ETPDesigner: Multi-Agent Orchestration for Interactive Multimodal Electronic Theater Program

Appendix

A Implementation Details

A.1 Poster Prompt Formulation

To ensure the expert system possesses robust domain knowledge for poster generation, we curated a comprehensive database of professional theater posters and their corresponding scripts. Utilizing a Vision-Language Model (VLM) paired with rigorous human verification and supplementation, we systematically distilled established poster design rules, including overall stylistic aesthetics, color palettes, and layout structures. These visual design rules were then explicitly mapped to specific script attributes such as theme, theatrical genre, and emotional tone. The resulting knowledge base categorizes design priors across three primary genres (Tragedy, Comedy, and Drama) and spans a diverse array of narrative themes, including romance, science fiction, and revenge. By encompassing all standard theatrical classifications, we guarantee the broad applicability and practical effectiveness of the expert system. Building upon this foundational knowledge, the cognitive agents dynamically synthesize precise instructions for the generative process. This tightly coupled text-to-visual mapping ensures that the generated assets not only exhibit high artistic quality but also strictly adhere to the underlying dramatic semantics. The detailed rule templates and structural paradigms for formulating these poster generation prompts are illustrated in Figure 3.

A.2 Iterative Refinement Strategy

To rigorously guarantee the aesthetic and semantic quality of the generated posters, we designed an automated, reference-guided iterative optimization mechanism. Concurrently with the initial poster generation, the system queries the **ETP-Pro** benchmark to retrieve the top three most similar images based on the script's genre, theme, and emotional tone. These retrieved images serve as the ground- truth reference standards for the subsequent evaluation.

The refinement process commences with a critical validation of the key visual elements. If the primary visual anchor of the generated image misaligns with the reference standards, the system Abandons the current generative strategy, formulates a new core visual concept, and regenerates the initial image. Conversely, if the key elements are semantically correct, the system proceeds to score the current image against the references and formulates corresponding optimization suggestions.

The evaluation is systematically structured around eight fine-grained indicators categorized into three core dimensions. First, Theatrical Aesthetics assesses abstraction, tension, and refinement. Second, Communication Efficacy evaluates the integration of visual elements and semantic clarity. Third, Thematic Fidelity measures the resonance and contextuality of the image relative to the source script. An example of the comprehensive evaluation report is illustrated in Figure 1.

Algorithm 1 Iterative Refinement Workflow for Theatrical Posters

Require: Source script $\mathcal{S}$, ETP-Pro Benchmark $\mathcal{D}$, Max iterations $N_{max} = 3$, Threshold $\tau = 6.5$
Ensure: The optimal finalized poster I_{final}

$n \leftarrow 0$
Candidate set $C \leftarrow \emptyset$
$I_{curr} \leftarrow \text{GenerateInitial}(\mathcal{S})$
Retrieve References: $R_{1,2,3} \leftarrow \text{QueryTop3}(\mathcal{S}, \mathcal{D})$
while $n < N_{max}$ **do**
  Step 1: Key Visual Validation
  $is\_valid \leftarrow \text{ValidateVisualAnchor}(I_{curr}, R_{1,2,3})$
  if not is_valid **then**
    Abandon strategy and update core concept
    $I_{curr} \leftarrow \text{Regenerate}(\mathcal{S})$
    continue {Restart evaluation for the new image}
  end if
  Step 2: Multi-dimensional Scoring
  $S_{dim}, S_{comp} \leftarrow \text{Evaluate}(I_{curr}, R_{1,2,3})$ {Dimensions $D_{aes}, D_{eff}, D_{fid}$}
  $C \leftarrow C \cup \{(I_{curr}, S_{comp})\}$
  Step 3: Threshold-based Execution Logic
  $P_{opt} \leftarrow \emptyset$ {Initialize optimization prompts}
  $all\_pass \leftarrow \text{True}$
  for each dimension $D_k \in \{D_{aes}, D_{eff}, D_{fid}\}$ **do**
    if $\min(S_{dim}[D_k]) \leq \tau$ **then**
      $P_{opt} \leftarrow P_{opt} \cup \text{GetSuggestions}(D_k)$
      $all\_pass \leftarrow \text{False}$
    end if
  end for
  if all_pass **then**
    {All exceed threshold, enforce refinement on the weakest link}
    $D_{min} \leftarrow \arg\min_{D_k} \text{Average}(S_{dim}[D_k])$
    $P_{opt} \leftarrow \text{GetSuggestions}(D_{min})$
  end if
  Step 4: Execute Optimization
  $I_{curr} \leftarrow \text{ApplyRefinement}(I_{curr}, P_{opt})$
  $n \leftarrow n + 1$
end while
Step 5: Final Selection
$I_{final} \leftarrow \arg\max_{I_{curr} \in C} S_{comp}$
return I_{final}

