Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai

Article

Verse-in-Wine: A Generative AI Framework for Chinese Calligraphy Painting with Drinking Culture

Ronghua Cai and James She *

The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China
* Correspondence: jamespmshe@hkust-gz.edu.cn

How To Cite: Cai, R.; She, J. Verse-in-Wine: A Generative AI Framework for Chinese Calligraphy Painting with Drinking Culture. *Transactions on Artificial Intelligence* **2025**, *1*(1), 248–264. https://doi.org/10.53941/tai.2025.100017

Received: 29 August 2025
Revised: 17 September 2025
Accepted: 9 October 2025
Published: 21 OctOber 2025

Abstract: This paper presents *Verse-in-Wine* , a generative framework that integrates Chinese classical poetry, traditional wine culture, and calligraphy painting through large language models (LLMs) and visual generation. Given user-selected intention keywords from culturally grounded categories, the system recommends poetic lines, maps them to symbolic wines and historical calligraphy styles, and synthesizes visually coherent outputs. A fully functional prototype was developed and evaluated through both automated and user studies. LLM-based evaluation across 300 samples achieved an overall score of 0.9165, while a user study with 100 samples yielded a comparable human rating of 0.8900, confirming both the system's cultural fidelity and usability. The framework demonstrates how generative AI can meaningfully engage with heritage aesthetics, linking related cultures for artistic expression.

Keywords: AI art; calligraphy painting; poetry; drinking; culture

1. Introduction

Chinese poetry, calligraphy, painting, and drinking culture represent distinct yet deeply intertwined threads in classical Chinese aesthetics. Across dynasties, literati and artists cultivated unique ways of expressing emotion, philosophy, and social rituals by composing verses while sipping wine and inscribing them in flowing brushwork. These cultural forms function as multimodal vehicles encoding moods, ideals, and worldviews. This longstanding aesthetic convergence is vividly reflected in historical artworks [1,2], which document the convivial, poetic lives of scholar-artists (Figure 1).

Figure 1. (**Left**): Classical Chinese paintings exemplifying the integration of poetry, calligraphy, and drinking culture. *Scholar Drinking Alone* by Chen Hongshou [1] , and a scene from *Eight Immortals Drinking* [2]; (**Right**): Our generated result.

Despite their historical synergy, they are often treated today as isolated artifacts—segmented across museums, literary anthologies, and commercial craft. Meanwhile, the rise of large-scale generative AI models offers new possibilities for cultural synthesis and digital co-creation [3]. Text-to-image (T2I) models [4] can evoke painterly

aesthetics, large language models (LLMs) [5] can emulate classical poetic structures, and neural stylization models [6] can simulate calligraphic styles. Yet, applying these technologies to culturally grounded, user-facing, multimodal experiences presents new challenges [7,8]. How can we ensure that generated poems align with the affective and symbolic intent of the user? How can AI-synthesized visuals preserve the aesthetic principles of classical Chinese painting and calligraphy? And how can different cultural modalities, poetry, wine, painting, and script, be recombined meaningfully in the digital realm? To explore these questions, This paper presents *Verse-in-Wine*, an interactive AIGC approach that reanimates the expressive ties between Chinese poetry, calligraphy and wine culture. Inspired by literati rituals and aesthetics, the approach guides users through a generative journey from selecting emotional or symbolic intentions to receiving AI-curated classical poems and ultimately generating synthesized calligraphy painting artworks. These creations are infused with user-directed symbolism, such as specific wine types or poetic moods, rendered through multimodal generation. The generated results are evaluated through automated alignment metrics and user studies, demonstrating how generative AI can mediate between tradition and innovation, fostering new forms of cultural engagement and co-creation.

The paper's core contributions are as follows: (1) It proposes *Verse-in-Wine*, a novel AIGC platform that connects classical Chinese poetry, wine culture, and calligraphy through a unified, user-interactive pipeline. (2) It develops a layout-aware calligraphy composition method based on YOLOv8, which significantly improves spatial aesthetics and overcomes limitations of heuristic placement. (3) It designs a culturally grounded user experience pipeline, from intention selection to multimodal output, and evaluates both content alignment and usability through automated and user-centric experiments.

2. Related Work

To situate this work within the broader research landscape, this section reviews related work in three areas: AI for cultural experience, AI for calligraphy generation, and AI for calligraphy painting generation.

2.1. AI for Cultural Experience

Generative AI is increasingly applied to cultural heritage for purposes of preservation, interpretation, and public engagement. Digital restoration of ancient manuscripts and inscriptions has seen considerable progress using vision-based techniques [9, 10]. Additionally, 3D reconstruction from sparse visual input allows for the recovery of cultural artifacts [11]. MLLMs have also been employed to analyze artworks such as paintings, offering multimodal insights into cultural content [12]. AI systems also enhance interactive engagement with cultural heritage. XR-based interfaces [13] and MLLM-powered chatbots [14] enable users to explore historical material in context-aware, conversational ways. Generative approaches extend to artistic recreation, including the synthesis of Chinese landscape painting [15], and the reinterpretation of intangible heritage elements such as wine culture and poetic symbolism. Yet, cultural alignment remains a critical challenge. Models must reflect historically grounded meanings—such as the symbolic role of wine in Chinese literati culture—while also preserving stylistic authenticity. Methods like layout-aware text-to-image generation [5] offer promising directions for improving the semantic and aesthetic coherence of such works.

2.2. AI for Calligraphy Generation

Chinese calligraphy generation has primarily followed two methodological pathways: style transfer from standard glyphs and skeleton-based rendering.

Style Transfer from Standard Glyphs . This approach regards calligraphy generation as an image-to-image translation task. Models such as Pix2Pix [16], CycleGAN [17], and U-GAT-IT [18] have been widely used for font style transfer. The *zi2zi* model [19] adapts Pix2Pix for Chinese font translation, while CycleGAN enables training with unpaired samples, extended by Chang et al. [20] for handwritten character synthesis. U-GAT-IT inspired the *MaLiang* model [21], which incorporated emotional conditioning. To reduce the cost of paired data, Zhou et al. [22] proposed an end-to-end calligraphy model that can utilize unpaired datasets. In more recent works, diffusion models have also been widely adopted as the backbone for style transfer, achieving state-of-the-art visual fidelity. For example, FontDiffuser [6] demonstrates high-quality stylized character synthesis, and *AnyText* [23] extends diffusion to text-controlled font generation, though challenges remain with character complexity and layout consistency.

Skeleton-Based Rendering. Another line of work first generates the character skeleton, followed by style rendering on the generated structure. Skeleton extraction can be achieved by curve fitting [24,25] or convolutional networks [26,27]. The *SCFont* model [28] exemplifies this two-stage design, combining skeleton transformation and rendering modules, with its foundation in earlier pipelines such as Lian et al. [29].

2.3. AI for Calligraphy Painting Generation

The generation of Chinese calligraphy painting artworks, which integrate calligraphy and painting within a single visual composition, represents an emerging and relatively underexplored area of research. Such works enable the fusion of textual expression with visual storytelling, providing a rich medium for conveying cultural themes such as drinking culture and poetic symbolism. One representative study is Polaca [30], which introduced a system for generating poetic Chinese landscape paintings incorporating calligraphy. While effective in synthesizing landscape imagery guided by textual input, the approach faces several limitations: it depends on poems containing concrete visual elements to guide image generation, the spatial integration of calligraphy with the background lacks aesthetic coherence, and the diversity of calligraphy styles remains limited. Subsequent work [31] sought to improve the visual integration by employing more abstract backgrounds, achieving better harmony between calligraphy and image but sacrificing detailed scenic representation. The generated result is shown in Figure 2. Overall, Chinese calligraphy painting generation continues to present open challenges in achieving flexible content alignment, stylistic diversity, and culturally resonant composition, making it a promising direction for further research in multimodal AI art generation.

