MTP: A Dataset for Multi-Modal Turning Points in Casual Conversations

Gia-Bao Dinh Ho[♡], **Chang Wei Tan**[♠], **Zahra Zamanzadeh Darban**[♠], **Mahsa Salehi**[♠], **Gholamreza Haffari**[♠], **Wray Buntine**[♡]

[♡]VinUniversity, Ha Noi, Viet Nam
[♠]Monash University, Melbourne, Australia
{bao.dhg, wray.b}@vinuni.edu.vn
{Chang.Tan, Zahra.Zamanzadeh, mahsa.salehi, gholamreza.haffari}@monash.edu

Abstract

Detecting critical moments, such as emotional outbursts or changes in decisions during conversations, is crucial for understanding shifts in human behavior and their consequences. Our work introduces a novel problem setting focusing on these moments as *turning points (TPs)*, accompanied by a meticulously curated, high-consensus, human-annotated multi-modal dataset. We provide precise timestamps, descriptions, and visual-textual evidence highlighting changes in emotions, behaviors, perspectives, and decisions at these turning points. We also propose a framework, TPMaven, utilizing state-of-the-art vision-language models to construct a narrative from the videos and large language models to classify and detect turning points in our multi-modal dataset. Evaluation results show that TPMaven achieves an F1-score of 0.88 in classification and 0.61 in detection, with additional explanations aligning with human expectations.

1 Introduction

Identifying key moments in videos, like highlight detection or moment retrieval, is crucial. This involves pinpointing moments through scene changes or specific descriptions using matching and strategic comparison processes. Turning point (TP) classification and detection enhance this by incorporating reasoning to identify significant conversational shifts. The challenge lies in the complex reasoning needed, evident in our data annotation where even human annotators require group discussions. Detecting these turning points is vital for post-analysis of conversations, recognizing moments that impact speakers' reactions. Understanding these moments enhances future interactions, particularly valuable in new or unfamiliar settings like therapy or negotiation, and offers strategies for successful outcomes.

Given limitations in existing multi-modal datasets and the novelty of our research, we aim to pioneer the creation of a novel high-quality dataset with turning points. Collecting four seasons of The Big Bang Theory TV series, with its eccentric characters likely causing turning points, we focus on 40 episodes from seasons 1 to 4, specifically on conversations.

This study makes several contributions: (1) Introducing Multi-modal Turning Point Classification (MTPC), Multi-modal Turning Point Detection (MTPD), and Multi-modal Turning Point Reasoning (MTPR) tasks in human casual conversation. (2) Curated a human-annotated **M**ultimodal **T**urning **P**oints (MTP) dataset for casual conversation, enriched with textual and visual cues depicting subjective personal states. (3) Proposing a novel framework for MTPC and MTPD, utilizing vision language models (VLMs) for narrative construction and large language models (LLMs) for effective reasoning in turning point detection. (4) The code and data are publicly available.[1]

2 Related work

Multi-modal datasets have been developed for understanding human conversations (Reece et al., 2023; Meng et al., 2020; Wang et al., 2023; Firdaus et al., 2020; Lei et al., 2018; Li et al., 2023; Shen et al., 2020). Each of them having limitations such as missing visual data, or providing just extracted features from it, missing context on shorter sequences, alignment issues and so forth. To address these gaps, we developed a multi-modal conversational dataset from TV series episodes, featuring video content with timestamp annotations, aligned transcripts, and video frames, with annotations for turning points.

Turning points are a special case of change points (Aminikhanghahi and Cook, 2017) sometimes indicating a trend change direction or substantial change in intent for human data. TPs in

[1] https://giaabaoo.github.io/TPD_website/

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 314–326

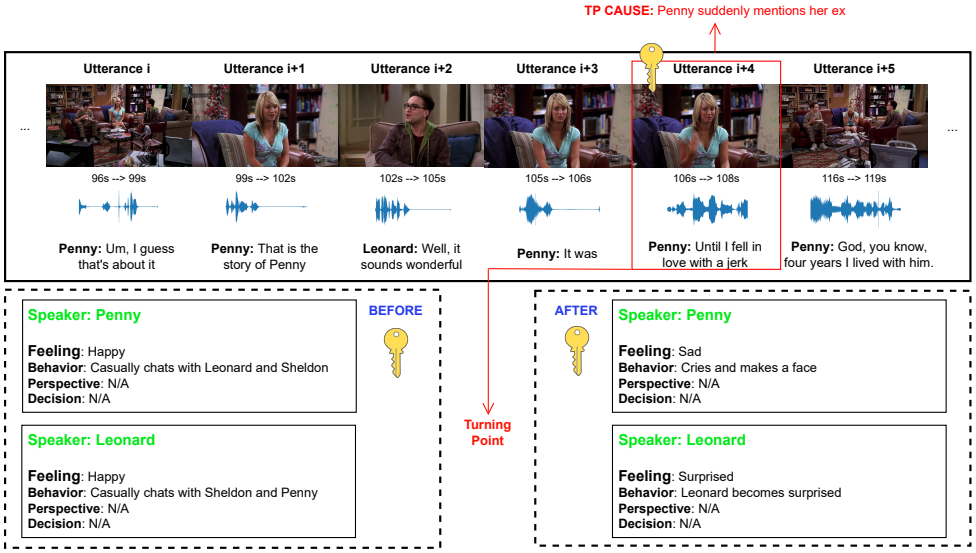

Figure 1: Considering this example: Everyone is chatting casually. A turning point occurs when Penny (female character) starts crying, caused by her mentioning her ex while sharing her personal stories with Leonard and Sheldon (two male characters). According to human commonsense, this should be considered a significant change in the conversation because it catches the attention of the people watching, and the speakers involved (Leonard and Sheldon become confused).