To govern the refinement loop, we implement a strict threshold-based execution logic with a baseline score of **6.5**. If any indicator within a dimension falls below or equals this threshold, the corresponding optimization suggestions must be strictly executed. If all indicators within a dimension exceed 6.5, its modification suggestions are temporarily pended. In the optimal scenario where all indicators across all three dimensions surpass the 6.5 threshold, the

Evaluation Report


```
"evaluation_report": {
  "target_script": "Hamlet",
  "genre": "Tragedy",
  "theme": "Revenge, Mortality, Madness",
  "key_visual_validation": {
    "status": "Pass",
    "detected_elements": ["Skull (Yorick)", "Poisoned Chalice", "Dark Castle Silhouette"],
    "feedback": "The primary visual anchors accurately reflect the tragic and fatalistic essence of the source script."
  },
  "dimensional_scores": {
    "theatrical_aesthetics": {
      "abstraction": 7.8,
      "tension": 8.5,
      "refinement": 6.2
    },
    ......
  "optimization_suggestions": {
    "theatrical_aesthetics": [
      "The overall aesthetic lacks refinement. Enhance the micro-details and texture of the central skull to give it a more hyper-realistic and polished theater-prop quality." ],
      .......
  },
  "system_decision": {
    "action": "Enforce_Optimization",
    "target_dimension": "theatrical_aesthetics",
    "reason": "Indicator 'refinement' (6.2) is below the strict execution threshold of 6.5."
  },
```


Figure 1: Evaluation Report of Theatrical Poster

system enforces a refinement step by targeting the dimension with the lowest average score to further elevate the visual quality. This iterative loop is bounded by a maximum of three iterations. Upon reaching this limit, the system aggregates all generated candidates and outputs the image with the highest comprehensive score as the final poster. This candidate preservation strategy effectively safeguards the final output against potential visual degradation caused by over-editing in later iterations. The visual progression and corresponding quantitative tracking of **iterative experiment** are detailed in Figure 2 and Table 1.

Figure 2: Visual evolution of theatrical posters during the iterative refinement process. The figure illustrates the intermediate generation results and the step-by-step visual optimization guided by the evaluation metrics.

A.3 Global Style Extraction

Upon finalizing the optimal theatrical poster, the system initiates the style extraction module to establish a unified visual identity for the entire electronic theater program. We deploy the Vision-Language Model to analyze the generated poster and distill its core aesthetic attributes, specifically focusing on the color palette and the overall stylistic paradigm, such as a retro artistic aesthetic. Crucially, this extraction process deliberately filters out specific instance-level details and localized structural content, capturing only high-level abstract visual semantics. The VLM then condenses these extracted aesthetic priors into a series of concise descriptive phrases. These phrases are collectively stored and formalized as a global style vector, denoted as s_{global}. This vector is subsequently propagated to all downstream generative modules to guide the synthesis of corresponding background layouts and character portraits, thereby guaranteeing rigorous stylistic consistency across all generated components. This explicit global anchoring effectively eliminates the stylistic drift commonly observed in independent generation steps. The detailed rule templates are illustrated in Figure 4.

Prompt for Poster Synthesis

You are a professional theatrical poster designer. Your task is to generate an abstract stage poster from structured inputs, translating dramatic semantics into a coherent visual design suitable for a theatrical production.

Input: You will be given three components. (1) A semantic JSON: "{visual_constraints_json}", containing style keywords, color palette, and composition rules. (2) A visual subject description: "{visual_subject_txt}", defining the abstract visual focus. (3) A text content block: "{text_content_txt}", including the full title, performance information, venue, director, cast, and production credits.

Generation Requirements:

- **Abstraction Constraint:** The poster must use symbolic, non-realistic visual language. Do not generate literal scenes, realistic characters, or detailed narrative depictions. All subjects must be abstract (e.g., silhouettes, geometric forms, fragmented shapes, or light-based structures).
- **Color System:** Strictly follow the color palette defined in the semantic JSON. Use soft gradients or painterly blending. The color scheme must reflect the emotional tone and reinforce the thematic keywords.
- **Composition and Layout:** Follow the composition rules from the JSON. Maintain a balance between visual tension and negative space. The visual subject must be integrated into the composition rather than dominating it as a literal object.
- **Background:** The background must be atmospheric and non-realistic, avoiding detailed environments.
- **Typography (Mandatory):** All text from "{text_content_txt}" must be included exactly, with no missing, altered, or additional words. The title must be prominently placed (top or center). Supporting text must follow a clear hierarchy (title → information → credits). Typography style must align with the theatrical tone.