Figure 2. Generated results of related calligraphy painting generation work. (**Left**): [31]; (**Right**): [30].

3. Proposed Verse-in-Wine Framework

The *Verse-in-Wine* platform is a generative AI system that integrates multimodal components for synthesizing artworks inspired by Chinese poetry, calligraphy, and drinking culture. As shown in Figure 3, the platform comprises an interactive frontend and a backend pipeline driven by language models, diffusion generation, and vision models.

3.1. Wine-Themed Poem Recommendation

The poem generation process begins with user intention selection, where up to five semantic keywords are chosen from six culturally grounded categories (Table 1). These intention terms reflect poetic moods, emotional states, philosophical ideals, and symbolic imagery commonly found in classical Chinese literature. The six-category system draws inspiration from traditional literary frameworks such as Sikong Tu's Twenty-Four Poetic Styles, aiming to reflect the rich semantic layers of classical poetry beyond basic sentiment analysis.

This classification enables users to express nuanced aesthetic goals in a way that aligns with historical poetic conventions. Categories like philosophical concept and literati ideal help direct the LLM toward generating verses that embody traditional values, imagery, and tone. By anchoring the input in culturally meaningful dimensions, the system enhances both the emotional coherence and thematic relevance of the generated content. The selected keywords are then embedded into a structured prompt designed for a LLM, guiding it to retrieve thematically appropriate poetic content centered on traditional wine culture.

To ensure semantic accuracy and generation quality, the prompt is structured into four logical components, referred to as Prompt Sections (PS): the main task directive (PS1), which clarifies the goal of retrieving real, verifiable classical lines related to wine and the input keywords; user input embedding (PS2), which explicitly inserts the selected intention terms; example injection (PS3), which illustrates a valid output with matching logic; and output format constraints (PS4), which enforce consistency and JSON readability. This modular design provides both interpretability and control, allowing the model to focus on cultural grounding while preserving diversity in retrieval. An overview of the four-section prompt structure is illustrated in Figure 4. The full prompt content is provided in Supplementary Materials.

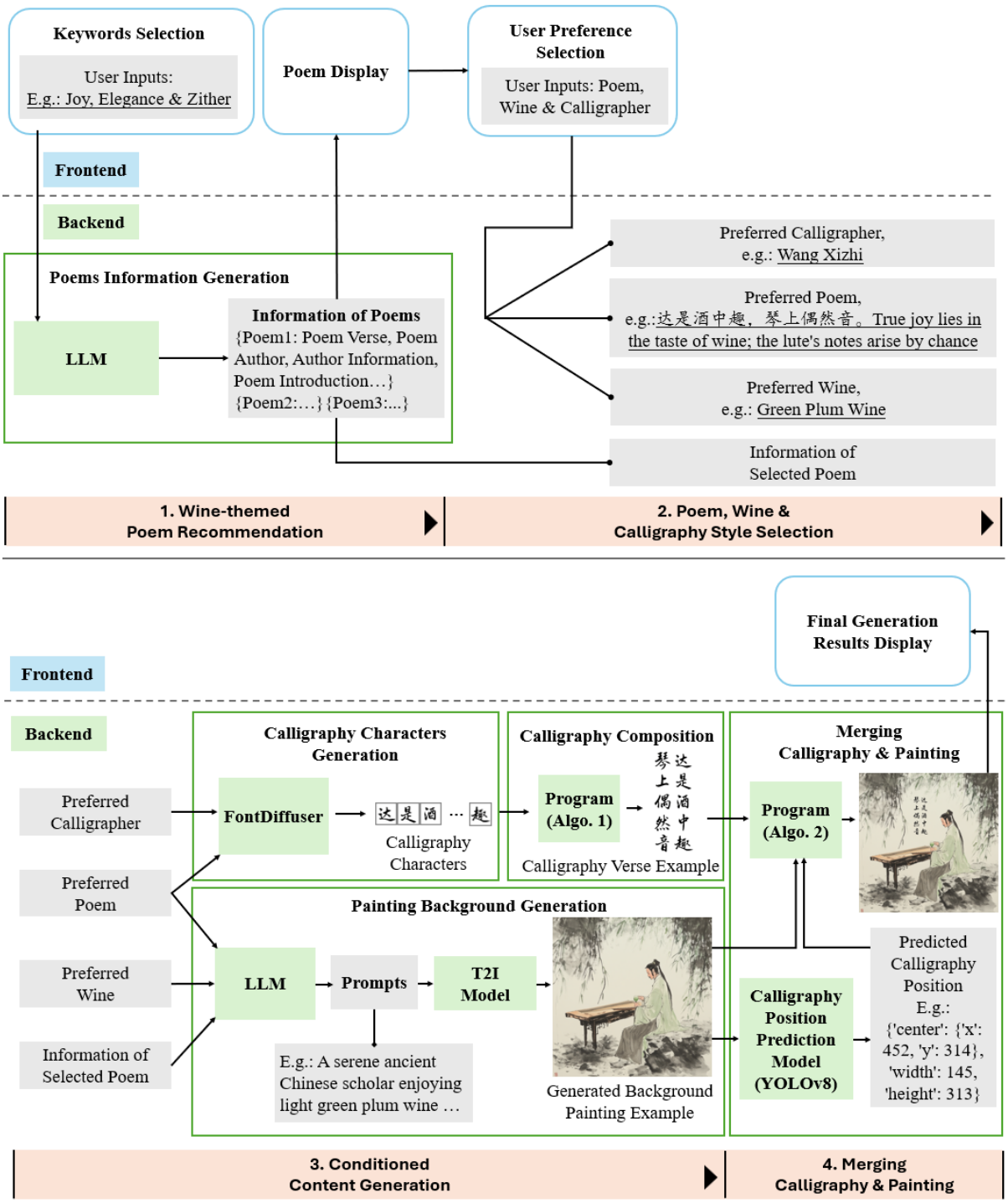


Figure 3. The proposed *Verse-in-Wine* framework.

Table 1. User inputs for wine-themed poem recommendation.

Category	Description	Example Options
Basic Emotions	Fundamental human emotions mapped to poetic tones	Joy, Surprise, Calm, Fear, Anger, Sadness
Poetic Mood	Derived from Sikong Tu's twenty-four poetic styles	Lofty Antiquity, Elegance, Graceful Ease, Composed Depth, Natural, Heroic Sorrow
Life Phase	Biographical states often depicted in poetry	Youth, Ambition, Frustration, Wandering, Retreat, Old Age
Philosophical Concept	Abstract ideas from Confucianism, Daoism, Buddhism	Dao, Impermanence, Emptiness, Harmony, Cycle of Life
Literati Ideal	Intellectual and emotional stances of classical scholars	Reclusion, Wine and Poetry, Nostalgia, Spontaneity, Recklessness
Cultural Symbol	Iconic visual-poetic imagery	Moon, Wine Goblet, Incense, Zither, Plum Blossom, Bamboo Shadow

PS1

```
You are asked to perform a controlled retrieval of classical Chinese
poetic lines centered on the theme of wine culture,... Below is a list of
intention keywords provided by the user, [Selected Keywords, E.g.: Joy,
Elegance, Zither].You should use them to guide your poem retrieval,
ensuring that the selected verses reflect these themes...
Your task is to:
Select at least three distinct poetic lines from real, historically
verifiable Chinese poems that align with both the wine-related theme and
the embedded user intentions...
```

