiv*, abs/2211.14730.
- OpenAI. 2024. [Hello gpt-4o](#). <https://openai.com/index/hello-gpt-4o/>. Accessed: 2024-09-02.
- Sinno Jialin Pan and Qiang Yang. 2010. [A Survey on Transfer Learning](#). *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. Conference Name: IEEE Transactions on Knowledge and Data Engineering.
- A. Patharkar, F. Cai, F. Al-Hindawi, and T. Wu. 2024. [Predictive modeling of biomedical temporal data in healthcare applications: review and future directions](#). *Frontiers in Physiology*, 15:1386760.
- M. Rabbi, P. Klasnja, T. Choudhury, A. Tewari, and S. Murphy. 2019. [Optimizing mhealth interventions with a bandit](#). In H. Baumeister and C. Montag, editors, *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*, Studies in Neuroscience, Psychology and Behavioral Economics, pages 277–291. Springer International Publishing, Cham.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *Preprint*, arXiv:2103.00020.
- Sohrab Saeb, Michelle Zhang, Christopher J. Karr, Stephen M. Schueller, Molly E. Corden, Konrad P. Kording, and David C. Mohr. 2015. [Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study](#). *Journal of Medical Internet Research*, 17(7):e175.

- Daniel L. Schacter, Donna Rose Addis, and Randy L. Buckner. 2008. [Episodic simulation of future events: Concepts, data, and applications](#). *Annals of the New York Academy of Sciences*, 1124:39–60.
- Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. [Ecological momentary assessment](#). *Annual Review of Clinical Psychology*, 4:1–32.
- Karan Singhal, Shekoofeh Azizi, Tien-Ju Tu, and et al. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620:172–180.
- Arthur A. Stone and Saul Shiffman. 1994. [Ecological momentary assessment \(ema\) in behavioral medicine](#). *Annals of Behavioral Medicine*, 16(3):199–202.
- Nguyen T. Thach, Patrick Habecker, Anika R. Eisenbraun, W. Alex Mason, Kimberly A. Tyler, Bilal Khan, and Hau Chan. 2025. Muhboost: Multi-label boosting for practical longitudinal human behavior modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Accepted. Available at <https://openreview.net/pdf?id=BAelAyADqn>.
- Barbara Tversky. 2005. Visuospatial reasoning. In Keith J. Holyoak and Robert G. Morrison, editors, *The Cambridge Handbook of Thinking and Reasoning*, pages 209–240. Cambridge University Press, New York, NY, US.
- Barbara Tversky. 2011. [Visualizing thought](#). *Topics in Cognitive Science*, 3(3):499–535.
- Rui Wang, Peilin Hao, Xia Zhou, Andrew T. Campbell, and Gabriella Harari. 2016. [SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students](#). *GetMobile: Mobile Computing and Communications*, 19(4):13–17.
- Rui Wang, Wenbo Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. [Tracking depression dynamics in college students using mobile phone and wearable sensing](#). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):43:1–43:26.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. [Medclip: Contrastive learning from unpaired medical images and text](#). *Preprint*, arXiv:2210.10163.
- X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell, J. Mankoff, and A. K. Dey. 2019. [Leveraging routine behavior and contextually-filtered features for depression detection among college students](#). In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, volume 3, pages 116:1–116:33.
- Xiangyu Xu, Prathyusha Chikersal, Jessica M. Dutcher, Yeganeh S. Sefidgar, Woojin Seo, Michael J. Tumminia, Daniel K. Villalba, Susan Cohen, Kelsy G. Creswell, John D. Creswell, Ali Doryab, Paula S. Nurius, Elizabeth Riskin, Anind K. Dey, and Jennifer Mankoff. 2021a. [Leveraging collaborative-filtering for personalized behavior modeling: A case study of depression detection among college students](#). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):41:1–41:27.
- Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023a. [GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling](#). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):190:1–190:34.
- Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve Riskin, Jennifer Mankoff, and Anind K. Dey. 2023b. [Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization](#). *Preprint*, arXiv:2211.02733.
- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021b. [LayoutLMv2: Multi-modal pre-training for visually-rich document understanding](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.
- Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. [Layoutlm: Pre-training of text and layout for document image understanding](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, page 1192–1200, New York, NY, USA. Association for Computing Machinery.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. [mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration](#). *Preprint*, arXiv:2311.04257.
- Sofia Yfantidou, Christina Karagianni, Stelios Efstathiou, and 1 others. 2022. [Lifesnaps, a 4-month multi-modal dataset capturing unobtrusive snapshots of our lives in the wild](#). *Scientific Data*, 9(1):663.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). *Preprint*, arXiv:2303.15343.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. [Extractive summarization via ChatGPT for faithful summary generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*,

pages 3270–3278, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. 2025. [Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting](#). *Preprint*, arXiv:2502.04395.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. [Domain generalization: A survey](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–20.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

This appendix provides supplementary material organized as follows:

- **Appendix A.1:** Implementation details, including experimental settings and hyperparameters.
- **Appendix A.2:** Comparison of VLM architectures evaluated in the ablation study.
- **Appendix A.3:** Detailed descriptions of datasets and baselines.
- **Appendices A.4–A.6:** LLM prompts used for meta-narrative generation, training target generation, and forecast extraction.
- **Appendix A.7:** Extended related work.
- **Appendix A.8:** Detailed ablation study with full tables and analysis.

