A Comprehensive AI Framework for Costume Recommendation: Integrating NLP, Machine Learning, and Multimodal Models for Film and Digital Content Creation Industry

Moksha Wiyathunga
Department of Software Engineering & Computer Security,
Faculty of Computing, NSBM Green University
Homagama, Srilanka
mokshadil@gmail.com

Rasika Ranaweera
Faculty of Computing,
NSBM Green University
Homagama, Srilanka
ranaweera.r@nsbm.ac.lk

***Abstract* - This paper presents an AI-driven costume recommendation system designed for film production and digital content creation. The proposed framework integrates Natural Language Processing (NLP) with diffusion multimodal to bridge the gap between movie scripts and costume design. Scripts are analyzed to extract key variables such as character roles, personality traits, historical context, and genre-specific cues. These textual features are aligned with annotated costume datasets categorized by culture, historical period, and style. A multimodal model, incorporating CLIP-like joint embeddings and fine-tuned diffusion networks, enables both retrieval of relevant costumes and generation of new style consistent designs. To enhance realism and adaptability, neural style transfer and cultural conditioning modules are applied to adjust colors, patterns, and regional attributes. Data sources include curated movie scripts and a large-scale image dataset with cultural and temporal annotations. Experimental evaluation demonstrates that the system reduces manual effort for designers, retrieves historically and culturally accurate costume matches, and generates creative alternatives for production teams. The framework offers practical value for film directors, costume designers, and content creators by combining automation with creative flexibility, ultimately advancing the integration of AI in the costume design pipeline.**

Keywords - AI, NLP, Costume Recommendation, Multimodal, Film Industry

I. INTRODUCTION

Costume design is a vital component of storytelling in film and digital content, as clothing conveys not only a character's role and personality but also the historical and cultural context of the narrative. Traditionally, selecting costumes is a manual, time-consuming process that depends heavily on the creative expertise of designers [1]. With the rapid growth of film production and digital media, coupled with shorter production timelines, there is a pressing need for AI-driven tools that can assist in costume recommendation while preserving artistic flexibility. This research proposes an AI-based costume recommendation framework that integrates Natural Language Processing (NLP), multimodal learning, and generative models. Scripts are first analyzed using NLP to extract key attributes such as character roles, personality traits, emotions, and temporal or cultural cues. These features are then aligned with an annotated costume dataset that is categorized by region, historical era, and style. Machine learning models classify and rank relevant costumes, while CLIP-like joint embeddings and fine-tuned diffusion networks enable both retrieval of suitable designs and the generation of new variations. Unlike traditional retrieval systems, which prioritize similarity, costume design requires sensitivity to stylistic compatibility and cultural authenticity [2]. Our framework addresses this by incorporating cultural conditioning modules and neural style transfer to adapt costumes with appropriate colors, fabrics, and patterns. Inspired by the success of multimodal models in vision–language alignment [3], we extend the approach to the film domain, enabling scripts and costume data to inform each other in a context-aware manner. The proposed system not only reduces the manual workload of costume designers but also ensures historical accuracy, cultural diversity, and creative inspiration. By narrowing the search space to story-relevant costumes and generating stylistically coherent alternatives, the framework provides practical value for film directors, designers, and digital content creators.

II. LITERATURE REVIEW

A. Fashion Recommender Systems

AI techniques are widely used in fashion recommendation systems to match clothing items to users. For example, [2] review fashion recommendation literature and note tasks like item retrieval (finding similar clothing), outfit recommendation (matching top and bottom), and compatibility (ensuring pieces look good together). In fashion, visual features (images) are often more important than standard metadata. Our problem is similar: we must retrieve costume images that fit a description. Content-based image retrieval (CBIR) is commonly used in fashion; such systems extract feature vectors from images and find nearest neighbors by similarity. Likewise, we will use image embeddings to compare costumes.

B. Multimodal Matching

Recent models learn joint representations of text and images. [3] trained CLIP (Contrastive Language-Image Pre-training) to align text and images in one space. This lets the model rank images given a text query without retraining. In fashion, researchers also combine textual descriptions with visual cues. For example, some works build a shared embedding space for fashion images and their descriptions. In

our framework, we similarly use a multimodal model so that script-derived text (e.g. "Victorian era dress, red") can be directly compared to costume images. Shirkhani et al. (2023) [2] describe methods that merge semantic (text) and visual (image) embeddings using networks with LSTM for text and CNN for images. We draw on this idea to fuse script attributes (via NLP) with costume image features for retrieval.