narrative analysis, as described by (Keller, 2020; Papalampidi et al., 2019, 2021), denote critical moments that shape the plot and segment narratives into thematic units. In psychology and social sciences, TPs are moments of significant change in individuals' perceptions, feelings, or life circumstances (Florida Association for Women Lawyers, 2003; Wieslander and Löfgren, 2023). Our research adopts the TP definition from (Keller, 2020) and (Papalampidi et al., 2019), focusing on crucial moments within conversations that significantly impact discourse elements in human-simulated dialogues from a TV series. Kumar et al. (2022) introduces Emotion-Flip Reasoning (EFR), which is the task of identifying past utterances in a conversation that triggered a speaker's emotional state to change, aiming to explain emotional shifts during dialogue. For clarification regarding the differences, we not only provide information on emotional changes but also on the causes behind those changes. We specifically focus on significant emotional shifts. Moreover, we consider changes in decisions, perspectives, and behaviors as they are deemed significant. Additionally, we provide visual-textual evidence for these changes.

3 Problem formulation

The context of a casual conversation is denoted as C, comprising m utterance-level videos $U = \{u_1, \ldots, u_m\}$. Each utterance video u_i is associated with a corresponding text transcript and a speaker name $\{t_i, s_i\}$. We consider turning points within the conversation, in accordance with Definition 1.

Definition 1 *A **turning point** in this context is a moment that belongs to an utterance in a conversation, triggered by an identifiable event (that is called the turning point cause). This moment marks the beginning of unexpected or significant changes in the subjective personal states of at least one participant (such as decisions, behaviors, perspectives, and feelings)* [2]*. We have annotated it with a timestamp and a textual explanation of its cause (Further elaboration on the definition is in appendix B.1).*

Our proposed problem inputs consist of utterance-level videos with corresponding transcriptions, speaker names, and timestamps bound to the transcript. The problem can be divided into three tasks. The first task, referred to as MTPC (*Multi-modal Turning Point Classification*), involves determining if a conversation includes any turning points (TP). The second task, MTPD (*Multi-modal Turning Point Detection*), focuses on pinpointing the timestamps of these turning points

[2]We identified these states through a process of group discussions, video analysis, and literature review in Section 2, focusing on the most common variables in the post-analysis of casual conversations.

in the conversations. A correct turning point is identified when the predicted timestamp falls within a time window threshold δ_t relative to the ground truth. The third task, MTPR (*Multi-modal Turning Point Reasoning*), aims to discern the reasons behind each turning point, presented as a textual description. This task is crucial for formulating potential solutions to address negative turning points and gaining insights into cultural norms. Regarding evaluation, the model's timestamp predictions can be assessed qualitatively. However, we believe that the textual causes should be evaluated by human experts. Currently, we have not identified a qualitative method for evaluating textual causes, considering it as a potential avenue for future research.

Total number of conversation videos	340
Total duration (h)	13.3
Total number of utterance-level videos	12351
Total number of words in all transcripts	81909
Average length of conversation transcripts	241.5
Maximum length of conversation transcripts	460
Average length of conversation videos (s)	1.9
Maximum length of conversation videos (m)	2.5
Total number of TPs videos	214

Table 1: Statistics of the MTP Dataset

4 The MTP Dataset

"The Big Bang Theory" (Lorre and Prady, 2007) provides a rich source of casual conversations, forming the foundation of our study. The eccentricities of its characters create a unique backdrop for sensitive moments crucial to our turning points analysis. Our three-stage process involves human annotators determining scene start and end times (Subsection 4.1), extracting videos for conversations. The second phase (Subsection 4.2) annotates turning points based on guidelines explained in appendix B, while the third stage annotates relevant information, such as visual-textual evidence for observed changes.

4.1 Scene boundary annotation

Since an episode can contain multiple scenes, but our focus is solely on studying conversations within each scene, we conducted scene boundary annotation. In the first phase, we initiated scene boundary annotation by providing videos (crawled from the internet), scene's tags, and their initial sentences extracted from Mirshafiee (2021) to annotators. They were tasked with accurately identifying the start and end times of scenes by watching the videos and using the first sentences as cues as explained in annotation details in appendix A.2.1. The statistics of the dataset can be found in Table 1.

4.2 Creating utterance-level videos

WhisperX (Bain et al., 2023) was employed to segment conversation C into utterance-level videos ($U = \{u_1, \ldots, u_m\}$) with precise timestamps ($\delta T = \{\delta t_1, \ldots, \delta t_m\}$) and transcripts ($T = \{t_1, \ldots, t_m\}$). We found that the speaker identifier is crucial for human annotators to locate the turning points. To address this, we utilized an online dataset (Bain et al., 2023) containing speaker identifiers for Big Bang Theory episodes. Using GPT embedding search and the LLAMA model for prompting, we matched each utterance transcript t_i to the corresponding speaker ID. Finally, human refinement was employed to ensure accurate alignment. This process resulted in triplets $\{t_i, \delta t_i, s_i\}$ for each utterance u_i in conversation C, with s_i representing the speaker for utterance i (further details are provided in appendix A.1).