Output: Generate a single image prompt (in plain text format, suitable for direct use in a text-to-image model). The prompt must clearly describe the abstract poster, including color system, composition, background, visual subject, and complete text layout, ensuring both strong visual aesthetics and strict textual accuracy.

Figure 3: Prompt for Poster Synthesis.

Table 1: Quantitative tracking of the iterative refinement process using VLM-as-a-Judge. The table presents the multi-dimensional scores of the intermediate images at each iteration step (K), demonstrating the progressive enhancement of visual and semantic quality.

Iteration	Theatrical Aesthetics			Communication Efficacy		Thematic Fidelity			Overall
	Abstraction	Tension	Refinement	Integration	Clarity	Resonance	Aptness	Contextuality	
$K = 0$	4.85	5.12	5.90	6.15	6.30	5.20	5.45	5.10	5.51
$K = 1$	5.60	6.35	6.20	6.45	6.15	6.05	6.25	5.95	6.13
$K = 2$	6.85	6.90	7.15	7.05	7.20	6.45	6.80	6.75	6.89
$K = 3$ **(Final)**	**7.82**	**8.05**	**8.15**	**7.95**	**8.20**	**7.88**	**7.92**	**8.10**	**8.01**

A.4 Implementation of Interactive Multimodal Character AI

To transition the static publication into an engaging multimodal experience, we developed a persona-grounded interactive module. This system orchestrates large language models, video diffusion, and custom speech synthesis to bring theatrical characters to life through a two-tiered interaction paradigm. Visual demonstrations of this interactive module and its corresponding user interfaces are presented in Figure 5 and Figure 6.

In the first tier, we establish the character's initial dynamic presence. We utilize **HunyuanVideo** to animate the static character portrait, generating a full-body action sequence that captures postures and movements strictly consistent with the character's specific temperament. Concurrently, we deploy **Qwen3-TTS** to clone voice characteristics derived from the stage directions and

Prompt for Style Extraction

You are an expert Art Director and Visual Analyst specializing in theatrical design. Your task is to analyze the provided theatrical poster and extract its core aesthetic attributes to establish a unified visual identity for a broader design project.

Your extraction must strictly adhere to the following constraints:

- **Focus EXCLUSIVELY** on high-level abstract visual semantics: the overall stylistic paradigm (e.g., retro, minimalist, surrealist, cyberpunk), lighting atmosphere (e.g., moody chiaroscuro, vibrant neon), and core color palette (e.g., monochromatic blue, high-contrast red and black).
- **Deliberately IGNORE** and filter out all instance-level details, specific character features, localized props, text, and layout structures.
- Condense your extracted aesthetic priors into a series of concise, highly descriptive phrases.

Output the result strictly as a JSON array of strings representing the global style vector. Do not include any conversational filler.

Example Output:
["vintage grunge aesthetic", "high-contrast sepia and crimson palette", "heavy grain texture", "dramatic low-key lighting"]

Figure 4: Prompt for Style Extraction.

generate a signature audio slogan. Within the electronic theater program, interacting with the portrait triggers the synchronized playback of this animation and the customized slogan, establishing an immediate emotional and visual connection with the user.

In the second tier, we enable deep, real-time narrative exploration through a voice-enabled conversational interface. We implement a Retrieval-Augmented Generation (RAG) framework, utilizing the source script as a localized knowledge base. This allows the persona-grounded LLM to accurately answer user queries regarding the character's background, biography, and plot events from a first-person perspective. To ensure low-latency engagement, the chat functionality operates via a streaming architecture, where the LLM's text output is piped directly into the Qwen3-TTS engine for real-time audio synthesis. During this conversational phase, the visual interface deliberately reverts to the high-fidelity static portrait. This design choice strictly prioritizes real-time auditory delivery and interactive fluidity by eliminating the substantial computational overhead of continuous video generation.

For the evaluation concern, we conduct comprehensive quantitative assessments, with results presented in Table 2: ***(1) Visual:*** *CLIP-I (CI)* and *ArcFace (AF)* assess similarity between static images and animations. ***(2) Textual:*** GPT-4 evaluates *Persona Consistency (PC)* and *Factual Accuracy (FA)*. ***(3) User Interaction:*** 15 testers rated *Voice-Persona Fit (VF)*, *Dialogue Naturalness (DN)*, and *Interactive Immersion (II)*. Both GPT-4 and user ratings use a 1-5 scale.