A.1 Implementation Details

This appendix provides detailed technical information about the implementation of the LE-Viz framework, expanding on the methodology presented in Section 3.

A.1.1 Problem Formulation Details

For comprehensive notation clarity, we provide additional details on our problem formulation:

For LE data collected from N individuals over a total period T , spanning K weeks with features measured daily (constituting T_{total} daily time steps), we define:

- Individual index: $i \in \{1, \dots, N\}$
- Daily time step index: $j \in \{1, \dots, T_{total}\}$

- Feature vector: $x_{i,j}$ (encompassing diverse modalities and types)
- Total number of weeks: K (typically $K = T_{total}/7$)

The k -week sliding window approach generates $K - k$ input sequences per individual. Each sequence starts at week s and comprises the raw data $\{x_{i,j}\}$ for all daily time steps falling within weeks s through $s + k - 1$, with s ranging from 1 to $K - k$.

A.1.2 Detailed Meta-Narrative Generation Process

The meta-narrative generation process employs a carefully designed prompt structure for the external LLM (GPT-4o). The prompt includes:

1. A context section explaining the nature of LE data and the purpose of the meta-narrative
2. A structured presentation of the k -week input data organized by feature, day, and week
3. Instructions to identify and emphasize:
 - Global patterns across the entire k -week window
 - Feature-specific trajectories and their significance
 - Co-variations and potential relationships between features
 - Contextual interpretation based on domain knowledge relevant to the specific domain (mental health or educational behavior)
4. Guidelines for creating a concise, informative summary (150-250 words) that integrates these elements

For feature-specific meta-text generation, we use a similar but more focused prompt structure targeting the individual feature’s trajectory. These summaries are typically 30-50 words and emphasize key changes, patterns, or anomalies in that specific feature’s values.

A.1.3 Visual Representation Generation Details

For generating the visual representations, we employ Python libraries (Matplotlib for charts and custom visualization code for heatmap chains) with standardized parameters:

Feature Trajectory Line Charts

- Figure dimensions: 640×480 pixels
- X-axis: Days within the k-week window
- Y-axis: Feature value (scaled appropriate to the feature's range)
- Line style: Solid line with circular markers at data points
- Title: Feature name
- Missing values: Represented by breaks in the line (visually indicating data gaps)
- Colors: Consistent color scheme across all features to aid VLM interpretation

Feature Heatmap Chains

- Figure dimensions: 1600×400 pixels (elongated horizontal format)
- Chain structure: Circular markers arranged horizontally, representing consecutive days
- Marker size: 80 pixels diameter
- Color encoding: Normalized feature values mapped to a color gradient:
 - Low values: Cool colors (blues)
 - Moderate values: Neutral colors (yellows, greens)
 - High values: Warm colors (reds)
- Labels: Day numbers displayed beneath each marker
- Title: Feature name and time period

Alternative Visual Organizations In our ablation study (Section 4.2), we explore an alternative visual organization where all feature visualizations for a given instance are combined into a single composite image:

- Multi-feature chart: All feature line charts stacked vertically in a grid layout
- Multi-feature heatmap array: All feature heatmap chains arranged in rows, forming a 2D array

This alternative organization is designed to test whether the VLM benefits more from processing all feature visualizations simultaneously as a unified spatial pattern versus the interleaved approach that associates each feature's visual with its specific textual context.

A.1.4 Language Generation Task Details

To generate training targets ($y_{i,s+k}^{\text{text}}$), we employ a structured approach:

1. For each k-week input sequence with known ground truth for week $s + k$, we craft a specialized prompt for GPT-4o
2. The prompt includes:
 - The complete k-week input data
 - The actual future state or outcome ($y_{i,s+k}^{\text{actual}}$)
 - Instructions to behave like an expert and generate a 3-4 lines narrative that:
 - Reflects on patterns and trajectories in the k-week data
 - Incorporates knowledge of the actual future state without explicitly stating it
 - Highlights features and trends that would plausibly anticipate or align with this future state
 - Uses domain-appropriate language and reasoning
3. The resulting narrative implicitly encodes the future state, serving as our target text

For example, if the actual future state is “low depression,” the generated narrative might focus on positive trends in social activity, improved sleep patterns, and stabilizing mood indicators—without explicitly stating “low depression” but rather describing a trajectory consistent with this outcome.