4.3 Multi-modal Turning Point Annotation

We assembled a team of three annotators, all of whom are proficient English-speaking students. Each conversation was then assigned to two annotators for annotation with clear guidelines (appendix B). The third annotator was designated as a judge responsible for reviewing the annotations and engaging in discussions with the first two annotators.

4.4 Turning Point Evidence Annotation

Once annotators identify turning points, they provide pre- and post-change details for a nuanced understanding. Clear explanations are required when annotators perceive no turning point, enhancing comprehension of situations considered unremarkable. Additionally, annotators timestamp moments of change in feelings, behaviors, decisions, and perspectives, substantiating observations with visual or verbal evidence.

4.5 Feelings Annotation

Annotators are asked to focus on emotions closely tied to turning points, ensuring clarity in decisions, behaviors, or perspectives before and after these moments. The incorporation of a feelings recognizer is motivated by recognizing emotions as vital markers in conversations. By highlighting feelings associated with turning points, annotators reveal

Methods	Turning point classification				Turning point detection		
	Precision	Recall	F1	AUC	Precision	Recall	F1
GPT-3.5	0.7	0.84	0.76	0.47	0.44	0.6	0.45
GPT-4	**0.81**	**0.96**	**0.88**	0.52	0.43	0.75	0.51
GPT-4 w/o tracking prompt	0.69	0.95	0.8	0.47	0.31	0.69	0.43
GPT-4 + few shot	0.71	0.95	0.82	**0.53**	**0.52**	**0.87**	**0.61**

Table 2: Performance metrics for turning point classification and detection using different comparison methods

emotional undercurrents shaping responses. We believe that proficient emotion recognition in the valence-arousal space aids in discerning significant changes in feelings, crucial for identifying turning points. However, due to resource constraints, we use common classes from the circumplex model of emotion (Russell, 1980) (see appendix A.2.3 for the model) instead of annotating valence and arousal for each emotion, enhancing precision and providing a structured framework for annotators to navigate human emotions systematically. An annotator selects frequent emotions from the circumplex model, defining a list including Positive (Happy, Excited, Calm, Relaxed, Alert), Negative (Anxious, Angry, Disgusted, Sad, Upset, Depressed, Frustrated, Confused), and Neutral/Transitional (Surprised, Neutral, Serious, Nervous) emotions.

4.6 Annotation consensus

After annotators completed their tasks, a group discussion session was organized to review and discuss conversation labels. The aim was to decide whether to keep, add, or delete turning points. This resulted in 340 conversations, with 214 having turning points and 126 without. Agreement was reached when annotators and the judge agreed on turning point labels, occurring in approximately 82% of the dataset's turning point events. If all three annotators identify three distinct turning points (though this scenario didn't happen), the sample would be deleted due to the lack of unanimous agreement. Typically, we retain annotations receiving at least two out of three votes for a turning point. In our review session, when annotators identified the same turning points but provided different yet reasonable evidence, we merged their before and after evidence (including emotions and behaviors changes).

5 TPMaven framework

We present TPMaven, a language model prompting framework engineered to identify and ground turning points in casual conversational videos. The framework comprises two key components: 1) a scene describer that captures the visual information and articulates the essence of each utterance; and 2) a robust reasoner that interprets instructions, locating and elucidating turning points. For the first component, we prompt the LLAVA model (Liu et al., 2023) as our scene describer to get the relevant visual description of the scenes (frames) in the conversations. For the second, various ChatGPT models are prompted with a system prompt, including the definition of TP and three prompts for turning point identification: a describing instruction, the conversation $C = \{< t_1, v_1, s_1 > , \ldots, < t_m, v_m, s_m >\}$, with v being the visual description, an optional tracking prompt to direct ChatGPT to track individual in the conversation, and a command prompt. Further details on the prompting templates for both components can be found in appendix C.

6 Experiments

We use LLAVA-7B (Touvron et al., 2023) to extract visual information in scene descriptions. GPT-3.5-1106 (a version of GPT-3.5 (OpenAI, 2022)) and GPT-4-1106 identify turning points, addressing context length issues. For assessing turning point localization, we focus on the positive set with 214 conversations. True positives are determined when predicted timestamps fall within δ_t = 20 seconds of ground-truth timestamps. During segmentation, we map GPT model outputs (utterance indices) back to timestamps for comparison (see more details in appendix D). The performance metrics, including Precision, Recall, F1 and Area Under the Curve (AUC) are reported for each method in Table 2. GPT-4, especially with few-shot learning, stands out as the most promising method for turning point classification, surpassing GPT-3.5 and GPT-4 without tracking prompts. We also found that the grounding output of GPT-4 is much concise in terms of tracking compared to other GPT models.

7 Conclusion

In conclusion, our research addresses the crucial task of recognizing pivotal moments in conversa-

tions, presenting a detailed taxonomy and a curated dataset called MTP. Our baseline framework, TP-Maven, utilizes vision-language and GPT models for classification and detection, demonstrating its performance across various metrics. While TP-Maven provides explainable predictions for sensitive moments, experimental results highlight the need to discern conversations with and without turning points. Future directions are in appendix E.

