A. EIBench

A.1. Practical Applications

EIBench's emphasis on *Emotion Interpretation (EI)* supports a variety of real-world use cases:

1. **Enhanced Emotion Recognition:** Most datasets label emotions but ignore *why* they occur. EIBench illuminates causal factors, further refining both accuracy and empathy in emotion recognition. Possible applications: customer service bots, mental health diagnostics, and interactive media, where *causal* triggers foster more context-aware responses.
2. **Adaptive Human-Computer Interaction (HCI):** Capturing *why* users feel certain emotions, EIBench-trained models provide adaptive, personalized experiences. Virtual assistants, interactive gaming, or user-facing platforms can tailor responses to precise emotional contexts.
3. **Psychological and Behavioral Studies:** Researchers can use EIBench's triggers to uncover patterns in emotional responses and factors shaping them. These insights inform clinical psychology interventions and broaden our grasp of human behavior.
4. **Deeper Social Media Analysis:** EIBench extends sentiment analysis by unveiling the emotional context behind online posts. This expanded layer of interpretation aids brands and organizations in tracking public sentiment more accurately, responding to feedback effectively, and managing their online presence with greater nuance.

A.2. Intended Audiences

EIBench aims to advance EI by capturing the subjective nature of emotional states. Addressing the dataset's challenges can lead to *empathetic* AI systems, enriching emotion-driven applications and enhancing human–computer interactions. Additionally, these insights may benefit tasks like humor understanding, harmful stance detection, and other domains that hinge on implicit emotion cues. Overall, EIBench paves the way for multifaceted, context-driven emotion interpretation, pushing the boundaries of next-generation EI research.

B. Baseline Models

B.1. Open-Source Models

Qwen-VL-Chat. Qwen-VL-Chat [3] is a multimodal large language model (LLM)-based assistant developed by Alibaba Cloud. It manages multiple image inputs, multi-round question answering, and uses bounding boxes for grounding. Through a 448×448-resolution visual encoder, Qwen-VL-Chat supports finer text recognition, document QA, and bounding box annotation. Additionally, it operates in English, Chinese, and other languages, enabling end-to-end recognition of bilingual text. Multi-image interleaved conversations allow image-to-image comparisons, enabling scenario analysis and multi-image storytelling.

Video-LLaVA. Video-LLaVA [37] acts as a baseline for Large Vision-Language Models (LVLMs) that handle both images and videos within a unified visual feature space. By aligning image and video representations, Video-LLaVA allows models to enhance performance across both modalities simultaneously, often outperforming methods restricted to either static images or video alone.

MiniGPT-v2. MiniGPT-v2 [9] is a versatile multimodal model supporting diverse vision-language tasks such as image description, VQA, and grounding. It reduces visual token sequence length by merging adjacent tokens, thus enhancing training efficiency at high resolutions. Trained in three stages—broad pretraining, task-specific fine-tuning on high-quality datasets, and multimodal instruction tuning—MiniGPT-v2 excels at chatbot-style interactions and complex multimodal tasks.

Otter. Otter [32] leverages *OpenFlamingo* [2] to perform multi-modal in-context instruction tuning. Each data instance in its *MIMIC-IT* [31] training set comprises an instruction-image-answer triplet along with relevant in-context examples. By conditioning the language model on image-caption or instruction-response pairs, Otter attains strong instruction-following skills and effectively learns from contextual exemplars.

LLaVA-1.5. LLaVA-1.5 [40] builds on CLIP-ViT-L-336px [48] with an additional MLP projection layer and integrates academic-task-focused VQA data. Compared to the original LLaVA, this version enhances cross-modal connections via an MLP connector and utilizes a broader set of VQA data. The 13B checkpoint for LLaVA-1.5 relies on around 1.2M publicly available data samples.

