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 centred around the solving of a puzzle, crime, or unexplained event, often involving a detective or investigator.

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

5. Action: Films that emphasise 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 dramatise 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 humour, satire, and lighthearted stories.

II. Genre Group

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

- **Drama and Emotional Focus**
  - **Biography**: Focuses on real-life stories and personal achievements.
  - **Drama**: Emphasizes realistic characters and emotional themes.
  - **Romance**: Centres around romantic relationships and emotional journeys.
- **Excitement and Suspense**
  - **Action**: High-energy scenes with physical feats and battles.
  - **Thriller**: Creates suspense and excitement, often involving crime or psychological elements.
  - **Adventure**: Exciting journeys and quests, often involving exploration and battles.
  - **Crime**: Revolves around criminal activities and law enforcement pursuits.
- **Historical and Real-World Events**
  - **History**: Portrays historical events and figures.
  - **War**: Focuses on armed conflict and the effects of warfare.
  - **Biography**: Real-life stories and historical figures.
- **Fantasy and Imagination**
  - **Fantasy**: Imaginary worlds with magical elements and mythical creatures.
  - **Adventure**: Can overlap with fantasy in terms of quests and exploration.
- **Music and Performance**
  - **Musical**: Songs integrated into the narrative to express emotions and advance the plot.
  - **Music**: Centres around the creation and performance of music.
- **Lighthearted and Family-Friendly**
  - **Comedy**: Intended to entertain and amuse with humour and satire.
  - **Family**: Suitable for all ages, focusing on family values and adventures.
- **Mystery and Intrigue**
  - **Mystery**: Solving puzzles, crimes, or unexplained events.
  - **Thriller**: Often involves suspense and investigation, overlapping with mystery elements.

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

TABLE I
Modality-wise Primary Genre (PG) and Secondary Genre (SG) prediction percentage

Genres	Video (V)		Audio (A)		Video_Audio (VA)	
	PG	SG	PG	SG	PG	SG
Action	19.94	9.06	10.24	8.8	15.61	6.82
Adventure	8.15	3.67	0.67	1.28	2.12	2.18
Biography	-	0.41	0.22	0.55	-	0.45
Comedy	14.61	10.18	30.29	9.17	15.08	11.55
Crime	6	4.58	10.24	12.10	4.5	7.91
Drama	7.58	28.72	9.35	29.42	1.85	27.36
Family	2.53	5	0.89	3.3	1.06	3.18
Fantasy	0.84	0.61	2	0.92	0.53	0.73
History	0.28	1.22	0.89	2.2	0.53	1.45
Music	4.78	0.92	12.69	6.51	8.99	2.82
Musical	-	1.12	0.67	1.1	-	1.82
Mystery	-	6.42	-	3.57	-	6.45
Romance	18.82	16.7	4.68	9.53	17.46	13.27
Thriller	14.89	10.59	13.59	9.9	23.81	11.73
War	1.69	0.81	3.56	1.65	8.47	2.27

III. Modality-wise genre prediction

Table I presents the individual genre prediction by Video (V), Audio (A) and Vedio_Audio (VA) modality. Both the

Primary Genre (P) and the Secondary Genre (SG) have been reported in the table. The prediction of the 15 genres is measured in percentage.

IV. Impact of our study

Our proposed model, FAMOS, aims to demonstrate that **genre prediction from video content can be achieved effectively and efficiently** without requiring additional textual metadata, expensive training pipelines, or user input. Below, we summarise the key dimensions of the impact that FAMOS brings to the field:

1. **Content-Based, Metadata-Free Genre Prediction:** Unlike traditional systems that depend on user-entered textual metadata (e.g., tags, reviews, descriptions), our approach uses raw video trailers, which are more readily available than structured datasets as the sole input. This allows genre inference in real-world scenarios where external metadata is missing or incomplete.

2. **Trailer-Based Efficiency:** Our system operates on short-length trailers instead of full-length movies, thereby drastically reducing computational cost and inference time. This design makes FAMOS suitable for a fast and effective system.