PS4

PS2

```
Below is an example of a valid response in JSON format based on the user's
intentions: Wine & Poetry, Reclusion, and Moon.
{
  "poem_line": "A pot of wine among the flowers, I drink alone with no kin
near.",
  "author": "Li Bai (Tang Dynasty)",
  "author_intro": "A romantic poet of the Tang Dynasty, honored as the
'Immortal of Poetry',...
  "poem_intro":...
}
```

PS3

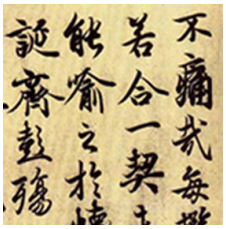
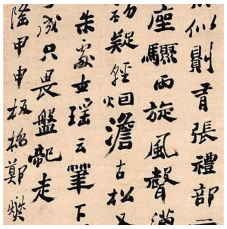
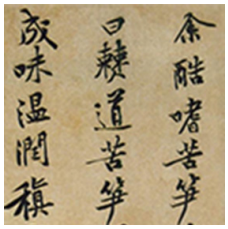
```
Each result must be output strictly in a JSON array format using the
following structure:
[
  {
    "poem_line": "A wine-related line from a real classical Chinese poem",
    "author": "Name and dynasty of the poet",
    "author_intro": "Brief biography of the poet",
    "poem_intro": "Short interpretation of the poem",
    ...
  },
  ...
]
```

Figure 4. Structured prompts for wine-themed poem recommendation.

3.2. Poem, Wine & Calligraphy Style Selection

After receiving three candidate poems generated based on their selected keywords, users are able to choose one poetic line they prefer. To further personalize the output, they also select one wine type and one calligraphy style, which function as semantic anchors for downstream generation. The prototype provides five wine types and five calligraphers, each associated with distinct emotions, visual features, and cultural meanings. Wine options vary in taste, aroma, and symbolic context—from the nostalgic depth of Yellow Wine to the floral elegance of Osmanthus Wine. The presence of wine vessels in the generated scene is not mandatory; rather, the wine type primarily influences the color palette, drinking mood, and artistic atmosphere of the output. The selected calligraphers represent diverse aesthetic temperaments, such as Wang Xizhi's refined semi-cursive and Li Qingzhao's lyrical grace. All options are presented with illustrations and descriptive attributes (Table 2), enabling users to make intuitive and culturally meaningful choices. The preference module is modular and extensible, allowing new wine or calligraphy styles to be added without affecting the core generation pipeline. This design supports future expansion toward more personalized and culturally diverse creative experiences.

Table 2. User-selectable cultural elements: (a) wine types and (b) calligraphy styles.

(a)					
Wine	**Yellow Wine**	**Osmanthus Wine**	**Green Plum Wine**	**Distilled Liquor**	**Rice Wine**
Aroma	Warm and rich with a hint of soy aroma	Fragrant osmanthus aroma	Fresh plum scent with a sweet-sour balance	Strong and pungent	Mild rice fragrance
Color	Amber or dark brown	Golden or light yellow	Clear light green	Colorless and clear	Milky white or clear
Taste	Sweet and mildly sour, smooth and mellow	Sweet and floral, smooth to the palate	Refreshing and tangy, long aftertaste	Strong but smooth with lasting impact	Soft and mildly sweet
Description	One of the oldest Chinese brewed wines, favored by poets to express emotions.	A floral rice wine often enjoyed during Mid-Autumn or poetic gatherings.	A summer wine used to relieve heat and inspire poetry in ancient times.	Distilled with ancient techniques, often drunk by warriors and bold men.	A family wine symbolizing reunion, often served at friendly banquets.
(b)					
	Wang Xizhi ♂	**Li Qingzhao** ♀	**Zheng Banqiao** ♂	**Su Shi** ♂	**Huang Tingjian** ♂
Description	The “Sage of Calligraphy”, known for fluid and lively semi-cursive script.	Renowned lyrical poet whose calligraphic style evokes elegance, introspection, and feminine grace.	Eccentric Qing-dynasty calligrapher known for bold clerical–semi-cursive style.	Literati-style calligraphy, unrestrained and expressive like his poetry.	Combines semi-cursive and cursive with bold structure and rhythmic expression.

3.3. Conditioned Content Generation

The painting generation process begins by transforming the selected poem line and wine profile into a culturally grounded visual scene. To achieve this, we construct a carefully designed prompt for a large language model, which interprets the poetic semantics and wine symbolism into concrete visual cues. Instead of generating literal imagery, the prompt guides the model to translate abstract poetic expressions into symbolic and atmospheric elements—such as seasonal context, emotional tone, and object arrangement. The poem line and wine type are embedded as key conditioning variables, while the model is instructed to produce detailed descriptions of the imagined scene, emphasizing metaphorical alignment and mood consistency. Due to space constraints, the full version of the structured prompt similar with Figure 4 is provided in the Supplementary Materials.

To render the calligraphy, we first extract the set of unique Chinese characters from the selected poem line, excluding punctuation and duplicate glyphs. These are individually synthesized by FontDiffuser, a few-shot generative model conditioned on reference images from the chosen calligrapher. The resulting character-level outputs preserve personalized brushstroke traits and stylistic fidelity. To reconstruct the full poetic line, our system reinserts duplicated characters, determines optimal line breaks using punctuation-based heuristics, and arranges the characters into a balanced vertical layout. Characters are stacked top-to-bottom within each column and arranged right-to-left to emulate traditional Chinese script presentation. Finally, all columns are merged into a single

transparent canvas. The complete character composition procedure is formally defined in Algorithm 1.

Algorithm 1 Calligraphy Generation and Composition

Require: Poem line T, Calligrapher ID C_{id},
FontDiffuser model FD
Ensure: Final composed calligraphy image $I_{composed}$
Extract unique Chinese characters U_{chars} from T
Initialize character image map D_{char_img}
Retrieve style reference image R_{style} for C_{id}
for all character $c \in U_{chars}$ **do**
  Generate $G_c \leftarrow FD(c, R_{style})$
  Apply alpha mask to G_c for transparent background
  Store masked G_c in $D_{char\_img}[c]$
end for
Initialize image sequence S_{images}
for all character c in original T **do**
  if c is Chinese **then**
    Append $D_{char\_img}[c]$ to S_{images}
  end if
end for
Determine line splits $L = \{L_1, L_2, \dots\}$ for S_{images} based on punctuation in T, balancing column heights
Initialize list of column images $C_{columns}$
for all line L_i in L **do**
  Create column image C_i by vertically stacking images in L_i, centering characters horizontally within the column
  Append C_i to $C_{columns}$
end for
Arrange column images $C_{columns}$ onto final canvas $I_{composed}$ from right-to-left
return $I_{composed}$

3.4. Merging Calligraphy & Painting

After generating the calligraphy image and Midjourney background, the final step involves intelligently compositing them into a visually harmonious whole. To identify appropriate placement regions within the generated painting, we adopt a vision-based layout prediction approach using the YOLOv8 object detection framework. To train this model, we manually annotated 100 synthesized images with one or more bounding boxes indicating ideal positions for calligraphy placement. The annotations follow a strict set of aesthetic guidelines (Figure 5), including avoiding visual focal points (e.g., faces, wine vessels), prioritizing low-density regions (e.g., mist, blank sky), and preserving compositional balance. Each annotation file adopts the standard YOLO format, specifying normalized bounding box coordinates and a single class label (0) for all regions. The training data pairs each image with its corresponding label file, as shown in the lower panel of Figure 5.