LLaVA-NEXT. Relative to LLaVA-1.5, LLaVA-NEXT [39] improves reasoning, optical character recognition (OCR), and world knowledge under high-resolution settings, reducing model hallucinations and capturing intricate image details. Training includes High-quality User Instruct Data and Multimodal Document/Chart Data, plus the flexibility to employ various LLM backbones (e.g., Mistral-7B [25] or Nous-Hermes-2-Yi-34B[1]).

B.2. Close-Source Models

Qwen-vl-plus. Qwen-vl-plus expands on Qwen-VL's capabilities for detailed recognition, text detection, and high-resolution image handling (e.g., millions of pixels, arbitrary aspect ratios). It performs competitively on a broad spectrum of visual tasks but is available only via an online API.

Claude-3. Claude-3 from Anthropic underscores safety, controllability, and ethics—distinguishing it from ChatGPT via adversarial training that reduces bias and harmful out-

[1] https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B

puts. Although ChatGPT also addresses safety, Claude emphasizes robust security measures and transparent documentation. While ChatGPT excels at broad NLP tasks, Claude's stringent ethical guidelines may favor use cases requiring higher compliance standards.

ChatGPT-4. ChatGPT-4 (ChatGPT-4o, ChatGPT-4V) is OpenAI's state-of-the-art LLM, proficient in text generation, conversation, translation, summarization, and more. It incorporates extensive pretraining to boost coherence and fluency. Like Claude, ChatGPT-4 has significant safety mechanisms for mitigating bias and harm, plus user-feedback loops to enhance performance. Its adaptability makes it effective for a wide array of applications, balancing general NLP strength with ethical safeguards.

B.3. Basic EIBench

EIBench is composed of two primary subsets—*Basic* and *Complex*. The *Basic* subset contains 1615 samples, each aligned with one of four primary emotion categories (*angry*, *sad*, *happy*, *excited*). Unlike the *Complex* subset, which may feature overlapping or multilayered emotions, the *Basic* portion focuses on a single dominant emotion per instance. This design choice allows models to learn and generalize from relatively direct emotional triggers before grappling with more intricate scenarios.

Annotation Approach. We follow the same *Coarse-to-Fine Self-Ask (CFSA)* pipeline as outlined in the main text. However, unlike *Complex* scenarios—where multiple viewpoints or confounding cues might need iterative clarification—the *Basic* subset typically converges on a single, primary trigger. Consequently, annotators can identify and refine emotional cues (e.g., facial expressions, objects, or contextual details) in fewer self-ask rounds, thus ensuring the reliability of each final annotation.

Scope and Limitations. Although each *Basic* sample focuses on one principal emotion, subtler undertones (e.g., mild frustration coexisting with sadness) can still arise. Annotators are instructed to emphasize the dominant emotion, but residual emotional nuances may remain. Models trained on the *Basic* subset alone often handle straightforward triggers well (e.g., "waiting in a queue," "a celebratory event"), yet may perform less effectively when encountering real-world complexities or mixed emotional contexts—challenges that are central to *Complex* EIBench.

Intended Use. The *Basic* subset is especially suited for initial baseline training, providing a gentle introduction for models to learn one dominant emotional cue per instance. Researchers can compare baseline performances on simpler triggers with the more layered triggers in the *Complex* subset. Additionally, the straightforward, readily identifiable causes in the *Basic* portion benefit educational demonstrations, helping novices grasp core mechanisms of emotion interpretation before tackling more advanced material. Overall, *Basic EIBench* offers a structured entry point to explain *why* a single emotion dominates a scene, complementing EIBench's broader aim of preparing models for more nuanced, overlapping emotional states.

B.4. Complex EI Subset

In contrast to the *Basic* subset, the *Complex* EI subset comprises 50 samples featuring overlapping or multilayered emotions (e.g., joy mixed with regret, anger intertwined with concern). Such scenarios push models to identify multiple coexisting triggers and navigate nuanced social or cultural cues (Figure 1(e)).

Scope and Design. Each *Complex* instance often involves layered triggers (e.g., work-related stress combined with family conflict), requiring multi-step reasoning; interwoven perspectives (e.g., two individuals each experiencing distinct emotional reactions), which force the model to untangle different motivations; and implicit contextual depth (e.g., cultural practices or off-screen backstories) that may not appear explicitly but remain crucial for understanding the emotional state.