A.5 Resource Consumption

ETP generation is an ***offline professional design*** task, not a high-throughput scenario. Our framework runs on a single NVIDIA RTX A6000 (48GB). The peak VRAM usage is ***38-42GB*** in FP16 precision. For real-world deployment, this automated process is highly practical, replacing days of manual design labor.

Table 2: Quantitative evaluation of the interactive module.

Method	Visual		Textual		Interaction		
	CI ↑	AF ↑	PC ↑	FA ↑	VF ↑	DN ↑	II ↑
Ours	0.85	0.78	4.56	4.82	4.35	4.60	4.72

B Additional Results

B.1 Qualitative Results

To further demonstrate the robustness, versatility, and broad applicability of ETPDesigner, we conducted extensive qualitative experiments across a diverse spectrum of theatrical genres and narrative themes. By testing our framework on texts ranging from classical tragedies to contemporary comedies, we systematically evaluated its capacity to handle complex visual metaphors and varying emotional undertones. Extended qualitative results and comprehensive visual comparisons are presented in Figure 7 and Figure 8.

Visual inspections reveal a fundamental divergence in aesthetic paradigms. While baseline outputs lean towards stylized cinematic realism, they frequently lack the structural negative space required for typographic integration and resemble disjointed promotional materials rather than a unified theatrical cast. In contrast, ETPDesigner consistently synthesizes abstract thematic posters, backgrounds with optimal functional void, and character portraits exhibiting rigorous stage fidelity. By emulating authentic stage photography, our framework ensures that all visual assets seamlessly integrate into a cohesive electronic theater program.

Figure 5: Temporal progression of the dynamic character animation. The sequence illustrates the fluid transition from a generated static portrait to a dynamic video, displaying consecutive frames sampled at $t = 1$s to $t = 5$s.

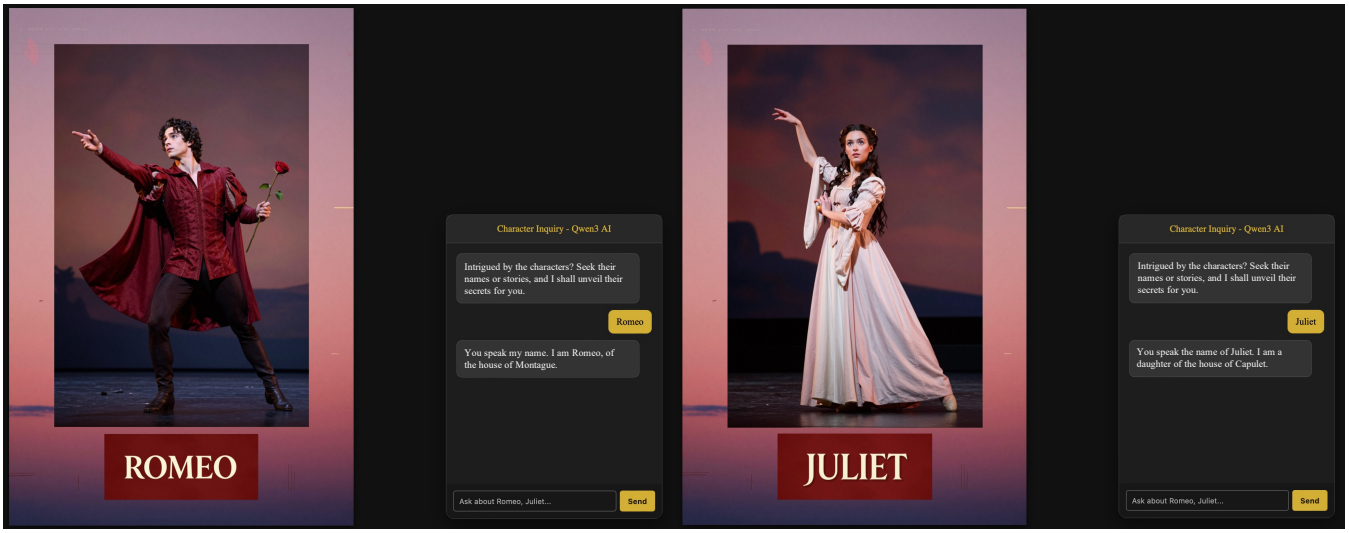


Figure 6: The interactive conversational interface. This module enables users to engage in real-time, voice-enabled narrative exploration with persona-grounded theatrical characters.