A.1.5 Prediction Extraction Process

The process of extracting specific predictions from generated narratives uses a structured extraction prompt with:

1. Instructions to behave like an expert identify key indicators in the narrative that suggest the target state
2. Options representing possible outcomes (e.g., “high depression” or “low depression”)
3. Requirements to provide a single, definitive answer (e.g., high or low depression)

A.1.6 VLM Adaptation and Inference Details

Base VLM Specification We utilize LLaVA-NeXT 7B (Liu et al., 2024a) as our base VLM architecture, which comprises:

- Vision encoder: SigLIP
- Language model: Qwen1.5-7B-Chat
- Vision-language connector: MLP-based projector

VLM Fine-tuning Configuration Our fine-tuning process employs the following specifications:

- Parameter-efficient fine-tuning: LoRA (Hu et al., 2021) with:
 - Rank: 64
 - Alpha: 128
 - Target modules: All attention modules in the language model
- Training objective: Causal language modeling with teacher forcing
- Optimizer: AdamW
- Learning rate: 1e-4 with cosine decay schedule
- Warm-up steps: 0.1
- Weight decay: 0.01
- Batch size: 4
- Training epochs: 20
- Mixed precision: bfloat16

VLM Inference Configuration During inference, we use the following generation parameters:

- Decoding strategy: Greedy search
- Maximum generation length: 2000 tokens
- Temperature: 0

LLM Fine-tuning Configuration We utilize Llama 3.1 8B Instruct (Grattafiori et al., 2024) for the baseline experiments. Our fine-tuning process employs the following specifications:

- Parameter-efficient fine-tuning: LoRA (Hu et al., 2021) with:

- Rank: 32
- Alpha: 16
- Target modules: All attention modules in the language model

- Training objective: Causal language modeling with teacher forcing
- Optimizer: paged-AdamW-8bit
- Learning rate: 1e-5 with cosine decay schedule
- Warmup-ratio: 0.1
- Batch size: 8
- Training epochs: 20
- Mixed precision: bfloat16

Training Hardware Training was conducted on $8 \times$ NVIDIA A40 GPUs (48GB each) with distributed data parallelism.

A.2 Comparison of VLM Architectures

This section provides a detailed comparison of the VLMs evaluated in the ablation study (Section 4.2): LLaVA-NeXT 7B (Liu et al., 2024a), MiniGPT-4 13B (Zhu et al., 2023), and mPLUG-Owl2 8.2B (Ye et al., 2023). Understanding their architectural differences, fine-tuning strategies, and multimodal capabilities is crucial for interpreting their varied performance on the LE-Viz task, particularly their ability to process interleaved image-text inputs and generate prospective narratives for generalization.

Table 3 summarizes the key characteristics of each VLM relevant to our application and the observed performance differences. The evaluated VLMs differ significantly in their core components and how they are fine-tuned for multimodal tasks. These differences provide insight into their varying performance on the LE-Viz generalization task. LLaVA-NeXT 7B, built on an instruction-tuned Qwen LLM and designed with robust native support for multimodal interleaving through its architecture and training, appears particularly well-suited for the LE-Viz framework’s requirements, enabling strong integration of textual and visual features. In contrast, MiniGPT-4 13B, despite a larger base LLM, relies on a simpler single-layer adapter and a fine-tuning strategy that freezes the base LLM during multimodal adaptation, which

likely hinders its ability to fully leverage the complex interleaved LE-Viz input structure for generalization. mPLUG-Owl2 8.2B offers good interleaving support and a flexible modular design but uses a raw LLaMA base LLM and a potentially less performant vision encoder compared to LLaVA-NeXT for this domain, which may impact its overall multimodal understanding and narrative generation quality for this specific task. The empirical results in Section 4.2 align with these architectural characteristics, demonstrating that VLMs with robust multimodal integration capabilities, strong base language models, and effective vision encoders (like LLaVA-NeXT 7B) are critical for effectively processing the LE-Viz multimodal representation and achieving superior generalization performance.

A.3 Datasets and Baselines

Datasets We utilize the following LE datasets, selected for their relevance to health and behavioral forecasting and their suitability for evaluating challenging generalization across different cohorts and time periods:

