I. Genre Definition

The whole experiment has maintained the definition of the genres as follows,

1. Biography: Films that depict the life story of a real person, often focusing on their major achievements and personal challenges.

2. Musical: Movies in which songs sung by the characters are interwoven into the narrative, often to express emotions and advance the plot.

3. Mystery: Films centered around the solving of a puzzle, crime, or unexplained event, often involving a detective or investigator.

4. Adventure: Movies are characterized by exciting and often dangerous journeys or quests, typically involving exploration or battles.

5. Action: Films that emphasize physical feats, including fights, chases, explosions, and other high-energy scenes.

6. War: Movies that focus on armed conflict, battles, and the personal and political effects of warfare.

7. Music: Films that revolve around the creation, performance, or significance of music, often featuring significant musical performances.

8. Fantasy: Movies set in imaginary worlds with magical elements, mythical creatures, and extraordinary adventures.

9. Romance: Films that focus on the romantic relationships between characters, often highlighting their emotional journey.

10. Thriller: Movies are designed to create suspense, excitement, and tension, often involving crime, espionage, or psychological elements.

11. Family: Films suitable for all ages, often focusing on family relationships, values, and adventures that appeal to both children and adults.

12. History: Movies that portray historical events, figures, and periods, aiming to recreate and dramatize past occurrences.

13. Crime: Films that revolve around criminal activities, the perpetrators, and the law enforcement officials who pursue them.

14. Drama: Movies that focus on realistic characters, emotional themes, and intense character development and interactions.

15. Comedy: Films intended to entertain and amuse, often through humor, satire, and lighthearted stories.

II. Genre Group

For the experiment, we grouped the 15 genres of the utilized dataset according to their semantic similarities for better analysis.

1. Drama and Emotional Focus

(i) Biography: Focuses on real-life stories and personal achievements. *(ii) Drama*: Emphasizes realistic characters and emotional themes. *(iii) Romance*: Centers around romantic relationships and emotional journeys.

2. Excitement and Suspense

(i) Action: High-energy scenes with physical feats and battles. *(ii) Thriller:* Creates suspense and excitement, often involving crime or psychological elements. *(iii) Adventure:* Exciting journeys and quests, often involving exploration and battles. *(iv) Crime:* Revolves around criminal activities and law enforcement pursuits.

3. Historical and Real-World Events

(i) History: Portrays historical events and figures. *(ii) War:* Focuses on armed conflict and the effects of warfare. *(iii) Biography:* Real-life stories and historical figures.

4. Fantasy and Imagination

(i) Fantasy: Imaginary worlds with magical elements and mythical creatures. *(ii) Adventure:* Can overlap with fantasy in terms of quests and exploration.

5. Music and Performance

(i) Musical: Songs integrated into the narrative to express emotions and advance the plot. (*(ii) Music:* Centers around the creation and performance of music.

6. Lighthearted and Family-Friendly

(i) Comedy: Intended to entertain and amuse with humor and satire. *(ii) Family:* Suitable for all ages, focusing on family values and adventures.

7. Mystery and Intrigue

(i) Mystery: Solving puzzles, crimes, or unexplained events. *(ii) Thriller:* Often involves suspense and investigation, overlapping with mystery elements.

Each group highlights how genres share common themes and elements, making them semantically similar.

III. FAMOS Features

The FAMOS features of the keyframes used in the whole experiment are as follows,

Focus (F): Focus refers to the central element or subject in a visual scene or image that draws the viewer's attention. It is what the image is primarily centered around or emphasizing.

Action (A): Action captures the movement or activity taking place within a scene. It involves the behaviors, gestures, or interactions of subjects within the image.

Mood (M): Mood refers to the emotional tone or atmosphere conveyed by a visual scene. It reflects the feelings or emotions that the image evokes in the viewer.

Object (O): Object refers to the physical items or entities present within the scene. These are the tangible elements, such as furniture, tools, or other visible items that contribute to the composition of the scene.

Setting (S): Setting describes the environment or location in which the scene takes place. It provides context about where the action occurs and the surrounding elements that shape the narrative of the image.

IV. Selection of VLM and LLMs in the experiment

In the experiment, for extracting the FAMOS [1] features the VLM model LLaVA [2] has been used. LLaVA stands out for several reasons in your context. It excels at processing both visual and linguistic cues, which is crucial for understanding movie frames, thanks to its vision-language alignment. Compared to other VLMs like BLIP-2 [3] and MiniGPT-4 [4], LLaVA is more efficient in handling diverse multimodal inputs without sacrificing performance. It also offers a balanced trade-off between computational cost and interpretability, unlike Flamingo [5], which is more resource-intensive. Additionally, LLaVA's open-source nature makes it accessible and affordable, providing robust support for keyframe analysis in our experiment.

Vicuna[1] has been chosen for generating visual summaries due to its balance of cost-effectiveness and capability. Unlike GPT-3.5, which is highly effective but limited by high costs and token restrictions, Vicuna, an open-source LLM, is more accessible while still offering strong performance in language understanding tasks. It has demonstrated notable accuracy in genre prediction from audio transcripts in previous studies [6], making it a practical alternative to GPT-3.5-turbo[2]. Furthermore, its ability to handle extensive input data without the token constraints of GPT-3.5 made it ideal for our experiment.

GPT-3.5-turbo has selectively been used in our experiment to predict genre from visual summaries and extract movie meta-information (e.g., director, cast) due to its superior reasoning ability and understanding of nuanced content. While it could have replaced Vicuna for both summary generation and genre prediction, doing so would have significantly increased costs. By limiting GPT-3.5 Turbo's use to these high-impact tasks, we balanced performance and cost, leveraging its strengths where they mattered most while keeping the experiment cost-effective.