B.2 Quantitative Results

While both stages utilize a VLM, they serve distinct roles. The internal critic identifies visual flaws by comparing outputs with retrieved reference images, whereas the evaluation VLM applies a fixed, unified scoring rubric to all methods. To further rule out potential same-model bias, we also re-evaluate all results using GPT-4o and Gemini-1.5-Pro. As shown in Table 3, these findings consistently align with our initial evaluation and the user study.

C User Study Setup

To further validate the practical effectiveness and subjective superiority of our framework, we conducted a comprehensive user evaluation. We formulated a rigorously structured questionnaire, instructing participants to evaluate and compare the generation results produced by ETPDesigner against those from state-of-the-art baseline models. To ensure a thorough and unbiased assessment, the evaluation protocol was meticulously categorized into four core dimensions. First, we evaluated the *Visual Appeal* of the theatrical posters, focusing on overarching aesthetic quality and artistic tension. Second, we assessed the *Layout Functionality* of the background images, specifically measuring their structural suitability and negative space distribution for typographic integration. Third, we examined the *Overall Consistency* of the character portraits, gauging their stylistic uniformity, visual harmony, and stage fidelity across different character assets. Finally, we measured the

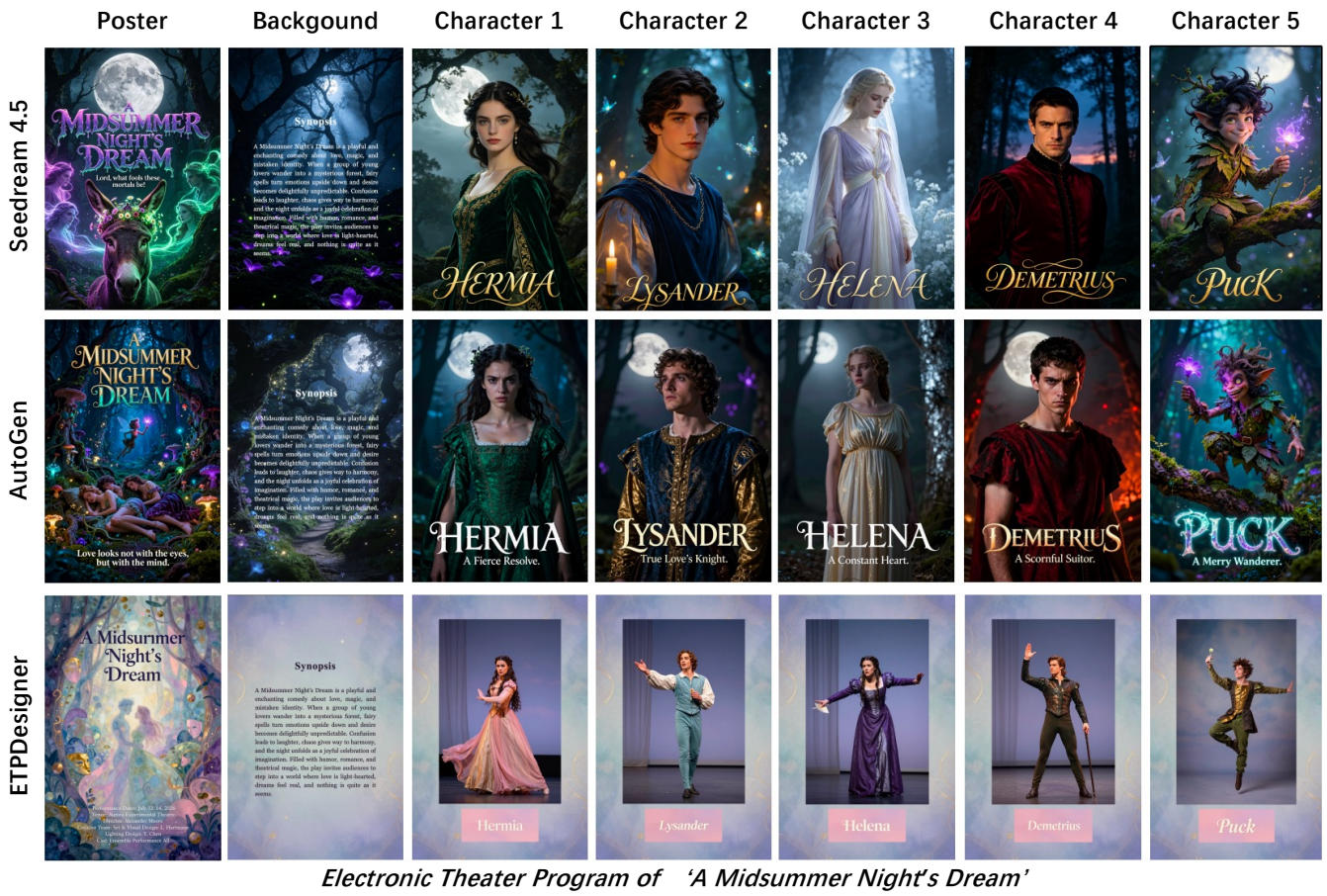


Figure 7: Qualitative comparison of generation results for A Midsummer Night's Dream across ETPDesigner, Seedream 4.5, and AutoGen.