- **GLOBEM (Xu et al., 2023a):** This is a widely used benchmark for longitudinal human behavior modeling and generalization. It comprises data collected from 497 unique participants across two institutions over four years (Year 1 & 2 from Institution A, Year 3 & 4 from Institution B), resulting in 661 person-years of data after initial preprocessing steps. Institutions A (pre-COVID) and B (post-COVID) represent distinct cohorts and time periods, with surveys including PHQ-4, BDI-II, and PANAS for depression assessment. We utilize a subset of 15 features based on prior work (Xu et al., 2023a; Thach et al., 2025; Kim et al., 2024), derived from mobile sensing data sources, including Location (variance, entropy, travel distance, duration of stay), Phone Usage (unlock counts, stats), Bluetooth (scan counts, unique devices), Call (duration stats, missed call count), Physical Activity (steps, active/sedentary duration), and Sleep (duration, episode stats). For the main evaluation, we use data from Years 1 & 2 from Institution A (344 person-years) for training and data from Years 3 & 4 from Institution B (317 person-years) for cross-cohort and cross-temporal generalization testing. Each person-year of data represents a 10-week observation period from which 6 sequences are generated using a 4-week sliding window predicting the subsequent week. This results in a training set of approximately 2226 sequences and a test set of approximately 2023 sequences. The task is binary mental health prediction based on a threshold applied to survey scores, resulting in a nearly balanced distribution.
- **LifeSnaps (Yfantidou et al., 2022):** This is a multi-dimensional LE dataset initially collected from 71 participants over 4 months, capturing unobtrusive snapshots of real-world human behavior in the wild. Data sources include Fitbit sensing data (e.g., activity, sleep, stress, heart rate), EMAs (e.g., mood, context), and validated surveys (e.g., psychological traits). The dataset includes over 35 distinct data types. For this work, we use a subset of relevant features from these modalities to predict a binary anxiety level in the week subsequent to a $k=1$ week observation window. After initial preprocessing steps, including filtering participants with significant missing values, a subset of participants was used for the evaluation splits. The specific cross-distribution split for evaluation involves training on data from 26 participants collected during the first 2 months of the study period and testing on data from 13 disjoint participants collected during the last 2 months, assessing cross-temporal and cross-participant generalization within the study cohort. Using a $k=1$ week window over these approximately 8-week periods yields a training set of approximately 112 sequences and a test set of approximately 64 sequences. This dataset serves to further validate cross-study generalization within the mental health domain using a different dataset structure, population, and data collection protocol.
- **MFAFY (Hayat et al., 2024a,b; Thach et al., 2025):** The Messages From A Future You (MFAFY) dataset captures aspects of first-year college students' academic journey over three consecutive semesters spanning two academic years (Year 1: Semesters 1 & 2; Year 2: Semester 3). It is a high-dimensional dataset comprising non-cognitive (28 dimensions, qualitative, e.g., motivation, engagement), cognitive (41 dimensions, quantitative, e.g., assessment scores), and background factors (9 dimensions, static qualitative, e.g., aca-

Table 3: Detailed Comparison of VLM Architectures Evaluated in Ablation Study.

Model	Architecture (Language, Vision, Adapter)	Fine-Tuning Process (Relevant Stages for Multimodal Adaptation)	Multimodal Interleaving Support	Relevance to LE-Viz Performance
LLaVA-NeXT 7B	- Language: Qwen1.5-7B-Chat (instruction-tuned LLM). - Vision: SigLIP-L-336px. - Adapter: MLP-based projector.	Two-stage: 1) Initial feature alignment (projector training), 2) Visual instruction tuning (joint projector + LLM tuning). Strong emphasis on multimodal instruction following.	Strong native support, designed to effectively handle sequences of interleaved images and text through its instruction tuning and architecture.	Its instruction-tuned base LLM, capable vision encoder, robust adapter, and joint LLM tuning provide strong multimodal reasoning and narrative generation capabilities. Native interleaving support directly aligns with the LE-Viz input structure, enabling superior integration of features and temporal patterns for generalization.
MiniGPT-4 13B	- Language: Vicuna 13B (instruction-tuned LLM). - Vision: CLIP-ViT-L-336px. - Adapter: Single linear projection layer.	Two-stage: 1) Alignment (projector training), 2) Fine-tuning with detailed image descriptions (projector + some generation layers, base LLM frozen). Focus on high-quality captioning.	Limited native support. Primarily designed for single image + text. Interleaving requires custom handling and is not natively optimized, potentially hindering effective information integration from multiple visual inputs.	Strong base LLM for generation. However, the simpler single-layer adapter and the strategy of freezing the base LLM during multimodal fine-tuning limit its ability to capture the complex temporal and relational structures in interleaved LE visuals, leading to weaker multimodal integration compared to architectures that fine-tune the LLM or use more complex adapters.
mPLUG-Owl2 8.2B	- Language: LLaMA-2-7B (raw, not instruction-tuned). - Vision: CLIP-ViT-L-14. - Adapter: Modular (Visual Knowledge, Abstractor) with two-layer MLPs and LoRA.	Two-stage: 1) Alignment (adapter training, LLM frozen), 2) Joint LoRA tuning (LLM + Abstractor) with multimodal instructions. Modular design.	Good native support via modular architecture and multimodal instruction tuning. Designed to handle diverse multimodal sequences.	Modular design and LoRA tuning offer flexibility in adapting to new tasks. However, starting from a raw LLaMA base (vs. instruction-tuned LLMs), potentially a less performant vision encoder (CLIP-ViT-L-14), and how its modular components interact with structured line chart inputs may limit its overall multimodal reasoning and narrative generation quality for this specific generalization task compared to LLaVA-NeXT.

demic meta-information). For forecasting student behavioral engagement, we predict a student’s lecture-related engagement status (binary: high/low) in the subsequent week, using a $k=4$ week observation window. Input features use only relevant non-cognitive dimensions. The binary target is derived by comparing the average of relevant non-cognitive features during weeks s through $s + k - 1$ with their average during week $s + k$. This task results in a nearly balanced binary distribution. For evaluation, the cross-year generalization split consists of a training set using data from 61 subjects in Year 1 (Semesters 1 & 2) and a test set using data from 35 subjects in Year 2 (Semester 3). Each subject-year/semester of data represents a 15-week observation period from which 10 sequences are generated using a 4-week sliding window predicting the subsequent week. This results in a training set of approximately 610 sequences and a test set of approximately 350 sequences.