3. **No Training, No Data Dependency:** FAMOS leverages the global reasoning capabilities of pretrained vision-language (VLM) and language models (LLM). Since we do not perform any additional training or fine-tuning, the model remains lightweight and resource-efficient, eliminating the need for large labelled datasets or compute-intensive model updates.

4. **Lightweight Yet Effective**: Despite its simplicity, the proposed system achieves high-quality results, demonstrating that pretrained models can be adapted for genre prediction in a plug-and-play fashion without sacrificing performance.

5. **Empirical Strength of Visual Content (Table IV):** The results in Table IV highlight the superior performance of visual features over audio features in genre prediction, reinforcing the relevance of our video-based approach and the importance of multimodal representation selection.

6. **Feature-Level Insights Supporting Real-World Relevance (Table VI):** Table VI further supports the semantic soundness of our feature selection: Setting and Mood individually perform well, and the pairwise combination of Setting and Object yields the best two-feature result. Ultimately, the complete combination of all feature sets in our proposed model achieves the best overall performance, showing how different aspects of video semantics contribute meaningfully to genre inference.

Together, these points clarify that our core contribution is a modular, data-independent, and efficient genre prediction system that can be deployed in practical applications involving large video corpora, including content tagging, cataloguing, search optimisation, and personalisation.

V. Selection of VLM and LLMs

1. Visual-Language Model (VLM): LLaVA

a) Why LLaVA?: We selected LLaVA for extracting visual semantics (FAMOS aspects) because it strikes an optimal balance between cost-efficiency, open availability, and vision-language alignment performance. Unlike large proprietary models like Flamingo (DeepMind) or MiniGPT-4, LLaVA is fully open-source and exhibits competitive accuracy in image-to-text understanding tasks with much lower computational requirements.

In the genre prediction context, the ability of a VLM to interpret scenes, actions, and visual mood cues is critical. LLaVA, built on a Vicuna backbone with CLIP-based visual encoders, has been shown to perform well on instruction-following benchmarks and can accurately describe frame-level video content, even when the trailer scenes are complex or noisy.

We considered alternatives like:

- BLIP-2: Highly effective, but often verbose and unstable across languages.
- MiniGPT-4: More accurate, but significantly slower and GPU-heavy, making real-time or batch processing difficult without high-end infrastructure.

Thus, LLaVA was selected not just for being performant, but also for being lightweight, reproducible, and capable of generalising across genres without fine-tuning.

2. Language Models (LLMs): Vicuna and GPT-3.5-turbo

b) a. Vicuna for Visual Summary Generation:: Vicuna (13B) is a strong open-source LLM trained on instruction-following datasets.

It supports long prompts and is less sensitive to input length compared to API-bound models like GPT-3.5.

We found it effective in generating genre-aware summaries from extracted FAMOS aspects, maintaining coherence without hallucination.

Its cost-free deployment makes it suitable for large-scale or iterative genre tagging.

c) b. GPT-3.5-turbo for Genre Classification and Metadata Extraction:: Despite its token-length limitations, GPT-3.5 is one of the most robust few-shot classifiers available today.

We have used it for zero-shot genre prediction and metadata extraction tasks, where its superior reasoning over summaries was especially valuable.

It consistently produced more accurate predictions than supervised models, as seen in Table VII(a), where GPT-3.5 achieved an F1 of (62.70 ± 0.31) with no training, comparable to traditional classifiers that require annotated data.

d) Why not use GPT-4, Claude, or Mistral?: GPT-4 and GPT-4o-mini provide marginally better performance (F1 = 66.15), but come at a significant cost and with access constraints.

Models like Claude 2 and Mistral 7B were evaluated offline and found less suitable for structured tasks like genre classification from summaries, often producing creative but less grounded outputs.

3. General Strategy: Modular, Lightweight, Zero-Shot

Our strategy with pretrained model selection was not to chase the largest or newest models, but rather to choose models that:

Align well with genre prediction as a zero-shot task (i.e., no labelled training data required).