Table 3: Evaluation results using GPT-4o and Gemini-1.5-Pro.

Method	GPT-4o					Gemini-1.5-Pro				
	TA ↑	CE ↑	TF ↑	EX ↑	WS ↑	TA ↑	CE ↑	TF ↑	EX ↑	WS ↑
Nano Banana	5.86	6.12	5.05	3.88	1.60	5.95	5.98	4.92	4.02	1.52
Seedream	6.08	6.40	5.42	3.75	1.80	6.18	6.25	5.26	3.89	1.68
AutoGen	6.12	6.54	5.58	2.15	2.76	5.98	6.42	5.70	2.24	2.91
Ours	**7.68**	**7.85**	**7.59**	**8.36**	**8.24**	**7.81**	**7.72**	**7.71**	**8.52**	**8.12**

Thematic Relevance of the complete electronic theater program to determine its comprehensive semantic alignment with the underlying dramatic script. For each of these dimensions, we developed a set of comprehensive, fine-grained questions to carefully guide the participants' subjective scoring process. The specific evaluation criteria and the detailed structural design of the survey instrument are visually presented in Figures 9, 10, 11, and 12.

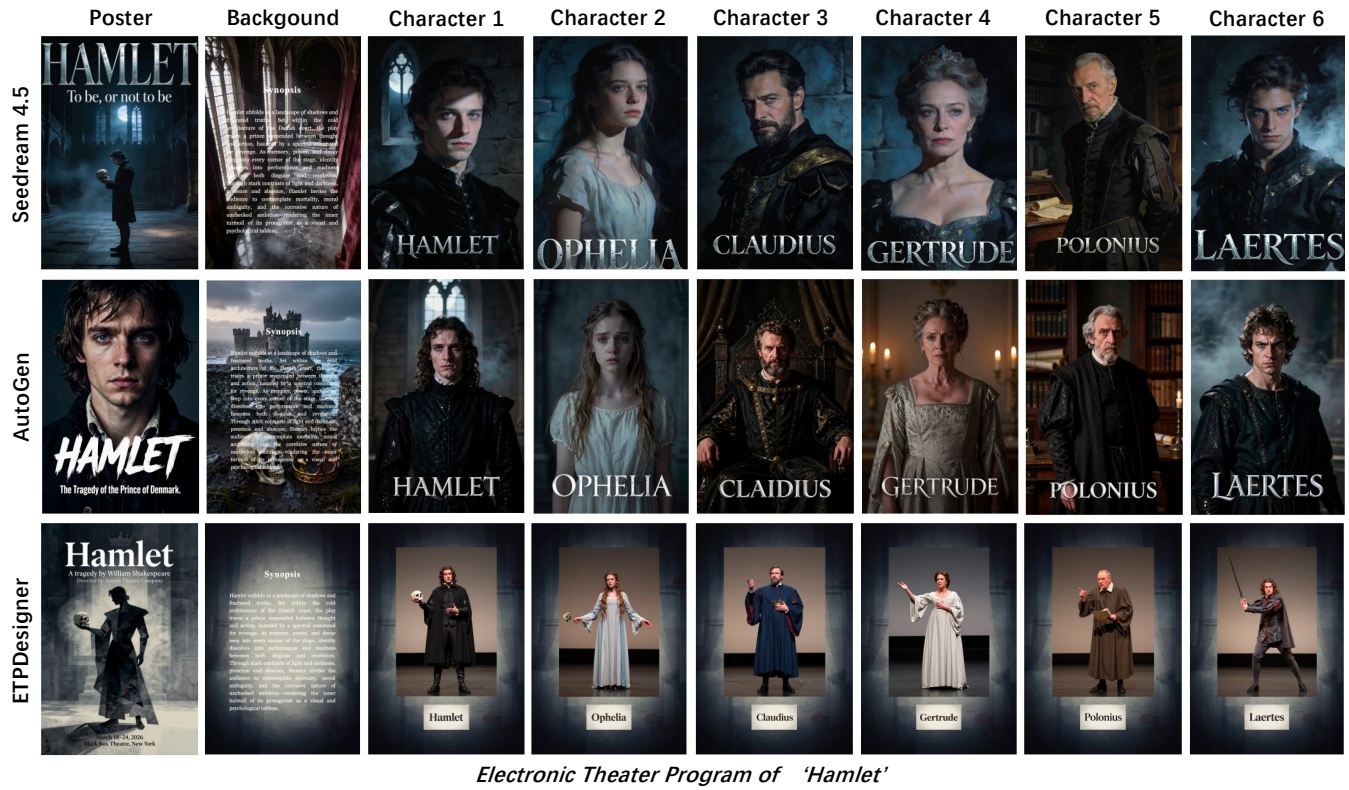


Figure 8: Qualitative comparison of generation results for Hamlet across ETPDesigner, Seedream 4.5, and AutoGen.

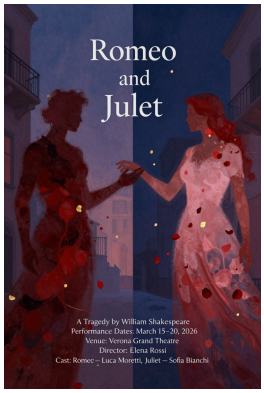
User Study for ETPs

Posters for theater programs must align with the play' s theme and possess a certain degree of theatrical abstraction and artistic quality.

Based on the plot summary below and the criteria in the prompt, select the method that produces the highest-quality result from the three options provided.

The play for this task is Romeo and Juliet.
Plot Summary: The Montagues and the Capulets are two families who have been feuding for generations. Romeo and Juliet fall in love at first sight and marry in secret. Romeo is banished for killing Juliet' s cousin, Tybalt. To escape an arranged marriage, Juliet takes a potion to feign death; Romeo, believing her to be dead, drinks poison at her grave and kills himself. When Juliet awakens, she follows him in death, and the two families are ultimately reconciled by this tragedy.