For all datasets, train/test splits are carefully cre-

ated to ensure strict separation of data from different cohorts or time periods for generalization evaluation, with 15% of the data from the training period (T) reserved as an in-distribution test set and 100% of the data from the distinct period (T') used as the OOD test set. Data preprocessing, including handling raw sensor data and feature scaling where appropriate, is applied consistently before handling missing data (via mean imputation, as detailed in Section 3.2) and generating the multimodal representations.

Baselines We compare the LE-Viz framework against baselines spanning three categories: time-series models, LLM-based text approaches, and a multimodal time-series model. We also reference the best reported traditional ML/DL result on GLOBEM for historical context.

- **Best Reported Traditional ML/DL Result on GLOBEM:** The best reported cross-distribution accuracy using traditional methods on the GLOBEM benchmark is 0.528 ± 0.024 (Xu et al., 2023a), achieved by the method from (Chikersal et al., 2021), which utilizes ex-

tensive feature engineering (breakpoint, slope) and combines sensor-specific gradient boosting models with Adaboost. The GLOBEM paper’s own proposed multi-task learning method, Reorder (Xu et al., 2023b) (which trains a 1D-CNN to solve a reordering puzzle pretext task alongside classification), reports an average cross-distribution accuracy of 0.525 ± 0.003 . None of the traditional methods evaluated achieve generalization performance significantly above random chance (0.50) for this challenging task.

- **Time-Series Models:** We evaluate two state-of-the-art transformer-based time-series architectures that process numerical sequences without language or vision components:

- *PatchTST* (Nie et al., 2022): A patch-based Transformer that segments time series into subseries-level patches and applies channel-independent processing with self-attention. PatchTST has achieved strong results on standard long-term forecasting benchmarks by capturing local semantic information through patching while reducing computational cost.
- *iTransformer* (Liu et al., 2024c): An inverted Transformer architecture that applies attention across variate dimensions rather than temporal dimensions, treating each feature’s entire time series as a token. This design captures multivariate correlations while maintaining computational efficiency. These models are evaluated under identical in-distribution and out-of-distribution conditions as all other baselines across all three datasets.

- **LLM Baselines:** To compare the impact of our multimodal input representation against text-based LLM approaches, we implement baselines using Llama 3.1 8B Instruct (Grattafiori et al., 2024). The LLM is fine-tuned on textual representations derived from the k-week input sequences, with the goal of generating the same prospective narrative output as LE-Viz. We evaluate four distinct strategies for converting the k-week LE data into a linear textual input:

- *LLM-Complete* (Hayat et al., 2024a,b): The complete, detailed LE sequence is verbalized into a natural language narrative.

- *LLM-Summary* (Thach et al., 2025): A concise narrative summary of the k-week sequence is generated using standard techniques (Zhang et al., 2023), intended to capture key information in a condensed form.
- *LLM-Meta* (Hayat et al., 2025): The textual meta-narrative representation introduced in ConText-LE, which produces an interpretative summary capturing global behavioral context and significant trends across all features. This representation also serves as the textual component of LE-Viz’s multimodal input (Section 3.2).
- *LLM-Clustered* (Kim et al., 2024): A structured representation where each feature is presented sequentially, followed by a list of its temporally ordered values (e.g., Feature A: $[v_1, v_2, \dots, v_k]$). This approach groups values by feature but maintains a 1D sequence input to the LLM.

The generative text output of these LLM baselines is evaluated using the same metrics and extraction procedures as LE-Viz.

- **Multimodal Time-Series Model:** We include Time-VLM (Zhong et al., 2025), a multimodal architecture that integrates numerical time-series inputs with vision-language modeling components. Time-VLM was designed for standard numerical time-series forecasting tasks (e.g., electricity, traffic, weather) and does not specifically accommodate the heterogeneous multimodal inputs (numerical, categorical, and qualitative features) characteristic of LE data. We include it as the most closely related VLM-based time-series approach to assess whether generic multimodal temporal architectures transfer to LE data forecasting. Time-VLM is evaluated under identical conditions as all other baselines across all three datasets.

A.4 LLM Prompt for Textual Meta-Narrative Generation

The following system prompt is used with GPT-4o to generate textual meta-narratives from raw LE data.

System Prompt:

You are an expert behavioral analyst tasked with evaluating a student’s weekly behavioral reflections over a 4-week course. The data includes daily pre- and post-lecture thoughts, confidence levels, peer comparisons, and future-oriented reflections.