Annotation Method. Compared to *Basic* cases, annotators adopted a more iterative *Coarse-to-Fine Self-Ask* flow to clarify overlapping cues and verify multiple triggers. This extra step ensures the final annotations encompass all relevant factors (e.g., social tension plus personal grief), rather than focusing on just the first visible cause.

Impact and Utility. The *Complex* subset highlights realistic emotional intricacies, fostering development of more robust *Emotion Interpretation (EI)* models. Beyond academic interest, these examples aid use cases in mental health diagnostics and advanced HCI, where single-label assumptions fail to capture genuine emotional complexity. Together with the *Basic* subset, these intricate scenarios enable a broader transition from straightforward emotion labeling to richer, more nuanced emotional understanding.

C. Human-in-the-Loop Data Cleaning

C.1. Addressing Hallucinations in VLLMs

Vision Large Language Models (VLLMs) can sometimes produce *hallucinated* triggers unrelated to the actual image content. Table 14 shows examples in which the model invents triggers (e.g., "Doing mountain biking") with no supporting evidence. Such hallucinations undermine dataset quality by misrepresenting the visual context. To mitigate these errors, we implement a human-in-the-loop cleaning process: annotators review the VLLM's outputs, remove triggers not clearly supported by the image, and note ambiguous regions for further inspection. By systematically weeding out these misinterpretations, we reduce biases introduced by VLLM-driven hallucinations.

C.2. Incorporating Commonsense Knowledge

Even when models avoid overt hallucinations, they may overlook *commonsense* cues essential to explaining an emotional state. Table 15 illustrates how human annotators augment triggers with contextual or cultural knowledge absent from raw VLLM outputs. For instance, the model may label an emotion as "angry" but omit a crucial real-life cause (e.g., "waiting for lost luggage"), prompting annotators to add relevant details. By explicitly integrating commonsense reasoning, the final dataset more closely aligns with real-world emotional triggers, thus enhancing the fidelity and utility of EIBench for emotion interpretation tasks.

D. Case Study of the VLLMs' EI Abilities

In this section, we present a detailed examination of how various Vision-Language Models (VLLMs) handle *Emotion Interpretation (EI)*, focusing on both *hallucinations* and *commonsense knowledge integration*. Tables 14 and 15 illustrate how a human-in-the-loop data cleaning process identifies and corrects inaccuracies or omissions in VLLM outputs.

Hallucinations in VLLMs. Table 14 shows instances where the VLLM-generated triggers deviate from the image content (e.g., "*Doing mountain biking*" when no bike is present), misrepresenting the scene and undermining dataset quality. By having human annotators remove or adjust these erroneous details, we mitigate biases that might otherwise skew emotion interpretation.

Commonsense Knowledge Integration. Table 15 highlights cases where VLLMs lack crucial background context (e.g., "*first Halloween experience*," "*first time to Beijing*"). Human annotators augment these triggers with necessary cultural or situational information, yielding more realistic and representative data annotations.

Basic vs. Complex EI. Figures 4 and 5 and the accompanying tables illustrate how emotional triggers distribute across *Basic* and *Complex* subsets. In simpler, single-emotion scenarios (Table 10), VLLMs often identify straightforward triggers (e.g., "*long wait*," "*enjoying the view*"). Meanwhile, *Complex* samples (Table 12) feature overlapping triggers or multiple emotional states, frequently exposing model challenges in capturing less obvious cues.

Detailed Model Responses. Tables 14–15 present user queries and ground-truth triggers, alongside raw VLLM outputs (e.g., Qwen-VL-Chat, LLaVA family, MiniGPT, Otter, and ChatGPT-4). Each response is evaluated by LLaMA-3 and ChatGPT for alignment with the annotated triggers. A common pattern emerges: Certain triggers (*e.g.*, metal claws, intense gaze) are detected reliably, while subtler elements (*e.g.*, wide-opening eyes, "defending gesture," "shrunk muscle") are overlooked or inconsistently recognized. Some VLLMs also invent erroneous triggers (e.g., "*concern about a meal he's preparing*") incongruent with the annotated details.