A



B

C

Q1: Which of the following best aligns with the play' s theme?

Option *A* | Option *B* | Option *C*

Q2: Which of the following best exemplifies dramatic abstraction?

Option *A* | Option *B* | Option *C*

Q3: Which of the following offers the greatest artistic appeal?

Option *A* | Option *B* | Option *C*

Q4: Which of the following is of the highest overall quality?

Option *A* | Option *B* | Option *C*

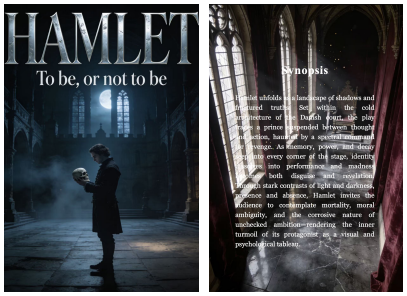
Figure 9: Survey instrument for Visual Appeal: Evaluating the aesthetic quality and artistic tension of the generated posters.

User Study for ETPs

The background image for the synopsis in the theater program must be consistent in style with the poster, leave appropriate white space for the text, and possess a certain degree of artistic quality.

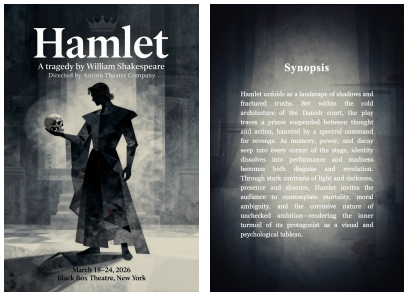
Based on the plot summary below and the criteria in the prompt, select the method that produces the highest-quality result from the three options provided.

The script for this task is Hamlet.
Plot Summary: Prince Hamlet of Denmark learns that his uncle Claudius murdered his brother to seize the throne and married his mother as queen. He feigns madness to seek revenge, using a play-within-a-play to expose the crime, but accidentally kills the minister Polonius and is forced into exile. Ultimately, in a duel, both he and Laertes are poisoned by their swords, and the queen mistakenly drinks the poisoned wine. Before his death, Hamlet kills his uncle, bringing the tragedy to a close.



A

B



C

Q1: Which of the following best matches the style of the poster?

Option *A* | Option *B* | Option *C*

Q2: Which of the following best utilizes negative space?

Option *A* | Option *B* | Option *C*

Q3: Which of the following is the most visually appealing?

Option *A* | Option *B* | Option *C*

Q4: Which of the following is of the highest overall quality?