Limitations

The dataset is designed for post-analysis to understand what captures the attention of viewers in videos and speakers during conversations. Due to resource limitations, we could only curate a single-lingual dataset focused on critical moments in English culture. Unfortunately, we had to opt for simple emotion annotation instead of the more informative valence-arousal space annotation, which would provide intensity and direction of emotions.

Furthermore, we faced challenges in evaluating the Multi-modal turning point reasoning task. While attempting to utilize another GPT-4 as an evaluator for explanations on some samples, followed by human verification, we encountered inconsistent results. Despite our belief that human evaluation is optimal, resource constraints prevented us from pursuing this approach. Emotion reasoning was excluded for the same reason.

Regarding scene-describing methods, we have employed LLAVA due to its cost-effectiveness. Although a faster version of GPT-4 was available (OpenAI, 2023) during the submission of this work, which could potentially improve scene descriptions, budget limitations hindered us from exploring its use.

In this problem, the input should simply be a video, and the output should consist of the turning points. However, at the time of conducting this research, we have not identified any reliable speaker identification method; therefore, this aspect may be addressed in our future research. As speaker IDs are crucial for tracking the states of each individual in the conversation, and it is reasonable to assume that speakers are known through the normal mental human annotation process, we believe it is justifiable to human-annotate that information instead of relying on an inaccurate speaker ID. The latter could lead to expected underperformance. It is important to note that turning points should also encompass non-verbal cues. Currently, we only consider verbal turning points that occur within an utterance. The case of online turning point detection, where turning points are identified in real-time, has not been explored in our research at this time. Additionally, we believe that the definition of a turning point can be broadened to encompass specific conversational contexts beyond casual discourse, such as political discussions. In these situations, even slight changes in subjective states can lead to significant norm violations. Conversely, in our scenario of casual conversations among friends, a much higher threshold should be considered to distinguish between meaningful event changes and insignificant ones.

Ethics consideration

Data life-cycle and access: Our dataset has been scrutinized and approved by the relevant institutional committees. All annotators have agreed to relevant terms and participated in training sessions. They were compensated at a rate significantly higher than the local minimum wage. The resources presented in this work are utilized for research purposes only. We have obtained all data copyrights pertinent to this paper. To ensure proper citation and prevent malicious application, we have prepared detailed instructions, licenses, and a data usage agreement document that we link in our project repository. Additionally, we intend to make our software available as open source for public auditing.

Copyrights Our dataset incorporates videos from 'The Big Bang Theory' television series for training AI models in natural language understanding tasks. The inclusion of copyrighted material raises important considerations regarding fair use and transformative use under copyright law. We assert that our use of these videos qualifies as fair use, as it is conducted for transformative purposes aimed at advancing scientific understanding and innovation. Specifically, our research involves the transformation of the original videos through linguistic analysis and modeling, contributing novel insights into conversational comprehension. Furthermore, our use of the videos is limited in scope and does not detract from the commercial market for the series. We provide appropriate attribution to the copyright owner of the show and take measures to ensure that the dataset is used responsibly and ethically within the research community.

Data bias: When pinpointing a crucial turning

point, the evidence reflecting subjective personal states (feelings, behaviors, perspectives, decisions) may exhibit variations. Annotators, expressing diverse viewpoints on the same event in human language, can contribute to this divergence. Consequently, the explanations and evidence surrounding the turning point may incorporate personal bias in articulating the matter. We advise future users of the dataset to be mindful of this potential bias.

Acknowledgements

This research is based upon work supported by U.S. DARPA under agreement No. HR001122C0029. The opinions, views, and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes, notwithstanding any copyright annotation therein. We appreciate all annotators for their contributions to this work. We would also like to thank Prof. Heng Ji for her valuable feedback.

References

Samaneh Aminikhanghahi and Diane J Cook. 2017. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Florida Association for Women Lawyers. 2003. Turning points. *F.A.W.L. Journal*, Summer. A Publication of the Florida Association for Women Lawyers.

Frank Keller. 2020. Analysing and summarizing movies via turning point identification in screenplays. Talk. The International Multimodal Communication.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. *Knowledge-Based Systems*, 240:108112.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Chen Li, Xutan Peng, Teng Wang, Yixiao Ge, Mengyang Liu, Xuyuan Xu, Yexin Wang, and Ying Shan. 2023. PTVD: A large-scale plot-oriented multimodal dataset based on television dramas. *ArXiv*, abs/2306.14644.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Chuck Lorre and Bill Prady. 2007. The big bang theory. *CBS*.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. OpenViDial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Mitra Mirshafiee. 2021. The Big Bang Theory series transcript. `https://www.kaggle.com/datasets/mitramir5/the-big-bang-theory-series-transcript`. Dataset on Kaggle.

OpenAI. 2022. Introducing chatgpt. Accessed: 2022.

OpenAI. 2023. Gpt-4v(ision) system card. OpenAI Technical Report.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717, Hong Kong, China. Association for Computational Linguistics.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2021. Movie summarization via sparse graph construction. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 15, pages 13631–13639.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*

on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. MemoR: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 493–502.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuxuan Wang, Zilong Zheng, Xueliang Zhao, Jinpeng Li, Yueqian Wang, and Dongyan Zhao. 2023. VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5036–5048, Toronto, Canada. Association for Computational Linguistics.

Malin Wieslander and Håkan Löfgren. 2023. Turning points as a tool in narrative research. *Narrative Inquiry*. Page 3.