Insights and Implications. These case studies highlight the complexity of moving from mere emotion *recognition* to *interpretation*. Straightforward triggers are typically recognized, but nuanced emotions often hinge on contextual, cultural, or implicit cues. Human review and data cleaning (Sections C.1–C.2) remain vital for honing outputs, particularly in ambiguous or subtle contexts. EIBench thus provides a structured environment for testing not only *Basic* scenarios but also the *Complex* interactions that more closely mirror real-world emotional landscapes.

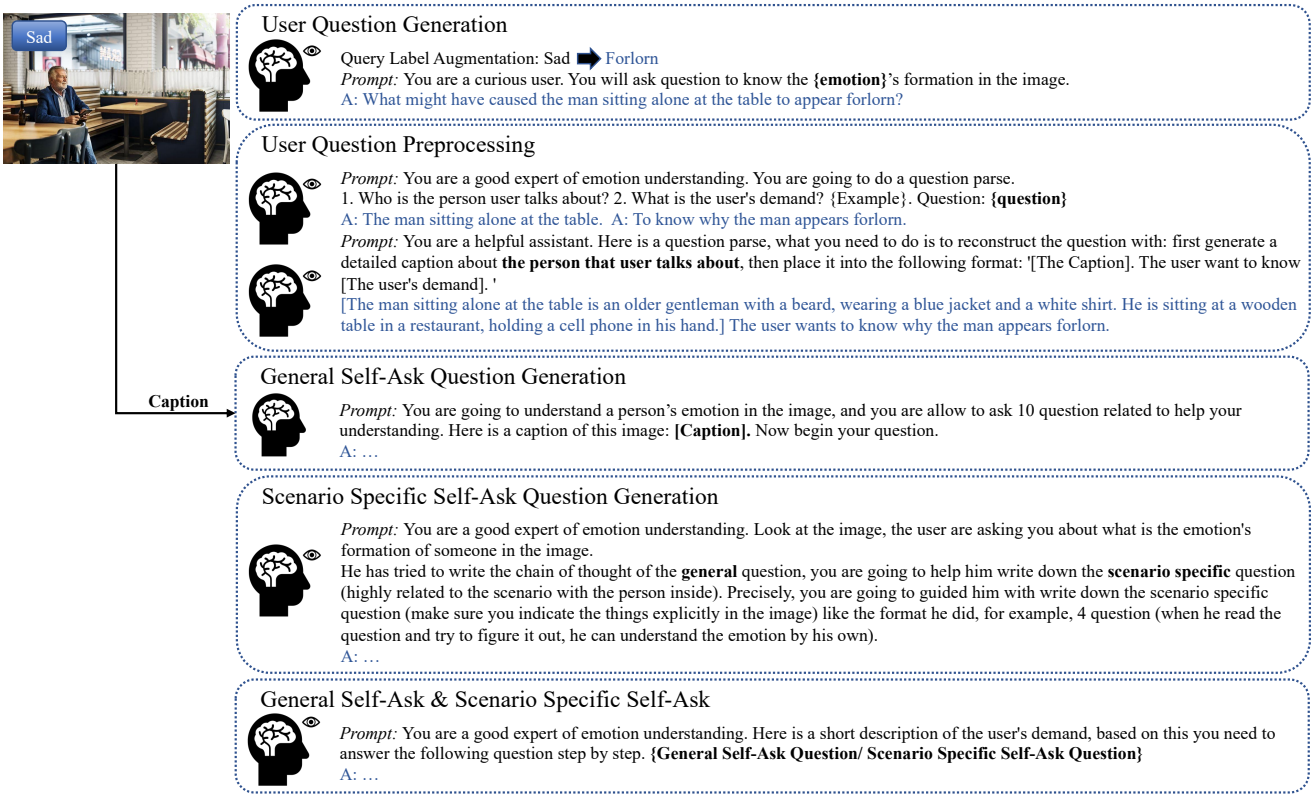


Figure 3. Pipeline of the VLLM-assisted dataset construction.