At inference time, the trained YOLOv8 model takes a Midjourney-generated image as input and predicts one or more bounding boxes indicating candidate regions for calligraphy placement. To ensure legibility and visual balance, we further evaluate the average grayscale value of the predicted region. If the region is found to be overly dark (i.e., below a luminance threshold $T_{\text{dark}} = 0.4$), the system dynamically switches the calligraphy color from black to white to enhance contrast against the background. Once the optimal region is selected, the vertical calligraphy image is resized to match the predicted box width and centered within the bounding box. The calligraphy is then composited onto the background using alpha blending while preserving brushstroke texture and spacing integrity. Since the system generates four unique background images based on the selected wine and poetic context, this layout and fusion process is performed four times—once for each visual background. The complete procedure is summarized in Algorithm 2.

The final four compositions are presented to the user through the frontend interface, each showcasing a unique visual combination of poem, background painting, and calligraphy layout. Alongside the images, the interface also displays the user's previously selected wine type and calligrapher, reinforcing the semantic context behind the generated outputs.

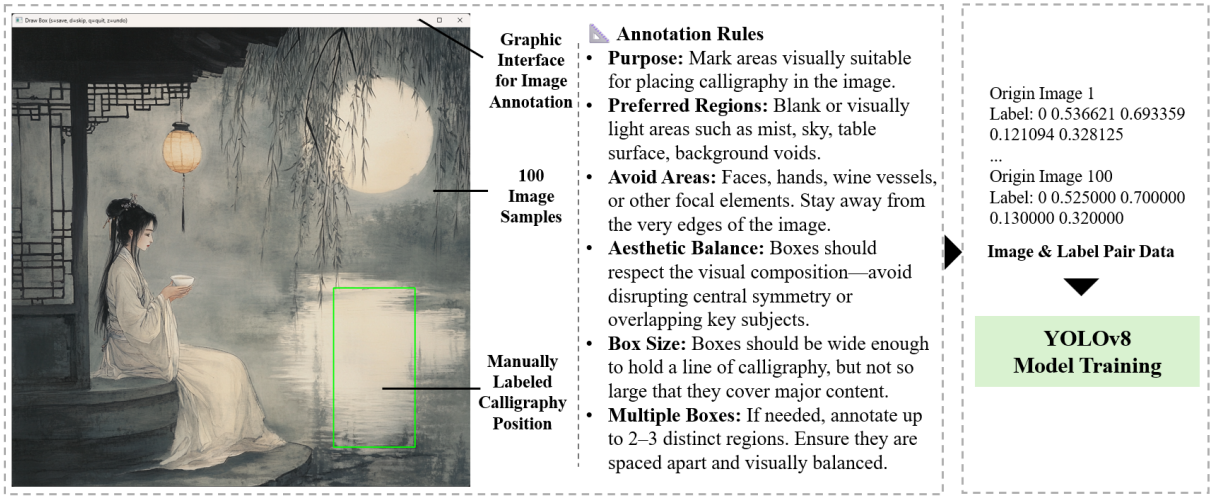


Figure 5. Image annotation and YOLOv8 training pipeline.

Algorithm 2 Calligraphy and Background Merging

Require: Input background image I_{bg}, composed calligraphy image $I_{composed}$, trained YOLOv8 model M
Ensure: Final composited image I_{out}

Predict bounding box $B = (x, y, w, h)$ from I_{bg} using model M
Extract region $R_{box} \subset I_{bg}$ corresponding to B
Compute mean brightness V_{mean} of R_{box} in grayscale
Define threshold $T_{dark} \leftarrow 0.4$ // Empirically chosen
if $V_{mean} < T_{dark}$ **then**
    Invert calligraphy color to white on transparent background
end if
Resize $I_{composed}$ to fit within (w, h) while preserving aspect ratio
Compute vertical offset y' to center $I_{composed}$ inside B
Overlay $I_{composed}$ on I_{bg} at position (x, y') using alpha blending
return I_{out}

4. Experiments for Model Selection

To ensure the reliability and cultural coherence of each generation stage, a series of experiments were conducted to evaluate model performance across three core tasks: wine-themed poem recommendation, text-to-image content generation, and calligraphy placement. Each subsection presents comparative results and informs the selection of default models used in the system pipeline.

4.1. LLM for Wine-Themed Poem Recommendation

To determine the most suitable large language model (LLM) for generating wine-themed Chinese poetry, we conducted a systematic comparison across five popular models: ChatGPT-4o, DeepSeek-R1, Gemini-2.5-Pro, Claude-Sonnet-4, and Grok-4. In each trial, the system randomly sampled 3 to 6 intention keywords covering emotional, symbolic, or cultural categories (e.g., “nostalgia”, “reclusion”, “plum blossom”), and embedded them into our standardized four-section prompt (see Figure 4). Each model was tasked with generating a classical-style poem based on these keywords, and the experiment was repeated 200 times per model. To evaluate output quality, we manually reviewed each generated poem according to two primary criteria: (1) semantic alignment, i.e., whether the poem meaningfully reflected the given intentions, and (2) factual correctness, i.e., whether the poem or the poem information contained fabricated or misattributed lines. Outputs with either misaligned content or invented historical text were marked incorrect.

The results, visualized in Figure 6, show that DeepSeek-R1 achieved a perfect score (100%), with ChatGPT-4o and Grok-4 following closely. Gemini-2.5-Pro and Claude-Sonnet-4 showed slightly higher error rates, often due to hallucinated classical content or weakened thematic correspondence. These findings led us to adopt DeepSeek-R1 as the default model in our prototype.

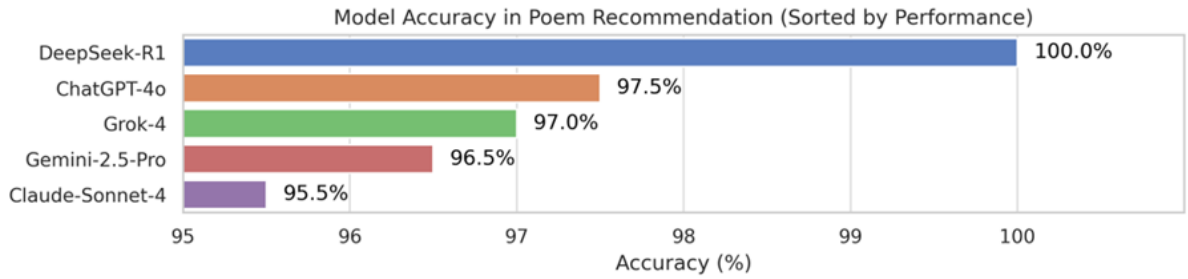


Figure 6. Accuracy of five LLMs in recommending wine-themed Chinese poems (200 trials each). Accuracy is measured as the percentage of semantically aligned and factually correct generations.

4.2. LLM & T2I Model for Conditioned Content Generation

To evaluate how different large language models (LLMs) and text-to-image (T2I) models contribute to culturally faithful visual generation, we designed a two-stage experiment. In the first stage, five LLMs—ChatGPT-4o, DeepSeek-R1, Gemini-2.5-Pro, Claude-Sonnet-4, and Grok-4—were tasked with converting a poetic line and wine context into a descriptive prompt suitable for Chinese painting synthesis. In the second stage, these prompts were fed into four T2I models—Midjourney v6, SDXL, FLUX, and ChatGPT-4o Vision—to generate images inspired by traditional Chinese aesthetics.