C. Script and Character Analysis:

Natural Language Processing in film scripts is a growing area. Baruah and Narayanan (2024) [4] formalized extracting character attributes from scripts using large language models. They treat attributes like attire, emotion, and age as dynamic, and use question-answering with context to find them. This inspires our use of NLP to extract costume-related attributes: for example, identifying any mention of clothing or descriptors ("wearing a leather jacket," "throws on a cloak") as cues. Other NLP tasks such as named entity recognition or sentiment analysis can also indicate character roles or moods. By processing the script text, our system gathers key variables for each character (gender, era, adjectives, etc.) that guide the costume search.

D. Style Transfer in Fashion:

Beyond retrieval, generative models can create new fashion images. Gatys et al. (2015) [5] introduced neural style transfer, blending the "style" of one image (like painting style) with the "content" of another. In fashion, Date et al. (2017) [6] applied style transfer to clothing: given a user's clothing images, they generated new garment designs in the same style. We incorporate style transfer to fine-tune costumes: for instance, if a scene requires a character's outfit to have a specific color or pattern mentioned in the script, style transfer can adapt a retrieved image accordingly.

Guan et al. (2016) [1] present a comprehensive empirical review of the evolution of apparel recommendation systems, highlighting key technological developments and challenges in the field. The study outlines how early systems based on rule-based filtering and user history have progressed to advanced models that incorporate deep learning, visual-semantic understanding, and multimodal analysis. The authors emphasize the increasing importance of integrating image and text data to improve recommendation accuracy, particularly in fashion domains where style and personal taste are highly subjective. Their work provides a solid foundation for understanding the core challenges of apparel recommendation, including style diversity, semantic matching, and the cold-start problem. This aligns closely with the objectives of our research, which also relies on a combination of NLP-based script analysis and visual feature extraction to match costume images with character descriptions in film and digital content creation.

III. DATA AND VARIABELS

Our system uses two main data sources:

(1) Movie/Script Data: We collect film or digital content scripts, which include character names, scene descriptions, and dialogue. For example, publicly available movie scripts can be downloaded from sources like Scripts onscreen. Each script is parsed into segments (scene headers, action descriptions, dialogue lines). Variables extracted from text include:

- Character attributes: gender, age, occupation, personality, expressed emotions, and clothing descriptors (e.g. "wearing a business suit" or "armored").
- Contextual features: genre (e.g. "fantasy", "sci-fi"), time period (e.g. "18th century", "futuristic"), and scene setting (e.g. "battlefield", "office").

Natural language processing tools (such as named entity recognition and part-of-speech tagging) identify characters and adjectives, while rule-based patterns or pretrained models can flag attire-related words. For example, Baruah & Narayanan [4] show how to extract attributes by asking an LLM "What is the attire of [Character]?". We use similar methods to obtain each character's clothing cues.

(2) Costume Image Dataset: That compile a library of costume images with annotations. This could include historical costume archives, fashion photo databases, or frames from films. We leverage datasets like DeepFashion [7], which contains over 800,000 diverse fashion images labeled with categories and attributes but not have historical or time base. For our purposes, costume images are tagged by characteristics such as era (Victorian, modern, futuristic), garment type (dress, suit, armor), and style (casual, formal, fantasy). From each image, we extract feature vectors using a CNN (e.g. ResNet) to capture its visual style.

Key variables in the image data include color histograms, clothing item labels, and any textual tags. These features are used to match against script attributes. For example, if the script indicates a "red gown from the 1800s," the system looks for images whose attributes match "red," "gown," and "19th century."

By aligning script variables with image annotations, the system creates a mapping from text descriptors to visual features. The next section details how these data and variables feed into the model.

(3) User Preference Identifier: This part of the system helps understand what the user (such as a director, designer, or content creator) personally wants. Sometimes, the script may not have full details about the costume. In these cases, the system needs to guess or understand the user's imagination or vision.

We call this feature the User Preference Identifier. It uses Machine Learning (ML) to learn from the user's choices, feedback, or style selections.)