Table 10. Visualization of basic EI dataset, an image is corresponded to one user questions.

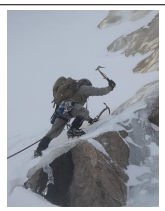
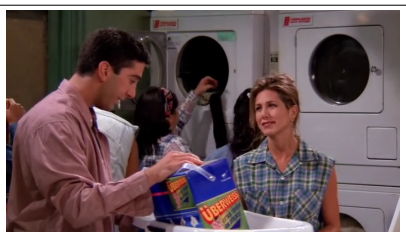
Examples of the Basic EI Dataset	
	
User Question Emotional Trigger	*What led to the formation of the arouse to the man in this image?* 1. Climbing a steep, snow-covered slope. 2. Physical effort and concentration. 3. Potential hazards and challenges. 4. Cold environment. 5. Determination to reach the goal.
	
User Question Emotional Trigger	*What do you think might have caused the person's delight as they look out the window?* 1. Snowy scene outside the car. 2. Smile on her face. 3. Enjoying the view. 4. Serenity of the winter environment. 5. Excitement of experiencing a snowy day. 6. Personal or emotional connections to snowy weather or winter scenes. 7. Fresh snowfall, brightness of the snow reflecting sunlight, or peacefulness of the scene.
	
User Question Emotional Trigger	*What do you think might have caused the man holding the box in the image to become lighthearted?* 1. Holding the "Uberweiss" box. 2. Smiling. 3. Friendly and approachable body language. 4. Positive and relaxed atmosphere of the laundry room. 5. Interaction with others in the laundry room.
	
User Question Emotional Trigger	*What might have caused the woman in the image to appear content and happy?* 1. Positive news about her health. 2. Pleasant interaction with a medical professional. 3. Comforting conversation with a friend or family member. 4. Good news about her health. 5. Positive relationship with the medical staff.
	
User Question Emotional Trigger	*What might have caused the woman in the image to appear irritated or angry?* 1. Service issue (mistake in order, long wait, problem with payment process). 2. Unpleasant environment (noise levels, cleanliness, presence of other customers). 3. Dissatisfaction with food or service. 4. Frustration or annoyance with the conversation or situation.

Table 11. Statistics of the Emotional Trigger Types (Basic Emotions).

Atmosphere	Social Interactions	Body Movements	Facial Expressions	Objects	Performances	Outdoor Activities	Clothing	Sports	Other
23.11%	17.17%	13.24%	9.40%	6.07%	5.06%	3.20%	3.08%	2.25%	17.41%