A representative example is shown in Figure 7, using the poem "争渡争渡，惊起一滩鸥鹭" ("Struggling to cross, struggling to cross, startling a sandbank of gulls and egrets") alongside the symbolic context of green plum wine. Across all trials, DeepSeek-R1 and ChatGPT-4o consistently produced prompts with strong semantic grounding and appropriate stylistic cues, such as mist, riverbanks, birds in flight, and natural scenery. Among T2I models, Midjourney v6 and ChatGPT-4o Vision most effectively translated these prompts into images that captured the spirit of classical ink painting. While this figure presents one illustrative case, we repeated the experiment with multiple poem and wine pairings to ensure generalizability. Across these broader tests, the combination of DeepSeek-R1 for prompt generation and Midjourney v6 for image rendering emerged as the most robust pairing, consistently yielding results that balanced aesthetic detail with stylistic fidelity. This model configuration was therefore adopted as the default in our final prototype.

4.3. Model for Calligraphy Placement

Accurate calligraphy placement plays a critical role in preserving the visual harmony and cultural authenticity of generated Chinese paintings. To evaluate different placement strategies, we compared our customized model—YOLOv8 trained on manually annotated calligraphy layout data—with several vision-capable large language models (e.g., ChatGPT-4o, Gemini-2.5-Pro, Claude-Sonnet-4) that support image-based reasoning. Each method was given the same generated background image and a calligraphy infomation of classical poem line, with the goal of determining the optimal location, scale, and orientation for overlaying the calligraphy text. As shown in Figure 8, three visual scenes are tested to demonstrate how each system handles layout in relation to subject focus, lighting, and surrounding detail.

While vision-capable LLMs are proficient in general reasoning about spatial composition, they are not specifically optimized for the conventions of Chinese calligraphy aesthetics. In our evaluation, they frequently positioned text over semantically salient elements such as faces, hands, or focal lighting areas, resulting in visual clutter or cultural incongruity. In contrast, the YOLOv8 model, trained on over 100 manually labeled examples following traditional placement rules, consistently selected low-texture, unobtrusive regions—such as mist, negative space, or distant background elements. It also adjusted vertical alignment and aspect ratio to better match the spatial rhythm of classical paintings. Notably, experiments show that all tested vision-capable LLMs demonstrate a general understanding of composition and can articulate reasonable placement strategies in natural language. When prompted to describe suitable regions for calligraphy insertion, these models often suggested appropriate areas. However, when tasked with producing explicit bounding box coordinates, their outputs became unreliable—often overlapping important elements or exceeding canvas boundaries, as illustrated in Figure 8. This highlights a key limitation: while LLMs can support human-in-the-loop design workflows, they currently lack the spatial precision required for fully automated, coordinate-based frameworks.

Figure 7. Prompt-to-image comparison across five LLMs (rows) and four T2I models (columns). Each prompt was generated from the poetic line "争渡争渡，惊起一滩鸥鹭" and paired with green plum wine. Midjourney and ChatGPT-4o Vision rendered the most stylistically coherent images, especially when driven by prompts from DeepSeek-R1 and ChatGPT-4o.

4.4. Generalizability Test of Trained YOLOv8 Model

To examine the generalizability of the trained layout prediction model, a test was conducted using images generated from text-to-image (T2I) systems other than Midjourney v6. The YOLOv8 model had been trained on 100 annotated images synthesized exclusively with Midjourney v6. For evaluation, additional images were generated with ChatGPT-4o and Gemini, and the YOLOv8 detector was applied to predict suitable regions for calligraphy placement before merging the calligraphy. Representative results are shown in Figure 9. The outcomes demonstrate that, although the training set consisted only of Midjourney-generated images, the YOLOv8 model can also be effectively applied to images from other T2I systems. The detected regions supported culturally appropriate calligraphy placement, maintaining balance while avoiding semantic focal points such as birds or vessels. This indicates that the model captures layout cues in a transferable way rather than overfitting to a single image generator.

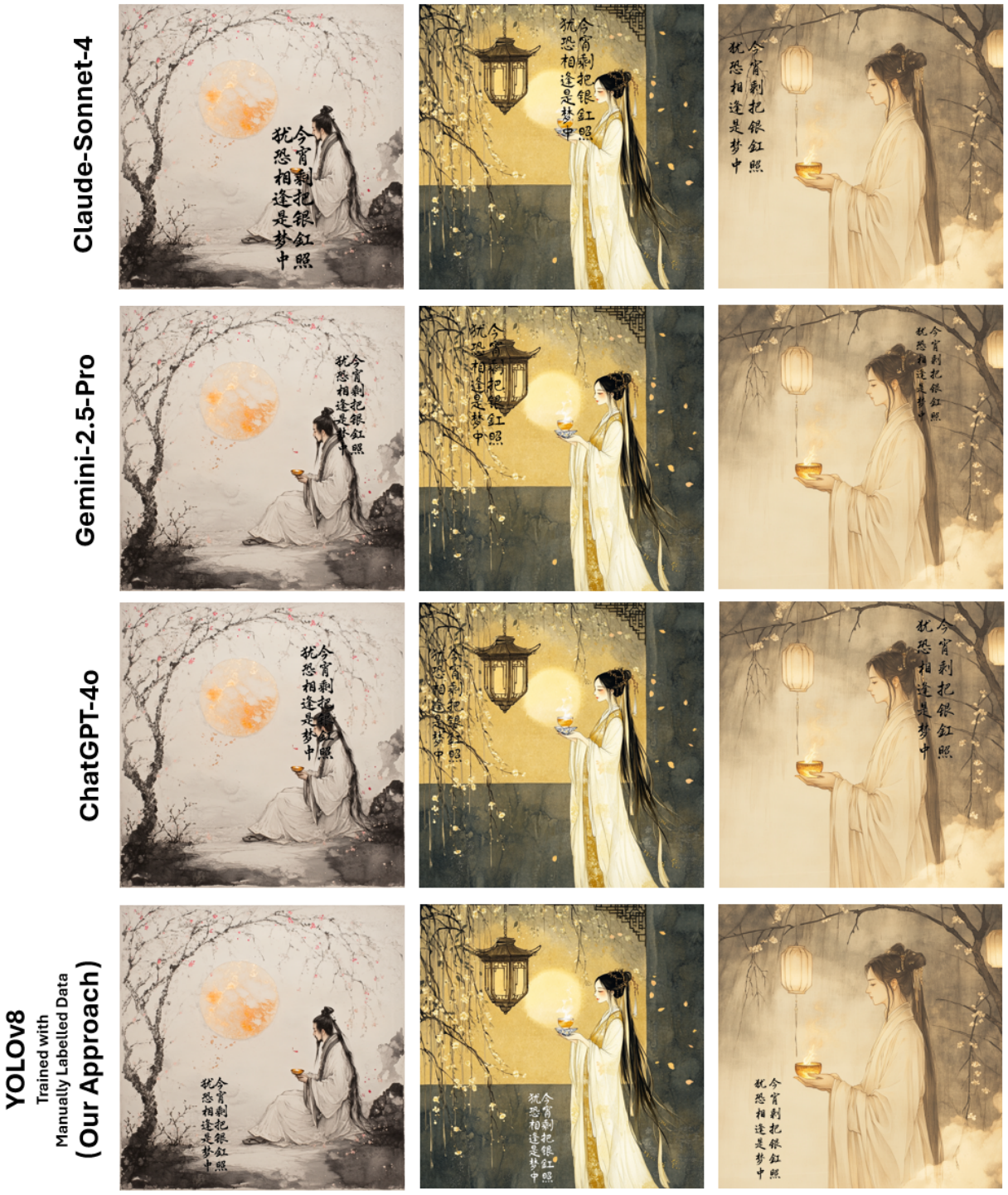


Figure 8. Comparison of calligraphy placement across vision-capable LLMs and a YOLOv8 model trained on annotated layout data. Each row shows how different systems overlay the same poem on generated paintings.