A MTP Dataset creation details

A.1 Preprocessing

In analyzing conversation C, we utilized WhisperX (Bain et al., 2023) to segment each video into m utterance-level videos ($U = \{u_1, \ldots, u_m\}$) with precise start and end timestamps ($\delta T = \delta t_1, \ldots, \delta t_m$) for each transcript ($T = \{t_1, \ldots, t_m\}$).

Speaker IDs for each utterance were annotated by a process of matching with the transcripts and speaker labels from the scenes in Mirshafiee (2021). For each utterance extracted by WhisperX, we need to find the row in Mirshafiee (2021) to extract the speaker name. This can be done by matching the corresponding transcript from WhisperX and the row from Mirshafiee (2021). Using GPT-3.5, we created an embedding file for each scene extracted from Mirshafiee (2021), where each line represents a text pair of utterance and corresponding speaker (u', s). Through an embedding search for each WhisperX-extracted utterance u_i, we retrieved the most similar sentence u'_i from the preprocessed Mirshafiee (2021) with its corresponding speaker s_i. We prompted LLAMA-7b with transcript t_i and the candidate sentence, including speaker names from the search model, to assign the speaker for each utterance. Recognizing potential unintended outputs from LLMs, human annotators meticulously verified speaker identification, ensuring accurate alignment with respective names in the transcripts.

A.2 Annotation

A.2.1 Scene Boundary

It is crucial to emphasize that our episodes consist of various scenes and transitions, requiring the annotation of scene boundaries. To streamline this task, we enlisted a team of students to view the videos. They were tasked with assigning scene tags and providing the initial sentence for each scene, serving as a prompt to expedite the process. This meticulous process resulted in the identification of 340 conversations, comprising a comprehensive 13.3 hours of video content for our study.

A.2.2 Turning Points

An example of our turning point annotation can be found in Table 3.

scene	A corridor at a sperm bank.
duration	150
conversation	1
TP_location	01:25
TP_cause	Sheldon shows his concerns about donating sperm
pre_point_feeling	neutral (1:24)
post_point_feeling	nervous (1:38)
pre_point_dbp	Leonard and Sheldon plan to donate sperms so that they can have extra money (1:45)
post_point_dbp	Leonard and Sheldon leave the room (2:29)
explanation	According to commonsense, there is a clear change in their decisions.

Table 3: A sample turning point annotation for conversation 1 in our dataset. **pre_point_dbp** and **post_point_dbp** stands for pre-point and post-point decisions, behaviors, perspectives respectively.

A.2.3 Feelings

Annotators are asked to focus on emotions closely tied to the turning points, ensuring clarity in decisions, behaviors, or perspectives before and after

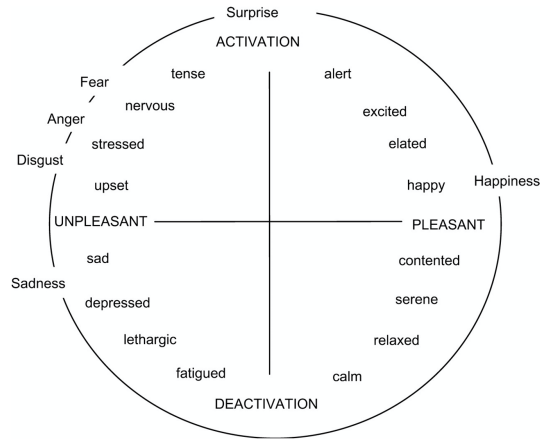

Figure 2: The circumplex model of emotions in (Russell, 1980)

these turning points. The intuition behind incorporating a feelings recognizer lies in the recognition that emotions serve as vital markers of key moments in a conversation. By focusing on feelings closely associated with turning points, annotators can illuminate the emotional undercurrents that shape individuals' responses and reactions. For instance, someone may say something offensive, but whether it forms a turning point depends on the other person's reactions. We also believe that a proficient emotion recognizer within the valence-arousal space proves valuable in discerning significant changes in feelings. Without knowing the intensity and direction of these changes, identifying turning points becomes challenging. To avoid overcomplicating the annotation process due to resource constraints, we opt for common classes in the circumplex model of emotion depicted in Figure 2 instead of annotating valence and arousal for each emotion. The circumplex model of emotion enhances this process by providing a structured dimension. This model maps emotions based on underlying dimensions such as valence and arousal, ensuring systematic classification. It not only enhances labeling precision but also offers annotators a practical framework to navigate the intricate landscape of human emotions.

A.3 Statistics

A.3.1 Different types of turning points

After annotating the data, we provide ChatGPT with all the causes of turning points and categorize the types in Table 4.

Types	Explanation
Emotional Outbursts	Sometimes, when someone gets really, really mad and can't control it, it can lead to a big, angry fight.
Changes in Decisions	Sometimes, the group has a plan, but suddenly they decide to do something different.
External Influences	Imagine someone new joins the conversation, and it completely changes how everyone feels or what they think.
Shifts in Perspective	Sometimes, everyone starts thinking one way, but later on, they change their minds and think differently.
Uncomfortable Situations	Imagine someone violating social norms, and it makes everyone feel uncomfortable or upset.
No Turning Points	- Even when someone says something mean, everyone reacts like they normally would, without any big changes. - Sometimes, during the conversation, nobody's subjective personal states change much; things stay pretty much the same.