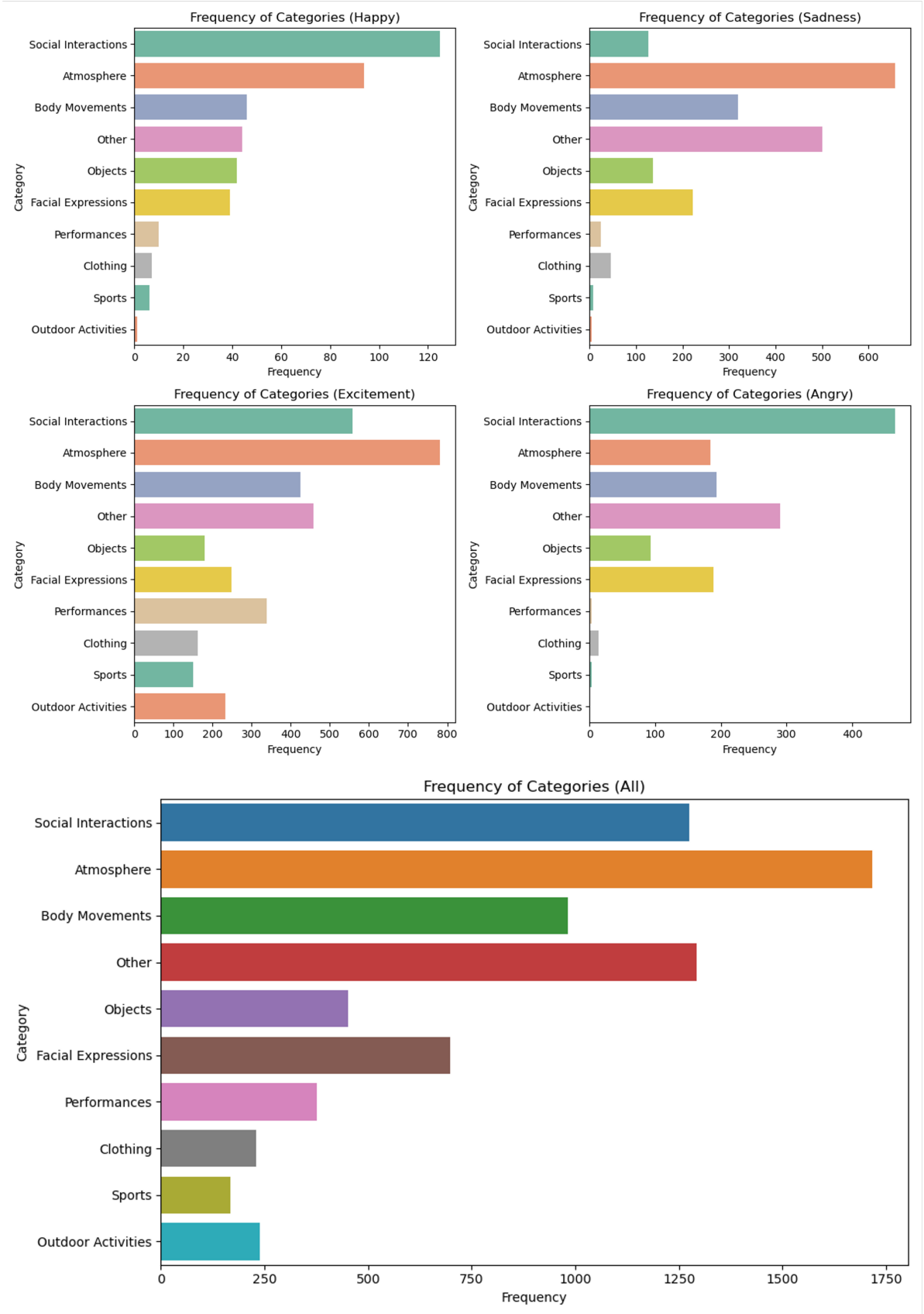


Figure 4. Visualization of the numbers of emotional triggers across different categories (Basic Emotions).

Table 12. Visualization of complex EI subset, an image is corresponded to multiple user questions.

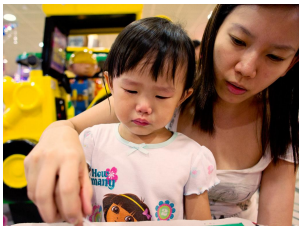
Examples of the Complex EI Subset	
	
User Question (1)	*Why does the kid in the background seem excited?*
Emotional Trigger	1. Head turning back. 2. Starring at the two playing with each other on the focus. 3. Sense of motion from the event. 4. Maybe excited about the desire to join them.
User Question (2)	*What do you think might have caused the kid in the background of the image to be confused?*
Emotional Trigger	1. Head turning back. 2. Two others acting abnormally. 3. Two others each holding a stick of corn. 4. Maybe curious about the event. 5. Maybe wondering about the motivation for the abnormality.
	