It should be noted that the trained YOLOv8 model is not detecting an objectively existing object category (such as cars or cats), but instead predicting "suitable regions for calligraphy placement." This is a subjective, aesthetics-driven task without strictly defined boundaries. During inference, the system directly selects the bounding box with the highest confidence score as the placement region, ensuring consistency in the pipeline. Consequently, conventional object detection metrics are naturally low (Box Precision = 0.231, Recall = 0.471, mAP50 = 0.199). Nevertheless, despite these numerical values, the predicted regions are visually appropriate and enable coherent integration of calligraphy into the generated backgrounds. At present, it remains difficult to quantitatively assess whether a calligraphy placement is aesthetically appropriate, since no established computational aesthetic framework exists for calligraphy painting. Future research may focus on developing such quantitative metrics, which could provide a more systematic validation of the predictions of the model.

Background Image Generated by ChatGPT **Background Image Generated by Gemini**

Figure 9. Generalizability test of the YOLOv8 model. Trained only on Midjourney v6 images, it provides suitable calligraphy placement on backgrounds generated by ChatGPT-4o (**left**) and Gemini (**right**).

5. Results & Evaluation

Figure 10 presents some examples of generated results with *Verse-in-Wine* framework. To systematically assess the semantic and stylistic quality of our generated outputs, we curated a benchmark set consisting of 75 classical Chinese drinking-related poems across major dynasties. For each poem, four distinct calligraphy compositions were generated by varying the background and style, resulting in a total of 300 samples.

Figure 10. Some Examples of Generated Results with *Verse-in-Wine* Framework.

These images were evaluated by GPT-4o along four dimensions: *Wine Style Alignment*, *Emotion Alignment*, *Object Alignment*, and *Historical/Cultural Accuracy*. Each image was scored using a carefully designed prompt, which provided the LLM with the poem line, its full context, and a detailed description of the selected wine. The prompt enforces strict output constraints in JSON format. The full colored prompt structure is provided in Supplementary Materials. We adopt the similar method of ELsharif et al [32], each dimension is scored on a five-point ordinal scale $\{-1, -0.5, 0, 0.5, 1\}$. The final overall score is computed as the mean of the four dimension scores. The scoring dimensions are briefly defined as follows:

1. **Wine Alignment** : Whether the visual cues reflect the wine's symbolic, sensory, and seasonal characteristics.
2. **Emotion Alignment**: Whether the emotional tone of the image matches the sentiment of the poetic line.
3. **Object Alignment**: Whether key poetic elements are accurately represented.
4. **Historical Accuracy**: Whether the image adheres to traditional Chinese cultural aesthetics and avoids modern artifacts.

To illustrate model performance variation, we include a representative low-quality output alongside the aggregated evaluation scores. As shown in Figure 11, the left image depicts a generated artwork that received low alignment scores due to visual inconsistency with the poetic context—specifically, the inclusion of modern tableware elements that conflict with the intended historical atmosphere. On the right, the bar chart summarizes the average LLM-based evaluation scores across 300 generated samples. The framework performs particularly well in *Emotion Alignment* (0.962), while maintaining high consistency across the other three cultural dimensions. The overall average correctness score reaches 0.916, demonstrating the system's capability to produce semantically coherent and culturally faithful outputs at scale. Highly scored examples can be found in Figure 10.

Figure 11. (**Left**): A representative low-scoring generated image (score across 4 dimensions in order: -1, -0.5, -0.5, -1), where modern objects such as wine bottles and bowls conflict with the intended poetic atmosphere. (**Rignt**): Average LLM-based evaluation scores across 300 generated samples.

6. Prototype Implementation & User Study

To validate the effectiveness of the proposed framework, a user-interactive prototype titled *Verse-in-Wine* was developed. This prototype corresponds directly to the frontend pipeline illustrated in the system architecture (see Figure 3), and provides a complete user-facing implementation of the poetry-to-painting workflow. As shown in Figure 12, the interface consists of four sequential modules: (1) **Keywords selection**, where users specify abstract intentions across multiple cultural and emotional dimensions; (2) **Poem display**, where the system presents LLM-recommended poetic lines with interpretive context for user comparison and selection; (3) **User preference selection**, in which the user selects their preferred calligraphy style and wine pairing to condition the final generation; and (4) **Final generated results display**, showcasing four stylized artworks alongside the chosen textual and aesthetic elements. The interface supports both Chinese and English, and emphasizes clarity, cultural expressiveness, and ease of interaction. The prototype can be access here.

A user study was conducted with 25 participants, each of whom completed a full creative session using the platform. The 25 participants were general users with graduate-level education backgrounds; several had prior exposure to Chinese literature or art. The study was exploratory and conducted internally under academic supervision. All participants were informed of the study purpose and voluntarily signed a consent form prior to participation. The questionnaire was designed in two parts. The first part included three binary-choice questions evaluating visual quality, usability, and time efficiency. The second part adopted a five-point ordinal scale aligned with the four semantic dimensions used in LLM-based evaluation: wine alignment, emotion alignment, object alignment, and historical alignment. Each participant evaluated four generated images, corresponding to the four dimensions, resulting in a total of 100 human-evaluated samples (25 participants × 4 images each).

All participants confirmed that the prototype was easy to use and produced high-quality visual outputs. Notably, 90% of users completed the entire creative process in under eight minutes, indicating strong usability. To assess semantic alignment, each participant evaluated four samples using a five-point ordinal scale across the same four dimensions used in LLM-based evaluation: wine alignment, emotion alignment, object alignment, and historical alignment. The resulting human scores are summarized alongside the LLM evaluation results in Table 3. While scores were generally consistent across both evaluation modes, slight discrepancies in wine and object alignment suggest nuanced differences in perception between human users and LLMs. The overall average human score

(0.890) closely matches the LLM average (0.9165), affirming both the system's semantic fidelity and its effectiveness in delivering a culturally coherent user experience.