Table 4: Different categories of turning points (TP) types were identified by prompting and providing ChatGPT with a list of TP causes from our dataset.

A.3.2 Emotional shifts

We also provide the analysis of the most common types of emotional changes before and after turning points in Figure 3.

B Turning points annotation guidelines

B.1 Further elaboration on the definition

Considering definition 1, we want to elaborate some important terms.

B.1.1 The term "*identifiable*"

This means the **event** can be recognized based on clear evidence.

Considering a conversation from Table 5, the identifiable events are:

1. Penny discovers Leonard and Sheldon entering Penny's apartment and confronts them about it.

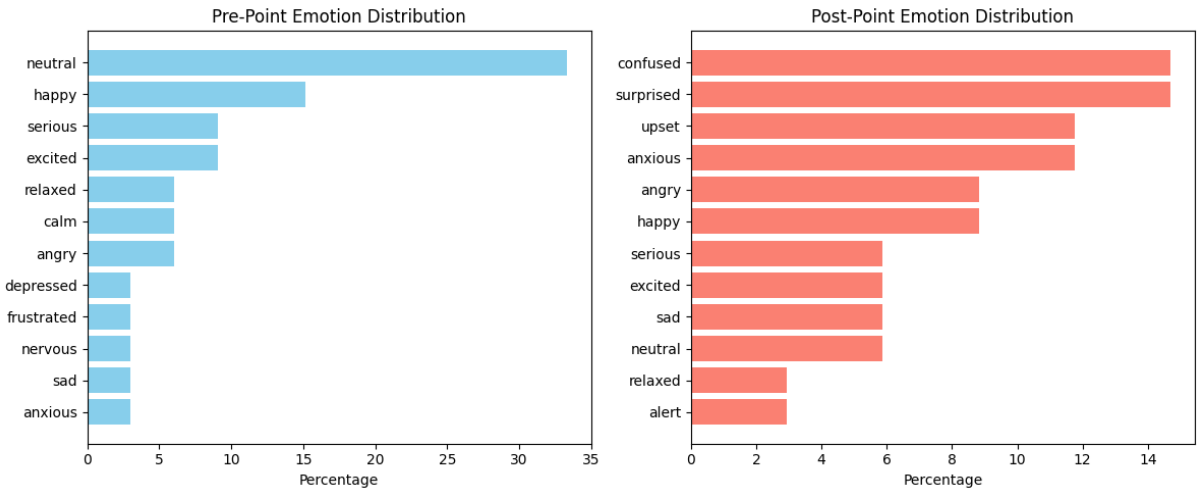

Figure 3: Emotional distribution of the top 20 most occurrences before and after the turning point in our dataset. This caption summarizes the analysis of emotions in relation to the most frequent occurrences, highlighting changes around the identified turning point in the dataset.

Leonard: Penny's up. Penny: You sick, geeky bastards! Leonard: How did she know it was us? Sheldon: I may have left a suggested organizational schematic for her bedroom closet. Penny: Leonard! Leonard: God, this is going to be bad. Sheldon: Goodbye, Honey Puffs, hello Big Bran. Penny: You came into my apartment last night when I was sleeping? Leonard: Yes, but, only to clean.

Table 5: A sample transcript of a conversation in our dataset

2. Leonard and Sheldon try to explain their actions and justify themselves.

B.1.2 The term "*subjective personal states*"

These encompass changes in a speaker's:

- **Decisions:** Choices made during the conversation.
- **Behaviors:** Actions taken during the conversation.
- **Perspectives:** Shifts in the way a speaker sees or understands a topic.
- **Feelings:** Emotional states.

B.1.3 The term "*Unexpected*"

The event should be surprising and deviate from the usual flow or expectations of the conversation.

B.1.4 The term "*Significant*"

The change should be of significance, impacting not only the individual but also affecting the dynamics of the conversation.

- It affects not only one person but also those around them.
  - Example: When Person A cries, it makes Person B cry too.
- The impact on the subjective personal states can differ, but it should make common sense.
  - Example: Changing your mind from staying in to going out is considered significant.
  - Example: Changes in how you act, like going from being neutral to getting into a debate or becoming more engaged, are considered significant.
  - Example: Going from feeling normal to feeling heartbroken is considered significant.

B.1.5 The term "*During*"

The annotators are asked to consider the evidence before and after that point in the current conversation only, not the potential consequences.

B.1.6 The goal of detecting TPs

In healthcare monitoring, we have two scenarios. For critical patients, we use a low sensitivity threshold to detect even subtle changes due to their sensitivity. For general patients, we employ a high

sensitivity threshold to identify only the most significant changes, avoiding unnecessary alerts.

Similar to general patient monitoring, our research objective is to identify important moments in casual conversations. We focus on recognizing changes that match our definition of significance while ignoring minor ones. This knowledge base serves as a valuable resource for developing applications, encompassing conversation analysis to mitigate miscommunication, study decision-making, and behaviors, and highlight key aspects of conversations.

B.2 Annotation Flows

The annotators are given a video of a conversation and asked to follow three phases of annotation.

B.2.1 First phase

In this initial phase, understand the content and flow of the conversation. Identify the topics, speakers, and main events without focusing on turning points.

B.2.2 Second phase

The annotators are asked to find an event in the conversation that causes a turning point, and then label the timestamps where the change occurs. There can be multiple turning points.