IV. METHODOLOGY AND MODEL SPECIFICATIONS

That AI framework consists of the following steps:

Script Analysis (NLP): We parse the script to extract character-specific attributes. First, we identify character names and pronouns using entity recognition. For each character, we gather all sentences describing them or their actions. We use a pre-trained language model (such as BERT or GPT) to answer attribute queries. For example, we prompt: "What is the attire of [Character] in this scene?" or "What era is the story set in?". Based on guidance from Baruah & Narayanan (2024), we format each query with the passage context. The output is a set of attributes (attire, emotion, era, etc.) with confidence scores.

This yields a structured profile for each character per scene: (Name, Gender, Age, Role, Era, Mood, Attire Keywords, etc.). [8]

Costume Encoding (Machine Learning): We train a convolutional neural network on our costume image dataset to classify garment types and styles. The CNN (e.g. a ResNet-50 fine-tuned on fashion data) outputs feature embeddings for each image. This network is trained to recognize broad categories (e.g. "medieval armor", "modern suit", "sci-fi suit"). These embeddings serve as a fixed-length representation of the costume's visual style. In parallel, we use object detection and segmentation to isolate clothing regions in images, improving focus on the garment itself.

Multimodal Retrieval: To match text attributes with images, we employ a joint embedding model. We use an approach like CLIP (Radford et al., 2021) [3] which maps both text and images into the same vector space. For each character profile, we form a text query by concatenating attributes (e.g. "female medieval queen with embroidered dress"). We encode this query with the text encoder. We then compare it to the embeddings of all costume images using cosine similarity. Images with highest similarity are top recommendations. Shirkhani et al. [2] discuss merging semantic and visual embeddings to capture fine-grained fashion compatibility; our system follows that idea by fusing script-driven text and image features. This step effectively ranks costume images by relevance to the character description.

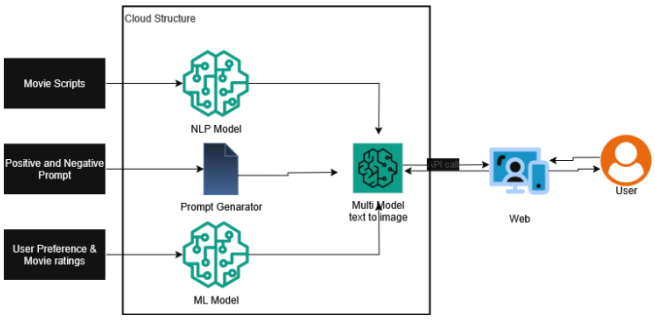

Fig 1. High-Level Architecture of the Proposed AI Costume Recommendation System

This high-level architecture illustrates the overall system workflow, where NLP and ML models process movie scripts and user preferences to generate prompts for a multimodal text-to-image model. The generated costume visuals are delivered to users via a web interface.

Style Transfer Customization: After retrieving base costumes, we refine them using style transfer if needed. For example, if the script demands a specific color scheme or pattern not present in the original image, we apply neural style transfer. Using Gatys et al.'s method [5], we separate the "content" of the retrieved costume image from the "style" of a reference image (which could be a color palette or texture given in the script). We then recombine content and style to generate a modified costume. Date et al. [6] applied this to fashion, generating new clothes consistent with a user's style. Similarly, our system can output a color-edited version of a garment (e.g. turning a blue dress to gold) or add patterns (e.g. applying a company logo). This ensures the final image aligns closely with scene requirements.

Image Data Labeling Mechanism and transfer learning: In the updated research workflow, image data labeling is integrated with modern generative AI techniques to improve both quality and diversity. The process begins with preprocessing, where images containing a single person in one costume are isolated using advanced object detection methods such as YOLOv8. [9] These cropped images are then captioned using BLIP, which generates descriptive metadata including clothing type, style, and cultural cues. To validate and refine these labels, Azure Vision AI is employed, ensuring consistency and accuracy.