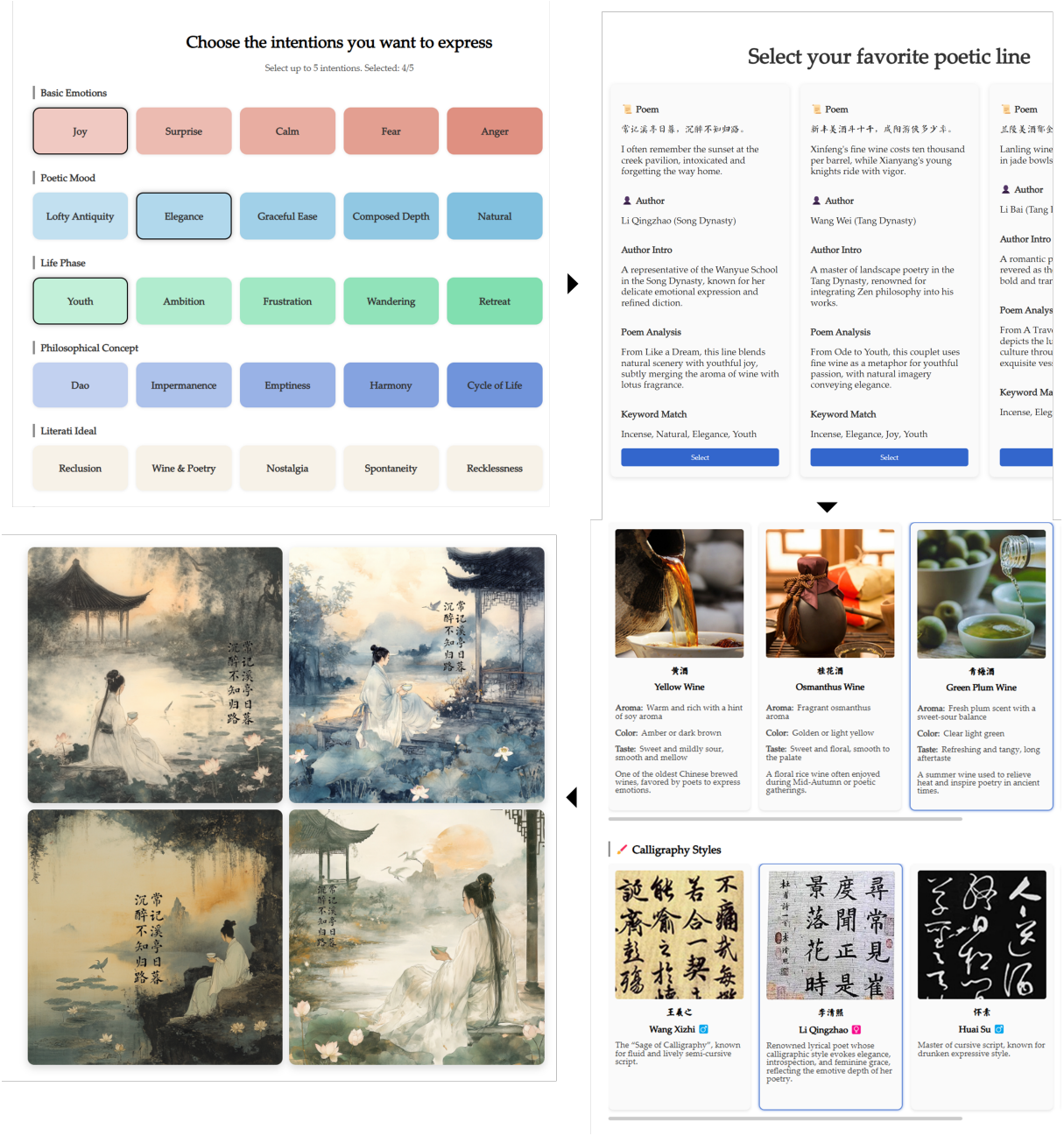


Figure 12. User interface of the prototype system based on *Verse-in-Wine* framework, demonstrating Keywords selection, poem display, User preference selection, and final generated results display.

Table 3. Comparison of LLM-based and Human Evaluation Scores Across Four Dimensions.

	Sample Amount	Wine Alignment	Emotion Alignment	Object Alignment	Historical Alignment	Overall Score
LLM Evaluation	300	0.864	0.962	0.898	0.942	0.9165
User Study	100	0.880	0.840	0.900	0.940	0.8900

7. Discussion

While the system demonstrates strong cultural alignment and user satisfaction, several aesthetic and generative limitations persist. One notable issue is the lack of stylistic variation in generated calligraphy. As expressive diversity is a core aspect of Chinese calligraphic aesthetics, this limitation constrains the system's ability to fully capture the liveliness of traditional brushwork.

To introduce variation, one approach is to assign different reference images to each character. However, as illustrated in Figure 13, this strategy causes stylistic inconsistency across characters, disrupting the visual coherence of the final output. In contrast, using a single reference image preserves unity but produces overly uniform strokes that lack expressive richness. This trade-off between variation and coherence presents a key challenge in calligraphy generation. Future research should explore adaptive mechanisms that balance individuality with stylistic consistency at the phrase level.

Another unresolved limitation lies in the system's inability to generate cursive script, which represents one of the most expressive and emotionally charged forms of Chinese calligraphy. Unlike regular or semi-cursive styles, cursive script emphasizes fluidity, speed, and gestural freedom, often sacrificing legibility for aesthetic spontaneity. These characteristics make it exceptionally difficult for current generative models, which rely on structurally aligned reference images, to replicate the dynamic and continuous stroke transitions inherent in cursive writing. This limitation is particularly significant in the context of Chinese wine culture. Historically, cursive script was closely associated with intoxication, spontaneity, and artistic release—values celebrated by poets and calligraphers alike. The brushwork of figures such as Huaisu or Zhang Xu exemplifies this fusion of emotional excess and visual abstraction. Integrating cursive generation into future models would not only enhance stylistic coverage but also deepen the semantic resonance between visual form and the cultural theme of intoxicated expression.

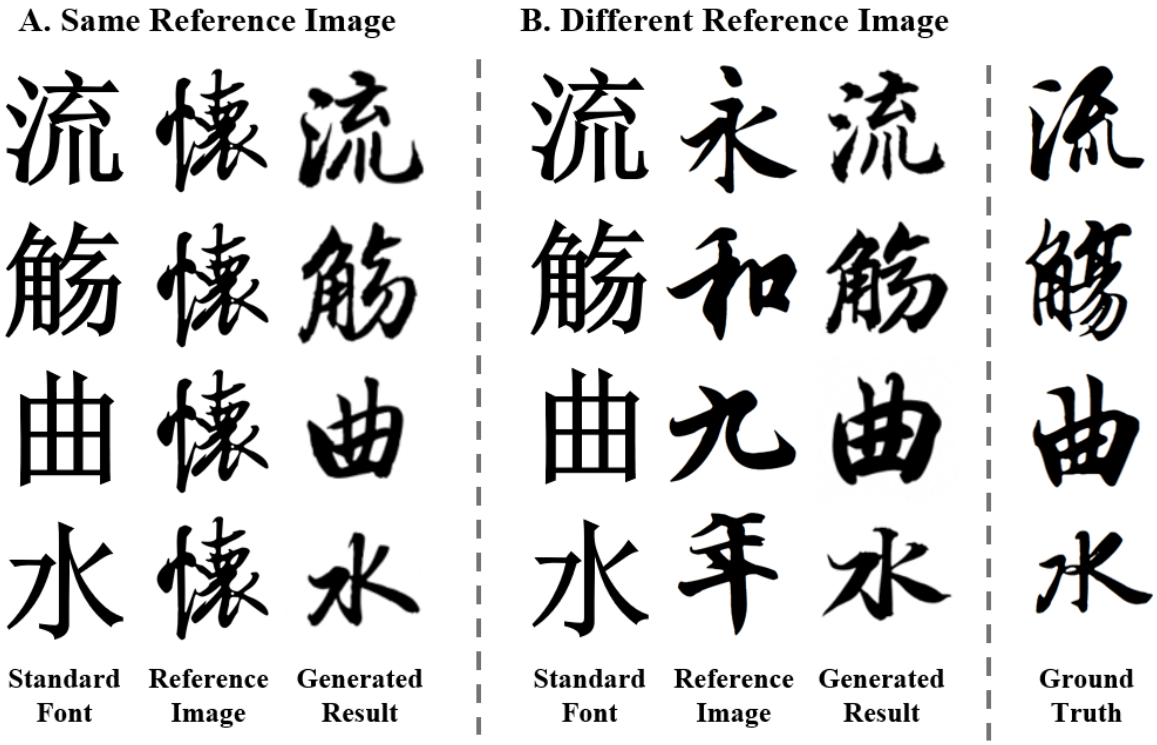


Figure 13. Comparison of calligraphy generation using different reference strategies.

8. Conclusions and Future Work

This paper presents *Verse-in-Wine*, a culturally grounded generative framework that synthesizes Chinese classical poetry, traditional wine culture, and AI-generated calligraphy painting into a coherent creative pipeline. Through a combination of semantic keyword embedding, large language model (LLM) guidance, and visual generation, the system enables users to explore poetically expressive and visually harmonious outputs. A fully functional prototype was developed to support end-to-end user interaction, incorporating multilingual UI design and structured poetic workflows. The system was rigorously evaluated through both automated and human assessments. The LLM-based evaluation yielded an average score of 0.9165 across four culturally specific dimensions, while a user study involving 25 participants and 100 creative samples reported a closely aligned average human rating of 0.8900. These results confirm not only the semantic and aesthetic quality of the generated content but also the system's usability and cross-modal coherence.