Your objective is to analyze the evolution of the student’s behavior and mindset across the 4 weeks. In your response:

- Identify and describe specific behavioral trends, such as shifts in confidence, motivation, or engagement.
- Reference specific weeks (e.g., “In Week 1...”, “By Week 3...”).
- Use precise language to describe changes, such as “X increased by Week 2”, “Y decreased from Week 1 to Week 4”, or “Z remained consistent until Week 3”.
- Avoid vague terms like “overall” or “in general” to ensure analytical precision.
- Provide a concise, natural, and evidence-based analysis in 3–4 sentences.
- Exclude any personal or identifying information from the response.

A.5 LLM Prompt for Textual Output Language Generation

The following prompt is used with GPT-4o to generate prospective narrative targets for training. The model receives the student’s behavioral data and ground-truth label, and produces an expert explanation supporting that label.

System Prompt:

You are an expert behavioral analyst. A student’s weekly behavioral reflections over a 4-week course are provided below, including daily pre- and post-lecture thoughts, confidence levels, peer comparisons, and future-oriented reflections:

{input_text}

The student’s behavior is labeled as ‘*{output_label}*’.

Write a clear, natural-language expert explanation – just a single 3–4 sentence paragraph explaining the behavioral trends that support the label. Be concise and insightful, as if communicating with another expert. Avoid vague terms like “overall” or “in general,” and exclude any personal or identifying information.

A.6 LLM Prompt for Interpretation of the Generated Narrative

The following prompt is used with GPT-4o to extract a binary classification from the generated prospective narrative.

System Prompt:

You are a student engagement expert. Based on the behavioral reasoning below, classify the student’s confidence level as either “High” or “Low”. You must choose one. No explanation.

Reasoning: *{reasoning_text}*

Output only: High or Low.

A.7 Extended Related Work

Our work intersects several areas of research in natural language processing, multimodal learning,

and cross-distribution generalization. This section reviews relevant literature on LLM-based approaches to longitudinal and structured data, vision-language models for non-photographic content, domain generalization, multimodal representation learning, and cognitively-inspired visual representation design.

Longitudinal Experiential Data Modeling.

Modeling longitudinal experiential data is important for applications in behavioral health (Nemati et al., 2022; Rabbi et al., 2019; Bae et al., 2017), mental health (Wang et al., 2018; Xu et al., 2021a, 2019; Chikersal et al., 2021), and education (Wang et al., 2016; Li et al., 2020). Traditional machine learning and deep learning approaches prioritize in-distribution performance and struggle with cross-distribution generalization, as highlighted by the GLOBEM benchmark (Xu et al., 2023b). These methods often neglect multi-dimensional relational structure (Xu et al., 2019) and face challenges with limited training data (Xu et al., 2023a).

LLMs for Temporal and Behavioral Forecasting. Recent work applies LLMs to LE data forecasting through various text serialization strategies. Kim et al. (Kim et al., 2024) investigate LLM prompting and fine-tuning on health datasets including GLOBEM, achieving promising in-distribution performance but focusing on within-dataset evaluation. Hayat et al. (Hayat et al., 2024a,b) apply LLM-based forecasting using diverse architectures on the MFAFY dataset, while Thach et al. (Thach et al., 2025) propose MuH-Boost for zero-shot multi-label prediction using LLMs. More recently, Hayat et al. (Hayat et al., 2025) introduce a contextual meta-narrative representation and formulate forecasting as prospective narrative generation, achieving strong cross-distribution results on GLOBEM, LifeSnaps, and MFAFY through text-only LLM processing. A common limitation of these text-based approaches is that they process data as one-dimensional token sequences, collapsing the inherent two-dimensional *Feature* $\times$ *Time* organization into a linear token stream where temporal adjacency is not preserved as positional proximity (Liu et al., 2024b). Our work builds on the meta-narrative and prospective narrative formulations of (Hayat et al., 2025) and investigates whether adding visual encoding can recover the structural information that text serialization loses.

Vision-Language Models for Non-

Photographic Visual Data. VLMs pre-trained on massive image-text pairs (Radford et al., 2021; Jia et al., 2021; Alayrac et al., 2022) have achieved strong results on natural image understanding tasks. Recent work extends VLMs to structured visual content including documents (Xu et al., 2020, 2021b), scientific figures, and charts (Masry et al., 2022; Islam et al., 2024). LayoutLM (Xu et al., 2020) jointly models text and layout information for document understanding, while ChartQA (Masry et al., 2022) introduces benchmarks for visual and logical reasoning about charts. These applications address the challenge of processing abstract, structured visual data where spatial organization encodes semantic relationships that do not depict physical scenes. Our work extends this direction by applying visual encoding to inherently non-visual numerical temporal data, transforming sequences into spatial representations that VLMs can process for cross-distribution generalization.