User Question (1)	*What may caused the little girl upset?*
Emotional Trigger	1. Crying. 2. Can not making handiwork. 3. The woman blamed her.
User Question (2)	*What may caused the little girl happy?*
Emotional Trigger	1. Crying but the women comfort her. 2. Can not making handiwork. 3. Woman help her finishing the work.
User Question (3)	*What may cause the woman angry?*
Emotional Trigger	1. The girl is not obedient. 2. The girl can't do handiwork. 3. The girl can't learn no matter how much taught. 4. Step-by-step instruction.
	
User Question (1)	*Why does the baby show the fear expression?*
Emotional Trigger	1. The man's scary outfit. 2. Afraid of the man. 3. The man's makeup. 4. Covering mouth with hand.
User Question (2)	*What make the baby surprise and happy?*
Emotional Trigger	1. Shocking face and gesture. 2. Staring at someone. 3. Sense of unbelievable. 4. A man colored in silver on the focus. 5. Maybe shocked to see something abnormal.
	
User Question (1)	*Why does this man in the picture look exhausted and annoyed?*
Emotional Trigger	1. Maybe lack of Sleep. 2. Closed-eyes. 3. Taking care of a young child. 4. Tired of the child. 5. Naughty child.
User Question (2)	*Why does this man being enjoyment and pleasure?*
Emotional Trigger	1. Enjoying spending time with his child. 2. Child lying in arms. 3. Satisfied with the moment. 4. Sense of company of family. 5. Engaging in playful activities.

Table 13. Statistics of the Emotional Trigger Types (Complex Emotions).

Atmosphere	Social Interactions	Body Movements	Facial Expressions	Objects	Performances	Outdoor Activities	Clothing	Sports	Other
10.81%	23.00%	19.37%	16.22%	8.55%	0.45%	3.60%	3.60%	0.9%	13.51%

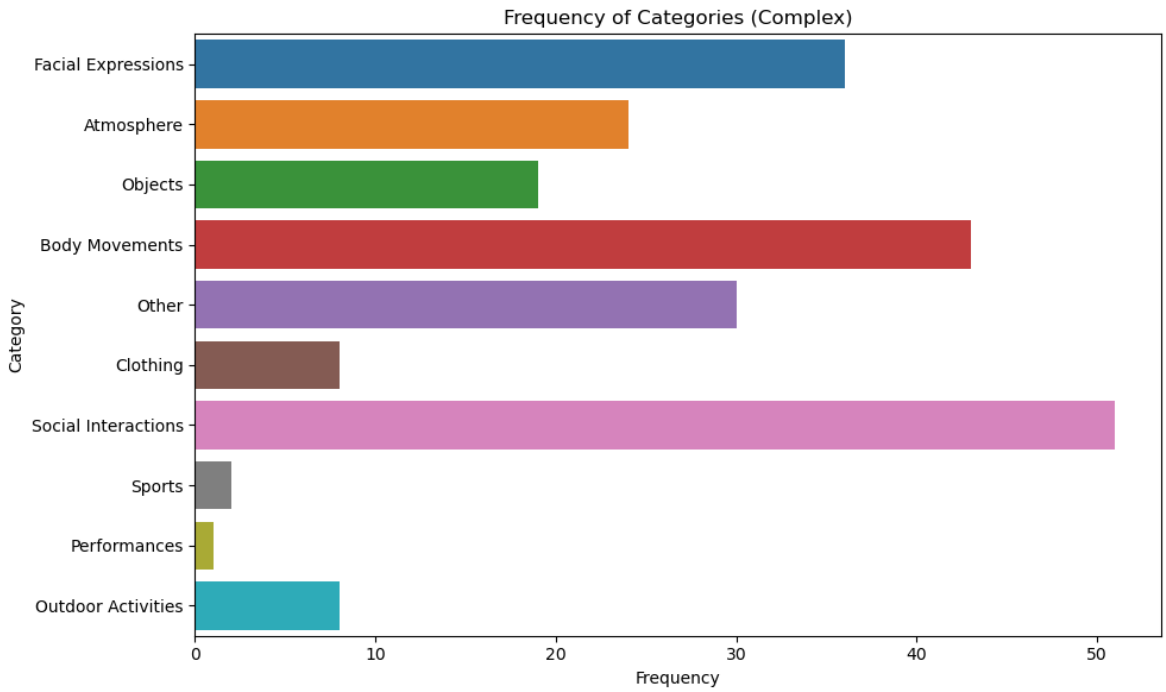


Figure 5. Visualization of the numbers of emotional triggers in the Complex EI Subset.