Beyond these promising outcomes, our qualitative findings reveal important challenges that shape future research. First, the generation of stylistically varied yet coherent calligraphy remains an open problem: attempts to inject character-level variation via multiple reference images disrupted visual unity, while single-reference generation produced overly uniform strokes. Moreover, the system does not yet support cursive script, which is

a style intimately associated with emotional spontaneity and intoxicated expression in Chinese artistic tradition. Addressing this gap is critical to deepening the framework's cultural and aesthetic resonance. Moving forward, we aim to develop fine-grained control mechanisms that allow expressive diversity while maintaining stylistic cohesion, and to extend the system's capabilities to include cursive generation via stroke-continuous modeling. We also plan to deploy the prototype in physical cultural venues, such as wine bars and gallery spaces, to explore its potential as a medium for public engagement and AI-mediated cultural storytelling.

Supplementary Materials

The additional data and information can be downloaded at: https://media.sciltp.com/articles/others/2510211543270664/TAI-25080348-Supplementary-Materials.zip.

Author Contributions

R.C.: data curation, writing—original draft preparation, reviewing and editing; J.S.:supervision, reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

No data is being made available.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-assisted Technologies

During the preparation of this work, the authors used generative AI to polish writing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

1. Chen, H. Elegant Scholar Sipping Wine, Ming Dynasty. Artwork preserved at the Shanghai Museum, Shanghai, China.
2. Chenghua. The Eight Immortals Drinking (partial), Ming Dynasty. Artwork preserved at the Palace Museum, Beijing, China.
3. Pourreza, R.; Bhattacharyya, A.; Panchal, S.; et al. Painter: Teaching Auto-regressive Language Models to Draw Sketches. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2–3 October 2023; pp. 305–314.
4. Rombach, R.; Blattmann, A.; Lorenz, D.; et al. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10674–10685.
5. Qu, L.; Wu, S.; Fei, H.; et al. LayoutLLM-T2I: Eliciting Layout Guidance from LLM for Text-to-Image Generation. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), Ottawa, ON, Canada, 29 October–3 November 2023; pp. 643–654.
6. Yang, Z.; Peng, D.; Kong, Y.; et al. FontDiffuser: One-Shot Font Generation via Denoising Diffusion with Multi-Scale Content Aggregation and Style Contrastive Learning. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 6603–6611.
7. Follmer, S.; Brade, S.; Wang, B.; et al. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *ACM Symposium on User Interface Software and Technology, UIST*; Association for Computing Machinery: New York,NY, USA, 2023.
8. Arawjo, I.; Swoopes, C.; Vaithilingam, P.; et al. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), Honolulu, HI, USA, 11–16 May 2024.
9. Zhang, Y.; Fang, Z.; Yang, X.; et al. Reconnecting the Broken Civilization: Patchwork Integration of Fragments from Ancient Manuscripts. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), Ottawa, ON, Canada, 29 October–3 November 2023; pp. 1157–1166.
10. Zhu, S.; Xue, H.; Nie, N.; et al. Reproducing the Past: A Dataset for Benchmarking Inscription Restoration. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 7714–7723.

11. Pan, J.; Li, L.; Yamaguchi, H.; et al. Reconstructing, Understanding, and Analyzing Relief Type Cultural Heritage from a Single Old Photo. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 7724–7733.
12. Bin, Y.; Shi, W.; Ding, Y.; et al. GalleryGPT: Analyzing Paintings with Large Multimodal Models. In Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 7734–7743.
13. Silva, M. Interaction with Immersive Cultural Heritage Environments: Using XR Technologies to Represent Multiple Perspectives on Serralves Museum. In Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Lisboa, Portugal, 10–14 October 2022; pp. 6920–6924.
14. Rachabatuni, P.K.; Principi, F.; Mazzanti, P.; et al. Context-aware chatbot using MLLMs for Cultural Heritage. In Proceedings of the ACM Multimedia Systems Conference (MMSys '24), Bari, Italy, 15–18 April 2024; pp. 459–463.
15. Zhou, A.L.; Zhang, K. Shanshui Journey: Using AI to Reproduce the Experience of Chinese Literati Ink Paintings. *Leonardo* **2024**, *57*, 370–378.
16. Isola, P.; Zhu, J.Y.; Zhou, T.; et al. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5967–5976.
17. Zhu, J.Y.; Park, T.; Isola, P.; et al. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
18. Kim, J.; Kim, M.; Kang, H.; et al. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
19. Cai, Y.T. zi2zi: Master Chinese Calligraphy with Conditional Adversarial Networks. 2017. Available online: https://github.com/kaonashi-tyc/zi2zi (accessed on 14 January 2024).
20. Chang, B.; Zhang, Q.; Pan, S.; et al. Generating Handwritten Chinese Characters Using CycleGAN. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 199–207.
21. Liu, R.; Yuan, S.; Chen, M.; et al. MaLiang: An Emotion-driven Chinese Calligraphy Artwork Composition System. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20), New York, NY, USA, 12–16 October; pp. 4394–4396.
22. Zhou, P.; Zhao, Z.; Zhang, K.; et al. An End-to-End Model for Chinese Calligraphy Generation. *Multimed. Tools Appl.* **2021**, *80*, 6737–6754.
23. Tuo, Y.; Xiang, W.; He, J.Y.; et al. AnyText: Multilingual Visual Text Generation and Editing. *arXiv* **2023**, arXiv:2311.03054.
24. Chen, Y.S.; Chao, M.T. Skeletonization application: Chinese calligraphy character representation and reconstruction. *J. Electron. Imaging* **2018**, *27*, 051202.
25. Chao, M.T.; Chen, Y.S. A compact representation of character skeleton using skeletal line based shape descriptor. In *Applications of Digital Image Processing XLII*; Tescher, A.G., Ebrahimi, T., Eds.; SPIE: San Diego, CA, USA, 2019; p. 99.
26. Wang, T.Q.; Liu, C.L. Fully Convolutional Network Based Skeletonization for Handwritten Chinese Characters. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 11868.
27. Wang, T.Q.; Jiang, X.; Liu, C.L. Query Pixel Guided Stroke Extraction with Model-Based Matching for Offline Handwritten Chinese Characters. *Pattern Recognit.* **2022**, *123*, 108416.
28. Jiang, Y.; Lian, Z.; Tang, Y.; et al. SCFont: Structure-Guided Chinese Font Generation via Deep Stacked Networks. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 4015–4022.
29. Lian, Z.; Zhao, B.; Xiao, J. Automatic generation of large-scale handwriting fonts via style learning. In *SIGGRAPH Asia 2016 Technical Briefs*; ACM: Macau, China, 2016; pp. 1–4.
30. Yuan, S.; Dai, A.; Yan, Z.; et al. Learning to Generate Poetic Chinese Landscape Painting with Calligraphy. *arXiv* **2023**, arXiv:2305.04719.
31. Cai, R.; She, J. Pop Calligraphy Artwork: AI Meets Guangzhong Wu on Social Media. In Proceedings of the 17th International Symposium on Visual Information Communication and Interaction, New York, NY, USA, 11–13 December 2024.
32. ELsharif, W.; Agus, M.; Alzubaidi, M.; et al. Cultural Relevance Index: Measuring Cultural Relevance in AI-Generated Images. In Proceedings of the IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 7–9 August 2024; pp. 410–416.