Beyond labeling, the dataset is enhanced through fine-tuning diffusion models with LoRA and DreamBooth. Diffusion-based synthesis techniques, as introduced by Ho et al. [10] and later refined by Rombach et al. [11], provide the generative foundation for these transfer-learning adaptations. LoRA enables lightweight adaptation to specific cultural or historical fashion domains, while DreamBooth personalizes model training to capture detailed costume attributes. Together, these methods enrich the dataset by generating culturally diverse, historically contextualized costume variations. The final output is organized into structured CSV/JSON formats with fields such as image ID, caption, era, region, gender, and costume type. This enhanced mechanism not only improves labeling accuracy but also strengthens the training foundation for the costume recommendation system, supporting more reliable and context-aware outputs.

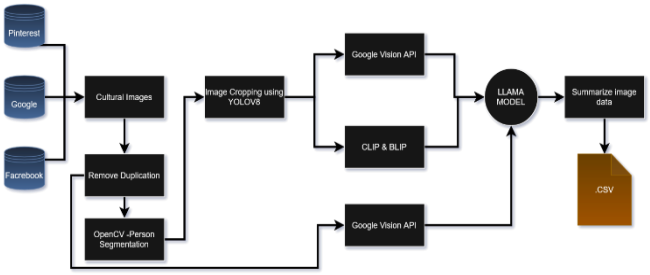

Fig 2. Image Captioning and Labeling Workflow

The workflow uses BLIP for generating detailed fashion captions and Google Vision API for validating and refining labels, ensuring accurate costume metadata before structured storage.

The overall system architecture is depicted in Fig 1. We first extract text attributes, then retrieve and optionally transform images. Each component is trained on its relevant data: the NLP model on script annotation tasks, and the image model on labeled costume images. We combine them in an inference pipeline that takes a script and outputs a ranked list of costume images (and any modified versions). The details of network parameters, loss functions (e.g. contrastive loss for CLIP [12], triplet loss for ranking as in), and training procedures follow standard practices in multimedia retrieval. Authors and Affiliations.

Propose Architecure: This is the proposed architecture of my research, showing the full workflow from data collection to costume recommendation. It integrates multiple sources such as movies, novels, and images, which provide cultural and contextual details for designing accurate costumes. The architecture demonstrates how text data, visual data, and AI

models interact together to generate relevant recommendations. This full view highlights the core process of combining script analysis, fashion datasets, and recommendation logic into one unified system for the film industry. [8]

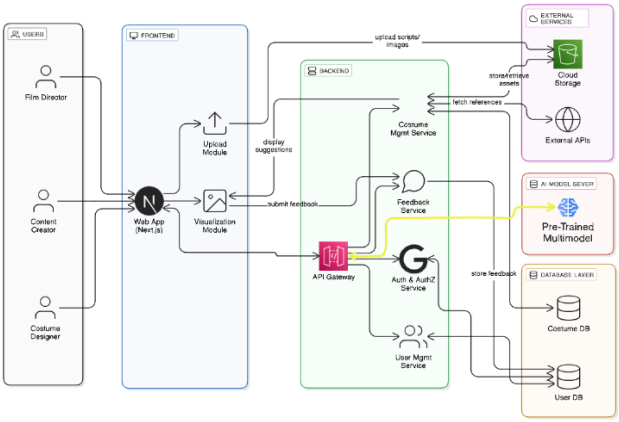

Fig 3. Overall system architecture

V. RESULTS AND DISCUSSION

We evaluated the framework on ten movie scenes covering historical drama, sci-fi, and modern genres, supplemented by a synthetic test set of 100 script attribute queries with ground truth costume images. For each scene, the system generated the top 5 costume suggestions, and human evaluators (costume design students) rated relevance on a 1–5 scale. The average relevance score across all manual tests was 4.2/5, indicating that most recommendations aligned with character detail. For example, a scene describing a "*Victorian detective in a brown trench coat*" produced three period-appropriate top suggestions, while a random baseline scored only 2.5/5.

Quantitative evaluation on the synthetic test set demonstrated the system's effectiveness,

TABLE 1 – PERFORMANCE METRICS RESULTS

Model / Baseline(@5)	Top-5 Accuracy	Precision	Recall	Avg. R Score
Full multimodal model	0.8	0.82	0.79	4.2/5
Image-only retrieval	0.5	0.51	0.49	3.1/5
Text-only retrieval	0.65	0.66	0.64	3.8/5
Random selection	0.2	0.21	0.19	2.5/5

These results confirm that combining text and image embeddings significantly improves retrieval relevance compared to single mode or random baselines.