Recommended Steps:

1. Evaluate each speaker separately.

2. Analyze changes in decisions, behaviors, perspectives, and feelings independently.

3. If a change meets the criteria of being **significant** and **unexpected**, mark the timestamp when the change starts. Also, write down a short summary of the event that started the change (the cause of the turning point).

   The change in the subjective personal states of a person can be caused by that person or another person, you should write down the event that caused the turning point (**who does what**). If it is caused by a person himself (by rethinking, etc.), you should write down something like "Penny realizes that ..." or "Sheldon decides to ..."

4. Please note the changes both before and after the turning point. While changes in decisions, behaviors, and perspectives are typically evident, when it comes to feelings, concentrate only on those that are closely linked to the turning point. The person whose subjective personal states change will have a clear pre-point and post-point decision or behavior or perspective. You should write who does what too. Additionally, if there is a change in feelings but no corresponding change in decisions, behaviors, or perspectives, please provide a clear explanation of why that change is significant. Since human emotions can change frequently, our focus should be on reasonably significant emotional changes within that context.

5. Mark the timestamp for the evidence associated with those changes in parentheses. The evidence can consist of verbal or non-verbal cues. For example, 'sad (1:05)' indicates that the evidence is located at 1 minute and 5 seconds into the video. At 1:05, a person might say something like, "I broke up with my girlfriend," which provides strong evidence of the feeling of sadness. Alternatively, at 1:05, there is a frame capturing his sadness expressed through his facial expressions.

Key Guidelines

- Decisions, behaviors, and perspectives are more likely to trigger a turning point, as it is defined to capture decisive moments in a conversation.

- When it comes to feelings, it's important to consider the context of why and how they change. This helps us conclude whether there's a significant shift influencing the emotional dynamics of the conversation.

- Ensure turning points are clear and memorable, leaving a lasting impression.

- If no significant moment is found in the first two phases, move on to the next conversation.

- Envision yourself as an impartial observer to identify surprising or attention-grabbing moments.

- Focus on sudden reactions indicating a noteworthy change in the casual conversation dynamics.

- Approach each video with fresh eyes, treating characters as unfamiliar individuals.

B.2.3 Third phase

If a point is labeled as a turning point and you believe it is not adequately represented by the pre-point, post-point, and TP_cause columns, please comment on the additional evidence you think is necessary for a conclusive determination.

If you are uncertain whether it qualifies as a turning point, provide a clear explanation, and express any concerns you may have.

C TPMaven framework

We present TPMaven, a language model prompting framework engineered to identify and ground turning points in casual conversational videos. The framework comprises two key components: 1) a scene describer that captures and articulates the essence of each utterance, providing a comprehensive understanding of the visual information; and 2) a robust reasoner that interprets instructions, skillfully locating and elucidating turning points, offering insightful explanations for shifts in the conversation.

C.1 Scene describer

Originally, our intention was to utilize the video-language understanding model Video-LLAMA. However, due to prolonged processing times, we opted for an expedited alternative, extracting a list of frames denoted as $F = \{f_1, \ldots, f_m\}$, wherein each frame corresponds to an individual utterance.

To expedite the process, we opted for LLAVA, a vision-language model that demonstrated satisfactory results in human evaluations and improved processing efficiency compared to Video-LLAMA. While GPT-4 integrated with images was considered, it was dismissed due to cost constraints. Subsequently, each utterance in the video is now denoted by a paired set $\{t, f\}$, where t signifies the transcript, and f represents a randomly selected frame during that utterance. Given that TV series consistently feature the speaker's face in every utterance, selecting a random frame serves as a sufficient baseline for capturing visual information. This approach is also computationally efficient.

The examination of visual stimuli within conversations yields rich evidentiary material, encompassing facial expressions and behavioral cues. These visual indicators are instrumental in constructing a comprehensive narrative of the discourse. Hence, we use this prompt: *"Give me the short descriptions of the actions, facial expressions, postures, gestures, potential emotions (with valence and arousal)"* to retrieve the relevant information (including actions and affective factors) that can help us to detect the turning points.

Given the verbosity of LLAVA's outputs and its potential impact on the context length of the GPT model, we employ a GPT-3.5 model for summarization. Eventually, we get a set of visual description for each utterance in the conversational

C.2 Reasoner

Pretrained language models (PLMs) store implicit knowledge about the world learnt from large-scale text collected around the internet (Petroni et al., 2019). There has also been previous attempts to use LLMs as a reasoner for a variety of tasks (Kojima et al., 2022). Our hypothesis is that if we are efficient at telling the story of the conversation to the LLMs and inspired from the CoT methods, if we can prompt a series of relevant prompt that can lead and guide the LLMs towards answering basic questions that it is trained on and is having in its internal knowledge, it can produce desireable results. Thus, we strive to break our tasks down.