Option *A* | Option *B* | Option *C*

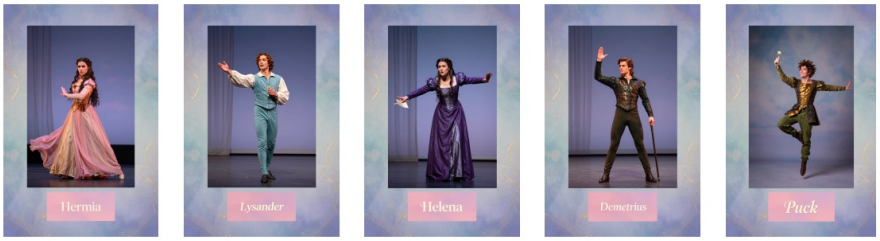
Figure 10: Survey instrument for Layout Functionality: Measuring the structural suitability and negative space of background.

User Study for ETPs

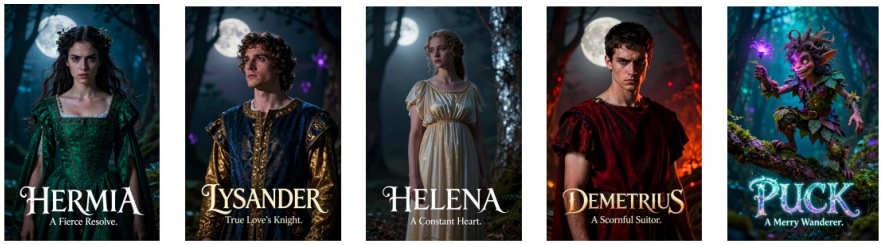
A series of character portraits for a theater program must meet the requirements of distinct character traits, theatrical exaggeration, and consistency among the characters.

Based on the plot summary below and the criteria in the prompt, select the method that produces the highest-quality result from the three options provided.

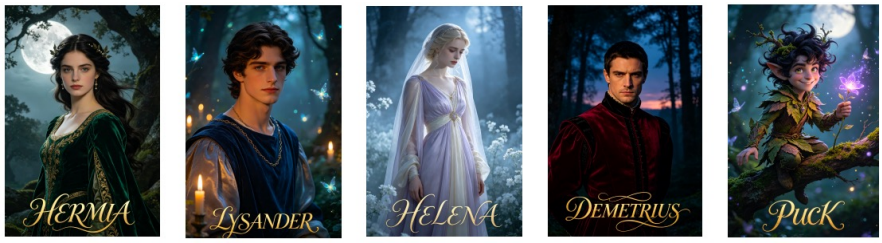
The script used for this task is A Midsummer Night's Dream.
Plot Summary: Hermia defies her father' s orders and elopes with Lysander; Helena, who is in love with Demetrius, follows closely behind. To tease the Queen of the Fairies, the King of the Fairies orders the fairy Puck to use "magic potion" to mix up the pairings, resulting in a chaotic love quadrangle. Eventually, the spell is broken, lovers are united, and the Queen reconciles with the King. The play concludes with a lively wedding, filled with humor and fantasy.



A



B



C

Q1: Which of the following demonstrates the greatest unity?

Option *A* | Option *B* | Option *C*

Q2: Which of the following characters' actions exhibits the most dramatic exaggeration?

Option *A* | Option *B* | Option *C*

Q3: Which of the following is the most realistic?

Option *A* | Option *B* | Option *C*

Q4: Which of the following is of the highest overall quality?

Option *A* | Option *B* | Option *C*

Figure 11: Survey instrument for Overall Consistency: Gauging stylistic uniformity and stage fidelity across character portraits.

User Study for ETPs

Finally, a complete theater program should exhibit overall stylistic consistency, artistic appeal, and theatrical exaggeration.

Based on the plot summary below and the criteria in the prompt, select the method that produces the highest-quality result from the three options provided.

The play for this task is Romeo and Juliet.
Plot Summary: The Montagues and the Capulets are two families who have been feuding for generations. Romeo and Juliet fall in love at first sight and marry in secret. Romeo is banished for killing Juliet' s cousin, Tybalt. To escape an arranged marriage, Juliet takes a potion to feign death; Romeo, believing her to be dead, drinks poison at her grave and kills himself. When Juliet awakens, she follows him in death, and the two families are ultimately reconciled by this tragedy.

A

B

C

Q1: Which of the following best demonstrates overall stylistic consistency?

Option *A* | Option *B* | Option *C*

Q2: Which of the following best demonstrates dramatic exaggeration?

Option *A* | Option *B* | Option *C*

Q3: Which of the following is most aesthetically pleasing?

Option *A* | Option *B* | Option *C*

Q4: Which of the following is of the highest overall quality?

Option *A* | Option *B* | Option *C*

Figure 12: Survey instrument for Thematic Relevance: Assessing the semantic alignment of the program with the script.