Table 14. Example of Hallucinations in VLLMs. Hallucinations are indicated in red, while other text is indicated in gray.

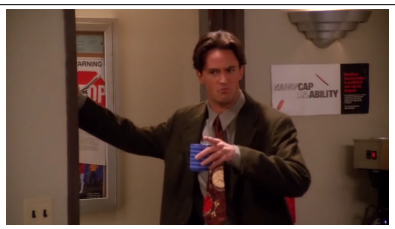
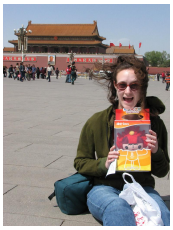
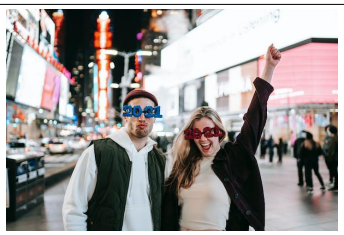
Examples of the Human Cleaning Process of Hallucinations	
	
User Question	*What might have motivated the man in the image to participate in this outdoor activity, given his gear and the environment?*
Emotional Trigger (Raw)	1. Determination and concentration. 2. Challenge of the race or trail. 3. Personal goals. 4. Desire to improve mountain biking skills. 5. Well-prepared gear. 6. Environmental factors (rocky slope, weather conditions). 7. Doing mountain biking.
	
User Question	*What could have caused the man in the image to appear outraged or hostile?*
Emotional Trigger (Raw)	1. Holding a black bag. 2. Animated conversation or gesture. 3. Furrowed eyebrows. 4. Open mouth. 5. Wide or squinting eyes. 6. Leaning forward or gesturing with hands. 7. Brown couch (as a place where he typically relaxes or discusses matters)
	
User Question	*What might have caused the man in the image to be angry or upset?*
Emotional Trigger (Raw)	1. KANO CAP ABILITY sign on the wall. 2. Feeling overwhelmed or pressured by his workload. 3. Undervalued or overworked in his professional role 4. Recent events or interactions in the workplace that have caused stress or frustration. 5. Tension and stress in his body language (posture, grip on the mug).
	
User Question	*What might have caused the man in the image to appear angry or frustrated?*
Emotional Trigger (Raw)	1. Disagreement with a family member. 2. Concern about a meal he is preparing. 3. Problem at work that he is thinking about while in the kitchen. 4. Serious or intense mood due to work-related issue or concern.

Table 15. The Human in the Loop process instills Commonsense Knowledge into the dataset. Text orange represents added commonsense knowledge.

Examples of Data Cleaning for Commonsense Knowledge	
	
User Question	*What might have caused the baby's delight in this image?*
Emotional Trigger	1. Halloween costume and bib with a pumpkin design. 2. Interaction with the person holding them up. 3. Festive atmosphere and attention from the person holding them up. 4. First Halloween experience.
	
User Question	*What led to the excitement on the woman's face?*
Emotional Trigger	1. A toy written "Beijing Welcome". 2. Taking a photo with Tienanmen Square. 3. First time to Beijing.
	
User Question	*What might have caused the man in the image to become excited and make a funny face?*
Emotional Trigger	1. Celebratory event or milestone related to the year 2021. 2. Excitement and joy. 3. Playful or lighthearted moment shared between the man and the woman. 4. Achievement or personal milestone. 5. Festive and celebratory atmosphere.
	
User Question	*Why does the kid in the background seem excited?*
Emotional Trigger	1. Head turning back. 2. Starring at the two playing with each other on the focus. 3. Sense of motion from the event. 4. Maybe excited about the desire to join them.