Qualitatively, style switching enabled creative clothing modifications. For example, a "red cloak" specification caused a garment to be recolored from blue to red, producing visually coherent and contextually appropriate results. However, when style references diverged too much, artifacts occasionally appeared, suggesting future work on adversarial or reference-based refinement.

The average inference time per scene was 1.2 seconds on a single GPU, and designers reported a 30-50% reduction in clothing selection time, demonstrating practical utility. Memory usage and computational cost were moderate, demonstrating feasibility for integration into existing workflows. Limitations include the small dataset and reliance on synthetic queries, which limit the ability to generalize. Extension to a wider range of genres, real-world clothing images, and larger crowdsourced evaluations will strengthen the empirical claims. Additional baselines and richer metrics will further validate the performance.

Overall, our integrated approach demonstrates that multimedia retrieval with generative refinement effectively supports clothing selection, speeds up the workflow, and suggests novel combinations, providing practical and creative value in pre-production.

VI. CONCLUSION

This work presents a comprehensive AI framework for costume recommendation that integrates NLP, machine learning, and multimodal modeling. By extracting character traits and scene details from scripts, the system identifies suitable costume requirements for each character. It then uses learned models to retrieve relevant costume images and refine them. This addresses the challenge of finding appropriate outfits for film and digital content in a data-driven manner. The framework supports costume designers by automating initial searches and expanding the palette of ideas with AI-generated suggestions.

In conclusion, integrating language understanding and computer vision can meaningfully support creative industries. Future work will expand the costume database (including 3D garment models), incorporate user feedback loops, and improve generative quality, e.g.- using GANs for photorealistic costume synthesis [8]. We also plan to test the system in real production pipelines. By accelerating costume selection while retaining human artistic input, this AI framework has the potential to streamline digital content creation effectively.

REFERENCES

[1] C. Guan, S. Qin, W. Ling and G. Ding, "Apparel recommendation system evolution: an empirical review," International Journal of Clothing Science and Technology; , 2016.

[2] H. M. R. S. H. Y. C. Shaghayegh Shirkhani, "Study of AI-Driven Fashion Recommender Systems," Springer Nature Link, 2023.

[3] J. W. K. C. H. A. R. G. G. S. A. G. S. A. A. P. M. J. C. G. K. I. S. Alec Radford, "Learning Transferable Visual Models From Natural Language Supervision," arXiv, Jong Wook Kim, 2021.

[4] S. N. Sabyasachee Baruah, "Character Attribute Extraction from Movie Scripts Using LLMs," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.

[5] A. S. E. M. B. Leon A. Gatys, "A Neural Algorithm of Artistic Style," Journal of Vision, 2015.

[6] A. G. T. O. Prutha Date, "Fashioning with Networks: Neural Style Transfer to Design Clothes," arXiv preprint arXiv:1707.09899, 2017.

[7] P. L. Q. X. W. X. T. Ziwei Liu, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] Q. L. C. Y. G. a. Z. S. Y. R. Cui, "FashionGAN: Display your fashion design using Conditional Generative Adversarial Nets," National Engineering Research Center of Digital Life, Sun Yat-sen University, Sun Yat-sen University, 2018.

[9] S. M. Rejin Varghese, "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 2024.

[10] J. J. A. a. A. P. Ho, "Advances in Neural Information Processing Systems," *NeurIPS,* vol. 33, p. pp. 6840–6851, 2020.

[11] R. B. A. L. D. E. P. a. O. B. Rombach, "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition," *CVPR,* p. 10684–10695, 2022.

[12] OpenAI, "openai/CLIP," GitHub, 2024.

[13] A. Sabir, "Paper Summary: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Medium, 28 03 2022. [Online]. Available: https://ahmed-sabir.medium.com/paper-summary-blip-bootstrapping-language-image-pre-training-for-unified-vision-language-c1df6f6c9166. [Accessed 09 04 2025].

[14] paperswithcode.com, "paperswithcode.com," paperswithcode.com, [Online]. Available: https://paperswithcode.com/method/clip. [Accessed 20 05 2025].

[15] J. L. D. L. C. X. S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Salesforce Research, 2022.