From the above steps, each conversation C consists of m utterances can now be represented as $C = \{< t_1, v_1, s_1 >, \ldots < t_m, v_m, s_m >\}$ with t_i, v_i and s_i being the transcript, visual description and speaker for an utterance i respectively. Our prompting template concatenates multiple sub components prompts, each with its own functionality in guiding the LLM:

- **describing_instruction** - *"Read this conversation. Each utterance includes the transcripts and visual descriptions."* - This is followed by filling the conversation in the form of a set of utterances U.
- **tracking_instruction** - *"Utilize a tracker for each person in the conversation. For each speaker, provide a concise list of their feelings, behaviors (based on the context and actions), decisions, and any perspective changes (include those with clear evidence from the conversation). Limit the list to a maximum of 256 words."*
- **commanding_instruction** - *"Identify the turning point events based on the initial conversation and track results if there are any. Begin by finding the turning point for each person."*

We also leverage the system role in the ChatGPT Completion API, which is the role that helps provide fixed high-level instructions to the whole system, by filling in the **system_content** field with this description: *"You are a trained chatbot that can find turning points in conversations. A turning point in a conversation is an identifiable event that leads to an unexpected and significant transformation in the subjective personal states (including decisions, behaviors, perspectives, and feelings) of at least one speaker during the given conversation."* - This prompt is used to fill in the **system_content** of the ChatGPT completion API.

C.3 Conclusion module

We provide GPT-4 with this prompt: "*For each found turning point in the prediction, find the starting utterance index only. Return a list of n utterance start indices corresponding to a turning point in the prediction. Follow strictly this format in your response: e.g. utterances = [utterance_5, utterance_25]. Return None if there is no turning point found. Limit the response to 50 words.*" and the conversation with utterance indices to retrieve the utterance indices that has turning points. Subsequently, we match these indices back to timestamps extracted in the pre-processing stage to compare with the timestamps' label.

D Experimental settings

D.1 Implementation details

For the scene describer, we utilize LLAVA-7B to extract visual information from an image. In the reasoning process, we leverage GPT-3.5-1106 and GPT-4-1106 versions to identify turning points. This choice is motivated by the large input size, mitigating potential context length issues encountered in conventional GPT turbo models from OpenAI. For the classification task, our primary evaluation metrics include Precision, Recall, and F1. Given the dataset's imbalance, we also incorporate the use of AUC. In the detection task, we focus on metrics such as P, R, and F1. To assess the performance of localizing turning points, we exclusively consider the positive set, comprising 214 conversations for evaluation. For each conversation, k turning points are detected by TPMaven. A true positive is determined if, for each ground-truth in the conversation, there exists a predicted timestamp falling within $\delta_t = 20$ seconds. This is done as the turning point event found by ChatGPT can belong to several consecutive sentences. Since the GPT model's output from the conclusion module consists of a list of utterance indices, we map it back to the timestamp from the utterance-level segmentation phase for comparison.

D.2 Discussion of the tracking prompts

Given the conversation video between Sheldon and Leonard in the first scene of the series (Season 01, Episode 01) (Lorre and Prady, 2007) (Please refer to our project website to watch the video[3]), different GPTs are utilized with the tracking prompt. The results are depicted in Figure 4, 5 and 6.

Tracking Results for Sheldon:

- **Feelings:** Neutral, concern, curiosity, seriousness.
- **Behaviors:** Standing, observing surroundings, holding a cell phone, talking on the phone, gesturing with hand, checking their phone, browsing the internet, engaged in a phone conversation, checking or looking at something on the screen of the cell phone.
- **Decisions:** None.
- **Perspective Changes:** None.

Figure 4: Tracking results using GPT-3.5

Sheldon:

- **Feelings:** Neutral, concern, curiosity, interest.
- **Behaviors:** Standing, observing, holding a cell phone, gesturing, browsing the internet, engaging in conversation.
- **Decisions:** None evident.
- **Perspective Changes:** Expresses the concept of quantum physics in a casual conversation, demonstrates interest in a t-shirt idea, and feels the need to explore the idea further.

Figure 5: Tracking results using GPT-3.5-turbo

Sheldon's Tracker:

- **Feelings:** Initially neutral but progresses to concern, curiosity, and eventually doubt regarding the plan.
- **Behaviors:** Engages in scientific discussion, pivots conversation to t-shirt idea, shows interest in the puzzle, reveals hesitance about donating sperm due to the possibility of subpar offspring.
- **Decisions:** Initially endorses the plan for improved internet, but ultimately decides against donating sperm due to ethical concerns.
- **Perspective Changes:** Rethinks the morality of sperm donation and the implications of potentially not having high IQ offspring.

Figure 6: Tracking results using GPT-4

E Discussing future works

In the course of conducting this research, we have identified several critical challenges that we believe are essential to address in future research on Multimodal turning point detection. The following areas present promising avenues for further exploration:

[3] https://giaabaoo.github.io/TPD_website/

Multi-lingual Multi-cultural Dataset

Addressing the nuances in conversations across different languages and cultures, where norms vary, requires the development of a comprehensive multilingual, multi-cultural dataset. Such a dataset would capture the intricacies inherent in linguistic and cultural differences.

Emotion Recognition in Valence-Arousal Space

The development of an effective emotion recognizer in the valence-arousal space holds the potential to enhance traditional time-series change point detection methods. Accurately identifying emotional shifts can contribute to the identification of candidate turning points.

Multi-modal Emotion Reasoning

Our dataset not only captures turning points but also annotates changes in emotions related to these points. Therefore, there is an opportunity to develop methods in emotion reasoning using this dataset.

Multi-modal Turning Point Reasoning

Providing the cause of the turning point and a causal chain of events related to feelings, behaviors, decisions, perspectives, etc., enables the development of a method or benchmark for turning point reasoning. However, a significant challenge lies in constructing a reliable evaluator to compare textual predictions from a model with the ground-truth explanations of turning points.