Temporal Data Modeling with Vision-Language Approaches. Time-VLM (Zhong et al., 2025) applies VLMs to time series forecasting by visualizing numerical sequences, achieving improvements on standard benchmarks through multimodal integration. While sharing the high-level concept of leveraging visual representations for sequential data, our work differs in several respects. First, we transform heterogeneous data combining numerical, categorical, and qualitative modalities, not purely numerical sequences. Second, we address cross-distribution generalization across cohorts and time periods, not in-domain forecasting. Third, we design visual encodings to preserve two-dimensional relational structure ($Features \times Time$) beyond simple temporal trajectories. Fourth, we formulate forecasting as prospective narrative generation, producing natural language outputs that can be parsed for specific predictions while maintaining interpretability.

Domain Generalization and Cross-Distribution Learning. Domain adaptation (Pan and Yang, 2010) and domain generalization (Zhou et al., 2022) aim to improve model performance on target distributions different from training distributions. Techniques including invariant representation learning, meta-learning, and data augmentation have been applied across various domains. However, success on complex

longitudinal behavioral data remains limited (Xu et al., 2023a). The multifaceted shifts occurring across human cohorts, time periods, and contexts pose challenges beyond typical domain shifts in static image or text classification datasets. Our work addresses this gap by designing visual representations that preserve structural properties invariant to distribution shift, combined with pre-trained visual features that transfer across domains.

Multimodal Integration and Representation Learning. Multimodal learning combines information from different data types (Baltrušaitis et al., 2017), with applications spanning medical imaging (Singhal et al., 2023; Lin et al., 2023; Wang et al., 2022), video understanding, and embodied AI. Effective multimodal integration requires designing representations that preserve complementary information across modalities while supporting cross-modal reasoning. Our approach uses linguistic representations (textual meta-narratives) to provide global semantic context and visual representations (spatially-encoded charts and heatmaps) to preserve local temporal dynamics. This complementary design allows VLMs to leverage both language understanding and spatial pattern processing capabilities for temporal reasoning.

Cognitively-Inspired Visual Representation Design. Our approach draws on cognitive psychology research, particularly Barbara Tversky’s work on visuospatial reasoning (Tversky, 2005, 2011). This research emphasizes that humans use external visualizations (diagrams, maps, charts) as cognitive tools that structure information, offload working memory, highlight relationships, and support complex reasoning by making abstract concepts spatially tangible. This functional view informs our transformation of numerical temporal data into multimodal representations designed to highlight two-dimensional relational structure. The use of complementary linguistic (abstract, global) and visual (structured, local) modalities draws on insights about how humans integrate different representational formats to support comprehension and reasoning under complexity.

A.8 Detailed Ablation Study

This section provides full results and analysis for the three ablation studies summarized in Section 4.2.

(1) **Impact of VLM Choice.** Using the op-

timal input configuration (Text: Meta-Narrative + Visual: Interleaved Charts), we compare three VLMs alongside the text-only LLM-Meta baseline on GLOBEM (Table 4). LLaVA-NeXT 7B achieves the highest generalization (Acc **79.37%** ID, **72.86%** OOD), surpassing both MiniGPT-4 13B and mPLUG-Owl2 8.2B despite their comparable or larger parameter counts. This result is instructive: not all VLMs benefit equally from visual encoding of temporal data. LLaVA-NeXT pairs a SigLIP-based vision encoder operating at 336×336 resolution with an instruction-tuned language backbone, producing stronger cross-modal integration for abstract visualizations. In contrast, MiniGPT-4 uses a frozen language model and mPLUG-Owl2 relies on a CLIP-based vision encoder, and both underperform the text-only LLM-Meta baseline on OOD F1. This suggests that a poorly aligned vision encoder can introduce noise that degrades generalization, and that VLM architectural choices, particularly vision encoder quality and cross-modal alignment, are important variables when applying visual encoding to structured temporal data.

The 7.12 percentage point OOD gap between LLaVA-NeXT (SigLIP-336px, 72.86%) and mPLUG-Owl2 (CLIP, 65.74%) is particularly informative because it persists despite frozen vision encoder weights in both cases. This isolates the contribution of visual representation quality from language model capacity: the performance difference is attributable to the vision encoder’s ability to process abstract temporal visualizations, not to differences in language model size or instruction tuning. Low-level visual primitives learned from natural images (edge detectors sensitive to trend slopes, texture analyzers responsive to pattern discontinuities, color gradient processors that track intensity changes) appear to transfer to temporal pattern detection in charts, but the fidelity of this transfer depends on encoder architecture and pre-training.

Table 4: VLM impact on GLOBEM generalization using Text (Meta) + Visual (Interleaved Chart).