V. Why Hindi movie only?

We focused on Hindi-language movies in our experiment due to the Bollywood industry's significant global presence and revenue generation, making it a vital area for research. While most previous work centers on Hollywood, Bollywood serves a substantial and diverse audience. The FlickScore dataset provided limited data for other regional languages, with few user interactions and outdated content, making them unsuitable for reliable recommendation tasks. Among 611 Hindi films, only 510 had accessible trailers, further supporting our decision to concentrate on Hindi films for better data reliability and representativeness in recommendation systems.

References

[1] V. Himakunthala, A. Ouyang, D. Rose, R. He, A. Mei, Y. Lu, C. Sonar, M. Saxon, and W. Y. Wang, "Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought," *arXiv preprint arXiv:2305.13903*, 2023.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[3] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[4] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[6] P. Mondal, S. Singh, Kushum, S. Saha, J. P. Singh, B. Singh, and N. Pedanekar, "Efficacy of large language models in predicting hindi movies' attributes: A comprehensive survey and content-based analysis," in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 947–950.

[1] https://lmsys.org/blog/2023-03-30-vicuna/

[2] https://chatgpt.com/

TABLE I
FIVE DIFFERENT PROMPTS USED IN THE PROPOSED STUDY

Prompt	**Remarks**
Prompt 1 = ”USER: \nCould you tell me about the Focus, Action, Mood, Objects, and the Setting of the scene? Here are the definitions of the different aspects: Focus (F): Focus refers to the central element or subject in a visual scene or image that draws the viewer’s attention. It is what the image is primarily centered around or emphasizing. Action (A): Action captures the movement or activity taking place within a scene. It involves the behaviors, gestures, or interactions of subjects within the image. Mood (M): Mood refers to the emotional tone or atmosphere conveyed by a visual scene. It reflects the feelings or emotions that the image evokes in the viewer. Object (O): Object refers to the physical items or entities present within the scene. These are the tangible elements, such as furniture, tools, or other visible items that contribute to the composition of the scene. Setting (S): The setting describes the environment or location in which the scene takes place. It provides context about where the action occurs and the surrounding elements that shape the narrative of the image. Specify each of the aspects separately with sub-headings.\nASSISTANT:”	
Prompt 2 = f”””Consider all the pairs of {a1} and {a2} and finally generate the movie theme-based single line {a1}that generalizes all the pairs of {a1} and {a2}. Take as much time as you want to analyze and generate the response. Do not allow any impact on your previous response to generate the current response. NOTE: GENERATE ONLY A SINGLE LINE RESPONSE. The {a1}-{a2} pair is as follows: ”{aspect_value}”. ””” prompt_template=f”””A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user’s questions. USER: {prompt} ASSISTANT: ”””	Here, a1 = $(\theta^{+} \setminus$ a2) and a2 = {Object}
Prompt 3 = f””” Suppose you are good at analyzing and summarizing the movie’s cinematography information. Now, consider a movie’s cinematography information given below, analyze it, and finally summarize it. Please summarize in small sentences and words. Use a maximum of 100 words strictly in response. The given cinematography information of movie = ”{responses}” Don’t allow any effect on your previous response. Take as much time as you require to generate the response. Generate the output as follows: Summary: ””” Prompt_template=f”””A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user. USER: {prompt} ASSISTANT: ”””	Here, response = $\mathcal{G}(\theta^{t})$
Prompt 4 (Primary Genre) = f ”””Suppose you are good at analyzing a movie’s summary and predicting the appropriate genres of the movie. Now, consider the text delimited by triple backticks, find the movie’s summary, analyze it, and finally generate a single genre that suits the movie summary. Choose the genre from the list provided below: [Biography, Musical, Mystery, Adventure, Action, War, Music, Fantasy, Romance, Thriller, Family, History, Crime, Drama, and Comedy] Don’t allow any effect on your previous response. Take your time as much as you require in generating the response and from the provided list, send the genre as you think is most suitable for the movie. Please don’t suggest any genres out of the provided genre list. Generate the output as follows: [] “‘{ text }“‘ ”””	Here, text = $Summary^{v}_{\theta^{t}}$
Prompt 4 (Secondary Genre) = f ”””Suppose you are good at analyzing a movie’s summary and predicting the appropriate genres of the movie. Now, consider the text delimited by triple backticks, find the movie summary of the trailer, analyze it, and finally generate a list of the most suitable genres of the movie summary. Choose the genres from the list provided below: [Biography, Musical, Mystery, Adventure, Action, War, Music, Fantasy, Romance, Thriller, Family, History, Crime, Drama, and Comedy] Don’t allow any effect on your previous response. Take your time as much as you require in generating the response and from the provided list, send as many genres as you think most suitable for the movie. Please suggest at least three genres and don’t suggest any genres out of the provided genre list. Generate the output as follows: [] “‘{ text }“‘ ”””	Here, text = $Summary^{v}_{\theta^{t}}$
Prompt 5 = f ””” Consider the movie information delimited by triple backticks and based on its name with the year of release, send its corresponding director name and casts in the appropriate format. Generate 5 most suitable casts. Don’t allow any effect on your previous response. Take as much time as you require to generate the response. Do not give any extra line. Only generate the output in the following dictionary format: {{ ’Director’: ’Director name’, ’Casts’: ’Casts names’ }} { movie_title_yor } ”””	Here, movie_title_yor = (item name, YoR)