Offer low latency and high throughput, allowing them to be scaled for real-world video libraries.

Are robust to multilingual and multimodal inputs, including Indian languages and trailer-specific visuals.

Contribute to a pipeline where each stage (aspect extraction, summary generation, genre prediction) can be independently evaluated or swapped out.

VI. How genre prediction helps in movie recommendation system

1. Addressing the Cold-Start Problem in Recommender Systems Genre is a high-level semantic descriptor that is often used in Content-Based Filtering (CBF) and hybrid recommender systems, especially to bootstrap new items into the system when no user interaction data is yet available (the "cold-start" problem)

FAMOS enhances this by:

Automatically extracting genres from trailers, even without metadata.

Replacing the traditional reliance on user-generated or manually curated tags (which are labour-intensive and error-prone).

Reducing human dependency in tagging, making systems more scalable.

By automating genre extraction, FAMOS expands the applicability of CBF approaches in cold-start scenarios, especially on platforms dealing with a high influx of new or user-generated video content.

2. Real-Time, Scalable Metadata Generation for RS Pipelines In large-scale RS applications (e.g., YouTube, Netflix), one key bottleneck is generating consistent, explainable metadata (genre, tone, mood, setting) across massive video libraries.

Our approach:

Uses short trailers instead of full-length movies, enabling low-cost inference.

Avoids supervised training and instead leverages generalised pretrained knowledge from VLMs and LLMs.

Can be integrated upstream in recommendation pipelines to enrich item profiles before collaborative filtering, matrix factorisation, or ranking.

This supports the construction of self-updating catalogues, helping recommender systems remain fresh, contextual, and linguistically inclusive.

3. Improved User Preference Modelling Using Genre-Level Semantics Our paper already demonstrates a working user preference prediction framework using genres (primary and secondary), directors, and cast. Table V shows that the genre component derived from FAMOS plays a dominant role in shaping user preferences. This matters because:

Genres map to user intent and interests better than low-level embeddings.

Fine-grained genre analysis (primary + secondary genres) allows personalised filtering, sentiment control (e.g., excluding horror), and diversity-aware recommendation.

FAMOS's extraction of detailed, interpretable genre semantics provides transparent and controllable signals to downstream RS models.

4. Contributions to Emerging Multimodal Recommender Systems Recent trends in RS research (e.g., YouTube MIND, Amazon's VLM-RS efforts) point toward multimodal and language-inclusive recommendations, especially with the rise of video content. FAMOS aligns with these trends by:

Unifying visual and textual modalities for genre inference.

Operating across Indian languages (Hindi, Bengali, Malayalam, Telugu), demonstrating applicability in multilingual markets.

Offering a modular architecture that can be adapted for other modalities, including posters or full movie transcripts.

Thus, the methodology is not only impactful in its own task (genre prediction) but also enables future research on vision-language-aware and culturally adaptive RS.

5. Bridging Human Perception with Model Reasoning Genres are human-interpretable concepts. Integrating genre prediction into recommender systems helps bridge the gap between:

Opaque model outputs (e.g., latent factors in matrix factorisation), and

Explanatory, user-facing metadata (e.g., "You're seeing this movie because it's a mystery-drama, like others you enjoyed").

This improves trust, explainability, and regulatory transparency in modern RS design—an increasingly important concern in AI-driven personalisation.

Summary The contribution of genre prediction to RS research can be concluded as:

A key enabler for robust content modelling,

A powerful tool for cold-start and zero-data recommendation,

A scalable solution for metadata enrichment, and

A foundation for explainable, multimodal, and inclusive recommender systems.

TABLE II
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 centred around or emphasising. Action (A): Action captures the movement or activity taking place within a scene. It involves the behaviours, 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 analyse 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, analyse it, and finally summarise it. Please summarise 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, analyse 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, analyse 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 the 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 lines. 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)