VLM	In-Distribution		Out-of-Distribution	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)
LLaVA-NeXT 7B	79.37	82.03	72.86	74.92
MiniGPT-4 13B	73.09	70.87	66.68	64.00
mPLUG-Owl2 8.2B	70.85	65.97	65.74	63.77
LLM-Meta Baseline	73.99	73.87	67.40	69.40

(2) Impact of Visual Organization. This ablation investigates whether the spatial arrangement

of visual information affects generalization, testing the prediction from our representational design that maintaining per-feature spatial modularity should benefit cross-distribution transfer. Using LLaVA-NeXT 7B on GLOBEM (Table 5), we compare the interleaved format (each feature visualization presented alongside its textual description) against a combined format (all feature visualizations arranged in a single image). Interleaved visuals consistently outperform combined formats by a wide margin. For Chart representations, the interleaved approach achieves **72.86% Acc / 74.92% F1** OOD versus 62.48% / 64.18% for combined, a difference of over 10 percentage points. Similar patterns hold for Heatmaps (70.98% / 73.16% vs. 66.73% / 66.23%).

This result bears on the question of whether LE-Viz’s gains stem from representational structure or from information quantity. Both formats contain identical visual information, but organizing it to preserve per-feature spatial adjacency produces a 10+ percentage point difference in OOD performance. If the gains were attributable to the visual channel simply providing additional data, spatial organization should not matter. The fact that it does supports the claim that *how* the information is spatially structured, not *how much* information is present, drives the generalization improvement. The underlying mechanism can be understood through attention distance: in the interleaved format, visual tokens for each feature appear adjacent to the corresponding textual description, allowing local cross-modal attention to establish feature-specific text-vision correspondences. In the combined format, all visual tokens are grouped together and separated from their textual counterparts, requiring the model to learn long-range correspondences that are less robust under distribution shift.

(3) Controlled Multimodal Component Analysis. To isolate each modality’s contribution, we perform controlled component analysis using the LifeSnaps dataset (Table 6). Three experimental conditions are compared under identical hyperparameters: (i) Text-only (vision encoder disabled), (ii) Vision-only (charts only, no textual input), and (iii) Multimodal (Text + Chart). Across all three VLMs, the multimodal configuration consistently achieves the highest accuracy and F1 in both ID and OOD settings. For LLaVA-NeXT 7B, the multimodal condition (71.9% OOD Acc) outperforms

Table 5: Impact of visual organization on GLOBEM generalization: interleaved vs. single combined visual representations.

Method	In-Distribution (Year 1 & 2 Test)				Out-of-Distribution (Year 3 & 4 Test)			
	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)
LE-Viz (Interleaved Chart)	79.37	83.33	80.77	82.03	72.86	75.58	74.28	74.92
LE-Viz (Interleaved Heatmap)	76.68	72.83	71.28	72.04	70.98	73.19	73.13	73.16
LE-Viz (Single Combined Chart)	67.26	76.27	66.67	71.15	62.48	62.16	66.34	64.18
LE-Viz (Single Combined Heatmap)	72.65	78.26	71.43	74.69	66.73	65.09	67.42	66.23

Table 6: Controlled multimodal component analysis on LifeSnaps comparing text-only, vision-only, and multimodal conditions across three VLMs.

VLM	In-Distribution (ID)			Out-of-Distribution (OOD)		
	Text-Only Acc (%) / F1 (%)	Vision-Only Acc (%) / F1 (%)	Text+Vision Acc (%) / F1 (%)	Text-Only Acc (%) / F1 (%)	Vision-Only Acc (%) / F1 (%)	Text+Vision Acc (%) / F1 (%)
LLaVA-NeXT 7B	65.5 / 68.9	58.4 / 61.3	82.4 / 80.0	67.0 / 72.8	57.3 / 59.9	71.9 / 76.4
MiniGPT-4 13B	63.1 / 62.4	60.8 / 63.6	73.2 / 72.2	61.7 / 61.0	56.8 / 61.0	68.4 / 71.2
mPLUG-Owl2 8.2B	64.5 / 63.8	58.3 / 60.8	70.3 / 69.8	64.3 / 63.6	55.8 / 58.4	69.8 / 71.4

both Text-only (67.0%) and Vision-only (57.3%).

Two observations are important here. First, Text-only outperforms Vision-only across all VLMs, indicating that the textual meta-narrative is the stronger individual modality and that current VLMs are more effective at extracting behavioral predictions from interpreted text summaries than from raw visual temporal patterns. Second, despite this asymmetry, the multimodal combination consistently exceeds text-only performance, indicating that the visual channel contributes complementary information that text alone does not capture. Combined with the visual organization results in Ablation (3), where identical visual content organized differently yields a 10+ point OOD gap, these findings point to complementary representational affordances as the source of the multimodal gain: text provides interpreted semantic context while spatial visual structure preserves local temporal patterns that linearized text discards.

On LifeSnaps OOD, Text-only achieves 67.0%, Vision-only 57.3%, and the multimodal combination yields 71.9%. The multimodal gain over Text-only (4.9 points) exceeds the Vision-only performance relative to chance, suggesting that text and visual representations capture partially non-overlapping aspects of the underlying temporal structure. This pattern of super-additive multimodal combination is consistent across all three VLMs, with gains ranging from 5.5 to 6.7 points over the respective text-only baselines.